Additional Analyses: Social Desirability Rating and LLM Responses

Main text findings point out the consistent differences between LLM responses and human responses. What could have caused this difference? A reasonable hypothesis is the influence of social desirability bias, where LLMs trained on human data may exhibit human-like social desirability biases, influencing their responses to align with socially accepted behaviors and traits. Previous research supports this hypothesis; for example, Hilliard et al. (2024) found that newer, larger-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 human-like social desirability biases when generating simulated data. In this section, we examined how the responses of human respondents and LLMs are influenced by social desirability ratings. Specifically, we used social desirability ratings to predict and examine the correlations with the item means of human data, LLM data, and their differences. By analyzing these predictions and correlations for the respective item means, we aimed to explore the characteristics and patterns of how social desirability ratings influence human and LLM data. Furthermore, through the prediction and correlation analysis of the differences in item means between human data and LLM data, we investigated whether social desirability ratings affect human and model data in the same way.

Method

Social desirability ratings 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 Ph.D. students and four Ph.D. holders 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.

Results

The results are presented in Table 1 and Table 2 (for the regression line chart, see Figure 1 and Figure 2). Both BFI-2 and HEXACO-100 results demonstrated a strong positive correlation between social desirability ratings and both LLM 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 an item's social desirability rating is neutral (4), most LLMs tend to provide a neutral response (3 on the scale). This suggests that LLMs have learned to associate neutral social desirability with neutral opinions (just like previous findings; Salecha et al., 2024). The slight deviations observed in Llama3 and GPT models indicate that different LLMs may have subtle biases or tendencies in their response patterns.

For the mean differences between human responses and LLM 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 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 responses also increased. For most human responses and LLM 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 responses, and when the rating was low, the item mean in human responses was lower than that in LLM 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 1

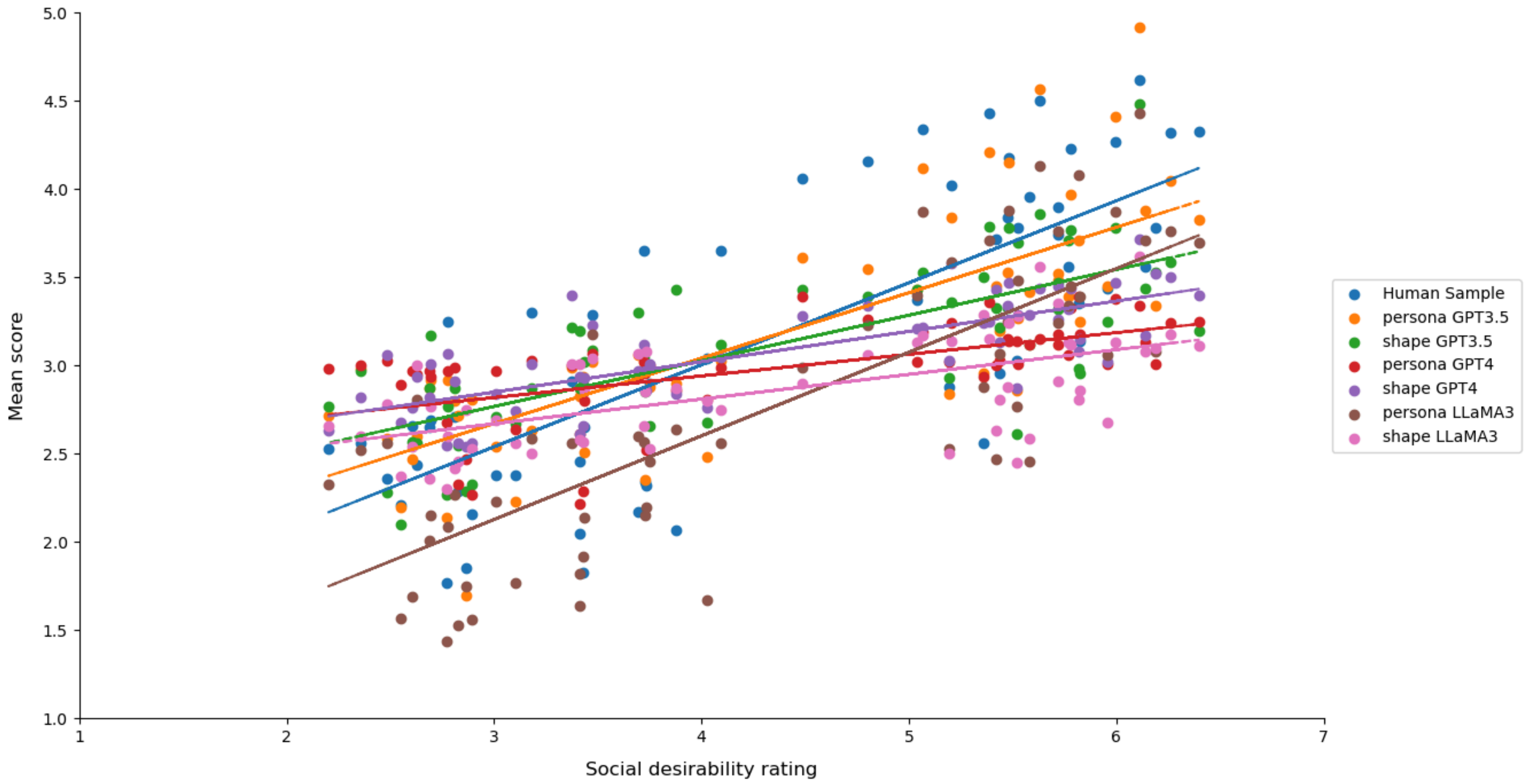
Regression Analysis and Correlation of Social Desirability Ratings: Human Responses vs. BFI-2 LLM Responses

	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	
			MSD$_{persona}$ GPT3.5		MSD$_{shape}$ GPT3.5		MSD$_{persona}$ GPT4		MSD$_{shape\ \mathrm{GPT4}}$		MSD$_{persona}$ Llama3		MSD$_{shape}$ 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* GPT3.5, *n* = 297 for *shape* Llama3, and *n* = 300 for other LLM responses. Some sample sizes are below 300 because certain generated data exceeded reasonable thresholds for specific items and were excluded from the analysis. MSD = mean score difference.

Figure 1

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



Note. n = 1,559 for human responses, *n* = 299 for *persona* GPT3.5, *n* = 297 for *shape* Llama3, and *n* = 300 for other LLM responses. Some sample sizes are below 300 because certain generated data exceeded reasonable thresholds for specific items and were excluded from the analysis.

Table 2

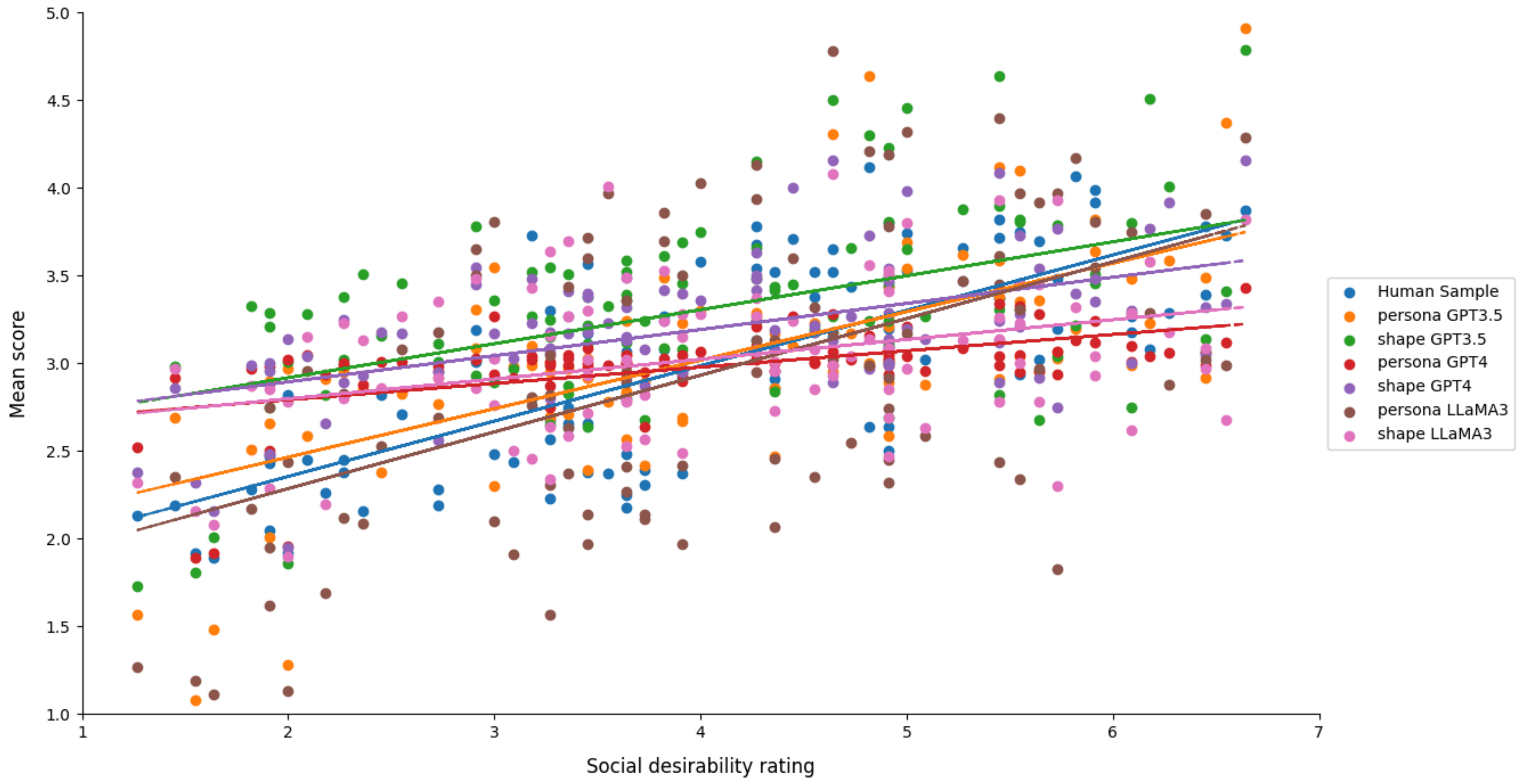
Regression Analysis and Correlation of Social Desirability Ratings: Human Responses vs. HEXACO-100 LLM Responses

	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	
			MSD$_{persona}$ GPT3.5		MSD$_{shape}$ GPT3.5		MSD$_{persona}$ GPT4		MSD$_{shape}$ GPT4		MSD$_{persona}$ Llama3		MSD$_{shape}$ 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* GPT3.5, and n = 300 for other LLM responses. Some sample sizes are below 300 because certain generated data exceeded reasonable thresholds for specific items and were excluded from the analysis. MSD = mean score difference.

Figure 2

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



Note. n = 7,204 for human responses, *n* = 298 for *persona* GPT3.5, and *n* = 300 for other LLM responses. Some sample sizes are below 300 because certain generated data exceeded reasonable thresholds for specific items and were excluded from the analysis.

References

Hilliard, A., Munoz, C., Wu, Z., & Koshiyama, A. S. (2024). Eliciting big five personality traits in large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.08341

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in big five personality surveys. *PNAS Nexus*, *3*(12). https://doi.org/10.1093/pnasnexus/pgae533