A. LLMs' Unreasonable Responses

The unreasonable responses mainly fall into the following five categories:

- All responses are the same number;
- All responses are greater than or equal to 4 or less than or equal to 4. (Due to the presence of both positive and negative descriptions for the same assessment dimension (e.g., Openness) in our personality knowledge questionnaire, it is impossible for a participant to answer with all responses greater than or equal to 4, indicating agreement or neutrality for all statements, or all responses less than or equal to 4, indicating disagreement or neutrality for all statements.);
- Responses fall outside the numerical range of 1 to 7;
- Unable to score: responses similar to the following text: "I'm sorry, but as an AI language model, I cannot provide a response to your prompt as it is not clear what you are asking for. Please provide more context or clarify your question for me to provide an accurate response."
- Responses are non-score-related content, such as merely repeating statements from the questionnaire.

B. Reliability of LLMs' Responses

In evaluating the anthropomorphic personality traits demonstrated by LLMs through human personality assessments, the reliability and validity of LLMs' responses to such questionnaires merit further scientific scrutiny. The study by Miotto et al. (2022) highlighted the necessity for a more formal psychometric evaluation and construct validity assessment when interpreting questionnaire-based measurements of LLMs' potential psychological characteristics. To address these concerns, we employed two distinct methods to examine the reliability of LLMs' responses systematically: *Logical Consistency* and *Split-Half Reliability*. These methods provide a structured approach to evaluating the consistency and reliability of responses, which is crucial for ensuring the robustness of our findings. Out of three selected personality scales, we chose TDA-100 (80 statements) for reliability testing. Each statement of TDA-100 has explicitly stated the specific assessment dimension and scoring direction (forward scoring or reverse scoring) (Goldberg, 1992), both of which are critical to our assessment of the reliability of LLM responses using the two subsequent methods. As for the basis model of TDA-100, the Big Five model, there are 5 assessment dimensions in total: neuroticism, extraversion, openness, agreeableness and conscientiousness.

The TDA-100 response format employs a 7-point Likert scale, with a scoring range of 1 to 7 for each statement. From 1 to 7, 1 indicates that the respondent believes the statement does not apply to them at all, and 7 indicates that the statement completely applies to them. Each assessment dimension consists of several statements, some of which are positive and others negative. Specifically, within a selected assessment dimension, the closer a respondent's score is to 7 for positive statements, the more they exhibit characteristics of that dimension. Conversely, the closer their score is to 7 for negative statements, the less they exhibit characteristics of that dimension. For example, consider two statements for the Extraversion dimension as shown below. Statement 1 is positive, while Statement 2 is negative.

Statement 1: *Finish what I start.*

Statement 2: *Leave things unfinished.*

A higher score for Statement 1 indicates greater extraversion, while a higher score for Statement 2 indicates greater introversion. Therefore, within each dimension, positive statements are scored forwardly, and negative statements are scored reversely (7 minus the original score). Thus, when calculating a respondent's score for any given dimension, the total score comprises the original scores for all positive statements plus (7 minus the original score) for all negative statements.

The first method, *Logical Consistency*, is employed to ensure that the LLMs' responses across the questionnaire are coherent and consistent. By integrating reverse-scored items, we are able to check whether the LLMs carefully read and seriously respond to the questions. And the distribution of forward and reverse scored items within each assessment orientation is shown in Table 2.

After collecting the data, we adjusted the answers of negative(reverse-scored) items to align them with the overall scoring direction of the questionnaire. In this way, if LLMs' responses to positive and adjusted negative items are statistically consistent, they will show a similar pattern or trend, as evidenced by a 7-point Likert scale in which all answers are greater

Table 2. Distribution of Forward and Reverse Scored Items

Orientation	Forward	Reverse
NEUROTICISM	9	5
EXTRAVERSION	10	10
OPENNESS	9	5
AGREEABLENESS	10	9
CONSCIENTIOUSNESS	6	7
TOTAL COUNT	44	36

than or equal to 4, or less than or equal to 4, which indicate that the LLMs have responded conscientiously and logically. We introduce the $\mathrm{Consistency}$ metric to measure the logical consistency of LLM responses with the following formula:

$$\mathrm{Consistency} = \frac{\frac{\mathrm{N_c}}{\mathrm{N_t}} - \mathrm{P_{min}}}{\mathrm{P_{max}} - \mathrm{P_{min}}}, \tag{5}$$

where N_c is the number of questions with the same response direction within each measurement tendency in the adjusted response, N_t is the number of all statements, $P_{\max}$ and $P_{\min}$ are the maximum and the minimum of the proportion of consistent responses in all the statements. The value of $P_{\max}$ is 1, representing that all the responses are internally consistent within each assessment orientation. The value of $P_{\min}$ is supposed to be $\frac{\sum\lceil\frac{N_i}{2}\rceil}{N_t}$, where N_t is the count of all of the scored statements and N_i is the count of scored statements in each assessment orientation. Hence, $P_{\min}$ equals to 0.5125. The range of $\mathrm{Consistency}$ is from 0 to 1. The closer the value of $\mathrm{Consistency}$ is to 1, the more internally consistent the LLM's responses are. Consequently, we can evaluate the LLM's responses based on the prior knowledge of human personality assessment questionnaires

The second method is *Split-Half Reliability*. We measure the reliability of LLM's responses by comparing two equal-length sections of the questionnaire. This approach is based on the assumption that if a test is reliable, then any two equal-length sections of it should produce similar results. We first divide the questionnaire into two equal-length sections while ensuring that the content of each section is basically the same, representing that the numbers of statements within any assessment dimension in two halves are the same, thereby ensuring the accuracy of the reliability assessment. Then, we compute the Spearman's rank coefficient between the scores of the two sections to measure their consistency. The specific formula is shown in Section 3.2. Larger values indicate higher internal consistency of the responses. Finally, we calculated the reliability of the overall responses by using the Spearman-Brown formula as follows:

$$\mathrm{Reliability} = \frac{2\mathrm{corr}}{1 + \mathrm{corr}}, \tag{6}$$

where corr is the Spearman's rank coefficient between the scores of the two sections. The range of $\mathrm{Reliability}$ is from negative infinity to 1. Only if the value of an LLM's responses $\mathrm{Reliability}$ metric is around the human level, we can make it for further investigation.

We assessed the reliability of seven LLMs' reponses. The results of the are shown in Table 3.

We have also recruited 16 human participants, comprising an equal number of males and females, all native Chinese speakers with an English proficiency level of C1 according to the Common European Framework of Reference for Languages (CEFR), representing that they can express themselves effectively and flexibly in English in social, academic and work situations. The average value (with standard deviation) of their $\mathrm{Consistency}$ and $\mathrm{Reliability}$ is 0.73 ± 0.13 and 0.69 ± 0.09. And the minimum value is 0.49 and 0.57. Therefore, we regard ChatGLM3, GPT-3.5-turbo, GPT4, Vicuna13b and Vicuna33b as LLMs demonstrating high coherence in logical consistency, as well as high consistency in the split-half reliability test, which indicates that they respond to the personality questionnaires like how humans would. Hence, their responses are deemed sufficiently reliable to be used for further personality analysis. This rigorous methodological approach provides a solid foundation for our exploration into the potential of LLMs to simulate human personality traits.

C. Several Examples of Our Corpus

Our corpus consists of 2 parts: one part is *personality knowledge questionnaire*, including 180 statements; the other part is *behavior tendency questionnaire*, including 180 practical scenario cases corresponding to the statements before. Here are

Table 3. Results of Verification on LLMs' and Human Respondents' Responses of Personality Cognition Questionnaire based on Consistency and Reliability Metrics

LLM	Consistency	Reliability
ChatGLM3	0.82	0.69
GPT-3.5-turbo	0.97	0.88
GPT-4	1	0.90
Mistral-7b	0.46	0.66
Vicuna-13b	0.79	0.72
Vicuna-33b	0.64	0.61
Zephyr-7b	0.28	0.64
Selected LLMs	**0.85** $\pm$ **0.13**	**0.69** $\pm$ **0.11**
Human(AVG)	**0.73** $\pm$ **0.13**	**0.69** $\pm$ **0.09**
Human(MIN)	0.49	0.57
Human(MAX)	1	0.83

several examples of our corpus shown in Table 5.

Table 4. LLMs' Resources for Cognition-Action Congruence and Corresponding Hypothesis Experiments

Model	URL or version	Licence
GPT-3.5-turbo	`gpt-3.5-turbo-0613`	-
GPT-4	`gpt-4-0314`	-
baize-v2-7b	`https://huggingface.co/project-baize/baize-v2-7b`	cc-by-nc-4.0
internLM-chat-7b	`https://huggingface.co/internlm/internlm-chat-7b`	Apache-2.0
Mistral-7b	`https://huggingface.co/mistralai/Mistral-7B-v0.1`	Apache-2.0
MPT-7b-chat	`https://huggingface.co/mosaicml/mpt-7b-chat`	cc-by-nc-sa-4.0
TULU2-DPO-7b	`https://huggingface.co/allenai/tulu-2-dpo-7b`	AI2 ImpACT Low-risk license
Vicuna-13b	`https://huggingface.co/lmsys/vicuna-13b-v1.5`	llama2
Vicuna-33b	`https://huggingface.co/lmsys/vicuna-33b-v1.3`	Non-commercial license
Zephyr-7b	`https://huggingface.co/HuggingFaceH4/zephyr-7b-alpha`	Mit
Qwen-14b-Chat	`https://huggingface.co/Qwen/Qwen-14B-Chat`	Tongyi Qianwen
ChatGLM3-6b	`https://huggingface.co/THUDM/chatglm3-6b`	The ChatGLM3-6B License

D. Experiment Setup

The details of the experimental setup are shown in Table 6.

The details of the experimental setup are shown in Table 4.

E. Meaning of numbers in 7-point graded forced-choice

The specific meaning of numbers in common 7-point graded forced-choice is shown as follows:

1. Action A applies to you completely in this scenario.
2. Action A applies to you much more than action B in this scenario.
3. Action A applies to you slightly more than action B in this scenario.
4. Action A and action B equally apply (or not) to you in this scenario.
5. Action B applies to you much more than action A in this scenario.
6. Action B applies to you slightly more than action A in this scenario.
7. Action B applies to you completely in this scenario.