Additional Analyses: Distinguishing Human vs. LLM Responses

Beyond differences in descriptive statistics and psychometric performance, the distinction between LLM responses and human responses can be further examined by mixing the two types of data for a classification task. If LLM can mimic human responses, it would be very difficult to differentiate between human respondents and LLM respondents in a combined dataset. Specifically, we used 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 responses and human responses, as well as HEXACO-100 LLM responses and human responses. LLM responses were considered the positive class, while human responses were considered the negative class (binary dependent variable). For the independent variable features (input features) of the logistic regression model, we have two approaches: (1) the mean and standard deviation of all items as independent variable features, and (2) individual items themselves as independent variable features. Using the mean and standard deviation of all items as input features can be seen as a simplified approach to processing all item features. It also serves as a higher-level aggregation analysis method, providing a general reflection of the overall statistical properties of the data.

We applied the same processing method to the BFI-2 and HEXACO data. To ensure a balanced analysis environment, we randomly selected 300 samples from the human response data to compare with the LLM responses in each group (with sample sizes close to 300) and conducted the analysis using logistic regression. To avoid selection bias, we repeated the random sampling of human data 50 times. For each analysis, we used five-fold cross-validation to ensure

the robustness of the results. Specifically, we divided the data into five parts, training and validating the model on each part to improve the reliability of the results. We report the average (*M*) and variability (*SD*) of the results obtained from the five-fold cross-validation process.

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 across the 50 replications for key performance metrics such as accuracy, precision, recall, and F1 score. These results have been comprehensively detailed in Table 1 and Table 2. If LLM responses were highly similar to human responses, the classification accuracy should be close to random. If the classification accuracy is higher than random, it indicates that LLM responses are distinct from human responses and they cannot replace human participants in psychometric research.

Results

Mean and Standard Deviation of All Items As Input Features

Table 1 shows the results when the mean and standard deviation of all items used as input features. For the BFI-2 data, a clear pattern emerged: 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 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.

For HEXACO-100 data, a similar pattern was observed 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 responses generated by the GPT series models using the *persona* and *shape* methods were larger, and the 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 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.

Individual Items Themselves As Input Features

Table 2 shows the results when the individual items themselves are used as input features. For the BFI-2 data, similar to the findings when the mean and standard deviation of all items were used as input features, the accuracy, precision, recall, and F1 score metrics of the *shape* method for the GPT-4 model were also lower than those of the *persona* method. However, for the GPT-3 model, they were comparable ($M_{\text{accuracy } persona \text{ GPT3.5}} = 0.92$ vs. $M_{\text{accuracy } shape \text{ GPT3.5}} = 0.93$, $M_{\text{precision } persona \text{ GPT3.5}} = 0.92$ vs. $M_{\text{precision } shape \text{ GPT3.5}} = 0.93$, $M_{\text{recall } persona \text{ GPT3.5}} = 0.92$ vs. $M_{\text{recall } shape \text{ GPT3.5}} = 0.93$, $M_{\text{F1 } persona \text{ GPT3.5}} = 0.92$ vs. $M_{\text{F1 } shape \text{ GPT3.5}} = 0.93$; $M_{\text{accuracy } persona \text{ GPT4}} = 0.95$ vs. $M_{\text{accuracy } shape \text{ GPT4}} = 0.89$, $M_{\text{precision } persona \text{ GPT4}} = 0.96$ vs. $M_{\text{precision } shape \text{ GPT4}} = 0.89$, $M_{\text{recall } persona \text{ GPT4}} = 0.95$ vs. $M_{\text{recall } shape \text{ GPT4}} = 0.89$, $M_{\text{F1 } persona \text{ GPT4}} = 0.95$ vs. $M_{\text{F1 } shape \text{ GPT4}} = 0.89$)

For HEXACO-100 data, a similar pattern was observed where the accuracy and precision metrics of the *shape* method in the GPT-4 model were also lower than those of the *persona* method, whereas in the GPT-3 model, they were comparable ($M_{\text{accuracy } persona \text{ GPT3.5}} = 0.98$ vs. $M_{\text{accuracy } shape \text{ GPT3.5}} = 0.98$, $M_{\text{precision } persona \text{ GPT3.5}} = 0.98$ vs. $M_{\text{precision } shape \text{ GPT3.5}} = 0.98$, $M_{\text{recall } persona \text{ GPT3.5}} = 0.98$ vs. $M_{\text{recall } shape \text{ GPT3.5}} = 0.98$, $M_{\text{F1 } persona \text{ GPT3.5}} = 0.98$ vs. $M_{\text{F1 } shape \text{ GPT3.5}} = 0.98$; $M_{\text{accuracy } persona \text{ GPT4}} = 0.96$ vs. $M_{\text{accuracy } shape \text{ GPT4}} = 0.95$, $M_{\text{precision } persona \text{ GPT4}} = 0.97$ vs. $M_{\text{precision } shape \text{ GPT4}} = 0.95$, $M_{\text{recall } persona \text{ GPT4}} = 0.96$ vs. $M_{\text{recall } shape \text{ GPT4}} = 0.95$, $M_{\text{F1 } persona \text{ GPT4}} = 0.96$ vs. $M_{\text{F1 } shape \text{ GPT4}} = 0.95$).

Comparison of Two Approaches

Comparing Table 1 and Table 2 reveals a striking improvement in accuracy when individual items are used as input features for logistic regression. This approach demonstrates a near-perfect ability to differentiate between LLM responses and human responses. The enhanced performance suggests that individual item-level features capture richer, more granular information than aggregate-level features, enabling more precise detection of differences

between LLM and human responses. This finding aligns with the descriptive statistics and psychometric performance analyses presented in the main text, which indicate that while LLMs can approximate higher-level aggregated patterns more closely, they struggle to accurately simulate finer-grained, item-specific characteristics.

Regardless of the method employed, although the classification accuracy for different LLM responses varied, all consistently exceeded 0.50, the threshold for random classification. This result underscores the significant distinctions between responses generated by LLMs and those produced by humans. The codes provided in the online supplementary materials can also be used to test whether the personality scale data (from the same scale) obtained online originates from human participants, for those who are interested.

Table 1

Classification Results between LLM Responses and Human Responses (Mean and Standard Deviation of All Items As Input Features)

	BFI-2								HEXACO-100							
	Accuracy		Precision		Recall		F1 score		Accuracy		Precision		Recall		F1 score	
LLM Responses	*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* GPT3.5, n = 297 for *shape* Llama3, and n = 300 for other LLM responses; for HEXACO-100, 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. The mean (*M*) and standard deviation (*SD*) represent the average and variability of the results obtained from the five-fold cross-validation process.

Table 2

Classification Results between LLM Responses and Human Responses (Individual Items Themselves As Input Features)

	BFI-2								HEXACO-100							
	Accuracy		Precision		Recall		F1 score		Accuracy		Precision		Recall		F1 score	
LLM Responses	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
persona GPT3.5	0.92	0.02	0.92	0.02	0.92	0.02	0.92	0.02	0.98	0.01	0.98	0.01	0.98	0.01	0.98	0.01
shape GPT3.5	0.93	0.02	0.93	0.02	0.93	0.02	0.93	0.02	0.98	0.01	0.98	0.01	0.98	0.01	0.98	0.01
persona GPT4	0.95	0.02	0.96	0.01	0.95	0.02	0.95	0.02	0.96	0.01	0.97	0.01	0.96	0.01	0.96	0.01
shape GPT4	0.89	0.02	0.89	0.02	0.89	0.02	0.89	0.02	0.95	0.02	0.95	0.02	0.95	0.02	0.95	0.02
persona Llama3	0.96	0.02	0.96	0.02	0.96	0.02	0.96	0.02	0.99	0.01	0.99	0.01	0.99	0.01	0.99	0.01
shape Llama3	0.92	0.02	0.92	0.02	0.92	0.02	0.92	0.02	0.97	0.01	0.97	0.01	0.97	0.01	0.97	0.01

Note. For BFI-2, n = 1,559 for human responses, n = 299 for *persona* GPT3.5, n = 297 for *shape* Llama3, and n = 300 for other LLM responses; for HEXACO-100, 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. The mean (*M*) and standard deviation (*SD*) represent the average and variability of the results obtained from the five-fold cross-validation process.

References

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.