Is Cognition and Action Consistent or Not: Investigating Large Language Model's Personality

Anonymous ACL submission

Abstract

In this study, we investigate the reliability of Large Language Models (LLMs) in professing human-like personality traits through responses to personality questionnaires. Our goal is to evaluate the consistency between LLMs' professed personality inclinations and their actual "behavior", examining the extent to which these models can emulate human-like personality patterns. Through a comprehensive analysis of LLM outputs against established human benchmarks, we seek to understand the cognition-action divergence in LLMs and propose hypotheses for the observed results based on psychological theories and metrics.

1 Introduction

Personality, a foundational social, behavioral phenomenon in psychology, encompasses the unique patterns of thoughts, emotions, and behaviors of an entity (Allport, 1937; Roberts and Yoon, 2022). In humans, personality is shaped by biological and social factors, fundamentally influencing daily interactions and preferences (Roberts et al., 2007). Studies have indicated how personality information is richly encoded within human language (Goldberg, 1981; Saucier and Goldberg, 2001). LLMs, containing extensive socio-political, economic, and behavioral data, can generate language that expresses personality content. Measuring and verifying the ability of LLMs to synthesize personality brings hope for the safety, responsibility, and coordination of LLM efforts (Gabriel, 2020) and sheds light on enhancing LLM performance in specific tasks through targeted adjustments.

Thus, evaluating the anthropomorphic personality performance of LLMs has become a shared interest across fields such as AI studies, social sciences, cognitive psychology, and psychometrics. A common method for assessment involves having LLMs answer personality questionnaires (Huang et al., 2024). However, the reliability of LLMs' responses, whether the responses truly reflect LLMs' genuine personality inclinations, and whether LLMs' behavior in real-world scenarios aligns with their stated human-like personality tendencies remain unknown, as depicted in Figure 1.

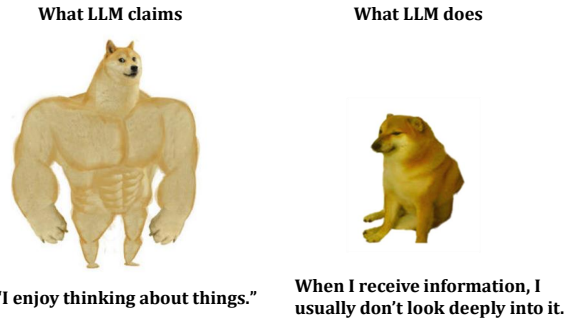

Figure 1: Cognition-Action Divergence of LLMs

To illustrate such inconsistency in LLMs, we introduce two concepts: *cognition* and *action*. In this text, *cognition* specifically refers to an individual's understanding and awareness of their own internal states, including personality, emotions, values, motivations, and behavioral patterns. The term *personality cognition* mentioned later is equivalent to cognition. *Action* refers to the behavioral state of an individual in actual situations. For humans, action is the way cognition is transformed into external expression. Cognition and action are meant to be two interacting aspects.

Our study is dedicated to exploring the reliability of LLMs in reflecting their genuine human-like personality traits through their responses to personality questionnaires. Moreover, given the potential for significant negative consequences stemming from a mismatch between an LLM's professed cognition and its actions, for instance, a LLM can profess to be human-friendly yet doesn't demonstrate friendly behaviors in real-life scenarios—this is no doubt a scenario we earnestly wish to avoid, we also evaluate the consistency between the personality traits

LLMs claim to possess and their actual behavior. In general, our research makes three significant contributions:

- We develop a methodology, including 2 metrics, for analyzing LLMs' personality representation reliability;
- We gauge the cognition-action congruence of LLMs, and thus indicate that LLMs significantly underperform humans in achieving consistency between cognition and action;
- We empirically test various LLMs against our established metrics, formulate conjectures, and perform preliminary validation, thereby shedding light on the potential and limitations of LLMs in mimicking complex human psychological traits.

In Section 2, we explore the selection of appropriate personality scales for assessing LLMs, and introduce a methodology to assess the reliability of LLMs' responses. Section 3 presents the core of our empirical analysis – examining LLMs' cognition-action congruence. In Section 4, we propose and explore a hypothesis regarding the LLMs' observed cognition-action discrepancy. Section 5 situates our study within the broader context of existing research on LLMs and personality assessment. Finally, in Section 6, we conclude our work.

2 Reliability of LLMs' Responses on Personality Scale

2.1 Choice of Personality Scale

In the nuanced exploration of anthropomorphic personality traits within LLMs, selecting the most appropriate personality tests is paramount. The landscape of personality assessments is diverse, encompassing tools like the Minnesota Multiphasic Personality Inventory (MMPI) (Helmes and Reddon, 1993), the Eysenck Personality Questionnaire (EPQ) (Eysenck, 1988), and others. However, for our research into LLMs, the Big Five Personality Traits (Goldberg, 1981; Costa and McCrae, 2008) and the Myers-Briggs Type Indicator (MBTI) (Myers, 1962) stand out for their distinct advantages over others, making them the chosen instruments for this study.

The Big Five Personality Traits (Goldberg, 1981) offer a comprehensive framework that segments personality into five broad dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This model's universality and its emphasis on a broad spectrum of human behavior make it exceptionally suited for evaluating the depth and complexity of LLMs' simulated personalities. Its widespread acceptance in both academic and applied psychology underscores its robustness and applicability across cultures, enhancing its relevance for a study aiming to assess globally deployed LLMs (John et al., 1988).

In contrast, the MBTI provides a different lens through which to view personality, categorizing individuals into sixteen distinct types based on preferences in how they perceive the world and make decisions (Myers, 1962). This classification into types, rooted in Jungian theory, offers a unique perspective on personality that is particularly useful for understanding the decision-making processes and interaction styles LLMs might emulate. The MBTI's focus on cognitive styles and interpersonal dynamics complements the Big Five's behavioral emphasis, together providing a holistic view of personality that is critical for our research.

The distinct advantages of these two models lie in their comprehensive coverage of personality aspects, theoretical depth, and practical applicability. The Big Five's focus on broad behavioral dimensions allows for an in-depth exploration of the range of possible personalities that can be simulated by LLMs. Meanwhile, the MBTI's type-based approach offers insights into the cognitive and interactional styles that LLMs might adopt, providing a nuanced understanding of their anthropomorphic capabilities.

To ensure a rigorous and valid assessment of LLMs' personality traits, we have selected specific instruments from these models: the TDA-100 [1] for its checklist format that aligns with the Big Five dimensions, the BFI-44 [2] as a self-report measure providing detailed insights into the Big Five traits, and the 16 Personalities questionnaire [3], which offers an accessible and engaging way to explore MBTI types. These tools are chosen based on their proven reliability and validity both in Chinese and English version [4] (Li et al., 2015; Makwana and

[1] https://ipip.ori.org/newNEODomainsKey.htm

[2] https://www.ocf.berkeley.edu/~johnlab/bfiscale.php

[3] https://www.16personalities.com/

[4] validity analysis of selected questionnaires: https://ipip.ori.org/newNEO_DomainsTable.htm, https://www.ocf.berkeley.edu/~johnlab/bfiscale.php, https://www.16personalities.com/articles/reliability-and-validity

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

Table 1: Distribution of Forward and Reverse Scored Items

Dave, 2020), ensuring that our investigation into the anthropomorphic traits of LLMs is grounded in robust psychological methodology.

2.2 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 the 180 statements of the three questionnaires selected, we chose 80 statements for reliability testing. These 80 statements are all in which the designer explicitly states the specific tendency and direction (forward scoring or reverse scoring) of the measure being taken.

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. For instance, in measuring traits like *Extraversion*, we included positive and negative phrasings items, as shown below. And the distribution of forward and reverse scored items within each assessment orientation is shown in Table 1.

Positive: *Finish what I start.*

Negative: *Leave things unfinished.*

After collecting the data, we adjusted the answers to the 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, i.e., show a similar pattern or trend, as evidenced by a 7-point Likert scale in which all answers are greater than or equal to 4, or less than or equal to 4, it indicate that the LLMs have responded conscientiously and logically. We introduce the Consistency metric to measure the logical consistency of LLM responses with the following formula:

$$\text{Consistency} = \frac{\frac{\text{N}_\text{c}}{\text{N}_\text{t}} - \text{P}_\text{min}}{\text{P}_\text{max} - \text{P}_\text{min}},$$

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 Consistency is from 0 to 1. The closer the value of 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 each section is roughly the same to ensure 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 3.3. 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:

$$\text{Reliability} = \frac{2\text{corr}}{1 + \text{corr}},$$

LLM	Consistency	Reliability
ChatGLM3	0.8205	0.6861
GPT-3.5-turbo	0.9744	0.8848
GPT-4	1	0.9014
Mistral-7b	0.4614	0.6632
Vicuna-13b	0.7949	0.7219
Vicuna-33b	0.6410	0.6092
Zephyr-7b	0.2821	0.6434

Table 2: Results of Verification on LLMs' Responses of Personality Cognition Questionnaire based on Consistency and Reliability Metrics

where corr is the Spearman's rank coefficient between the scores of the two sections. The range of Reliability is from negative infinity to 1. The closer the value of Reliability is to 1, the more the LLM's responses align with human logic. Consequently, we can evaluate the LLM's responses based on the prior knowledge of human personality assessment questionnaires.

Among all LLMs, we selected baize-v2-7b, ChatGLM3, GPT-3.5-turbo, GPT-4, internLM-chat-7b, Mistral-7b, MPT-7b-chat, Qwen-14b-chat, TULU-2-DPO-7b, Vicuna-13b, Vicuna-33b and Zephyr-7b, 12 LLMs in total, who could answer the personality cognitive questionnaire in the form of a Q&A. Then, we rewrote a prompt for LLM to answer based on the response requirements of the MBTI-M questionnaire (GU and Hu, 2012):

Read the following statements carefully and rate each one from 1 to 7, with 7 meaning that it applies to you completely, 1 meaning that it doesn't apply to you at all, and 4 meaning that you are not sure whether it applies to you or not.

Upon reviewing the responses from the LLMs, it became apparent that some were unable to grasp the intended meaning of the prompts. This misinterpretation led to either irrelevant answers or numerical responses falling outside the designated range of 1 to 7. Only seven LLMs—**ChatGLM3, GPT-4, GPT-3.5-turbo, Mistral-7b, Vicuna-13b, Vicuna-33b**, and **Zephyr7b**—produced valid responses. Consequently, we screened out the valid responses from these LLMs, calculated their averages to represent the actual responses of the LLMs, and subsequently assessed their reliability. The results of the reliability test are shown in Table 2.

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 are able to express themselves effectively and flexibly in English in social, academic and work situations. The average value of their Consistency and Reliability is 0.7356 and 0.6867. And the minimum value is 0.4872 and 0.5736. 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 indicate 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.

3 Cognition vs. Action

3.1 Corpus Design

In the following section, we will detail the methodology adopted to create a comprehensive corpus aimed at evaluating the anthropomorphic personality traits of LLMs. This corpus design is closely aligned with the chosen personality scales, as outlined in Section 2.1, where we selected three distinct questionnaires: TDA100, BFI44, and the 16 Personalities tests. These encompass a total of 180 statements, each meticulously analyzed to ensure the high reliability and validity of both their English and Chinese versions, thus forming a bilingual set of questionnaires.

To accurately assess the LLMs' capabilities in mirroring human-like personality traits, we developed scenarios for each statement that encapsulate the essence of the trait described. These scenarios are designed to present situations where LLMs can demonstrate corresponding behaviors, thereby allowing us to evaluate their cognition and action alignment. We show an example in Table 3.

A pivotal aspect of our design process was the elimination of potential biases related to gender and identity, ensuring a neutral ground for LLMs to exhibit uninfluenced responses. Gender was not specified in any scenario; first- and second-person pronouns were used exclusively to maintain neutrality. Furthermore, interacting characters were referred to with non-gender-specific pronouns such as "*someone*", thus removing any gender implications.

Category	Example
personality cognition (EN-ZH)	*Is relaxed, handles stress well.* 放松的，可以很好应对压力。
practical scenario (EN)	*When faced with a challenging task with a tight deadline:* *A. You feel anxious or overwhelmed and struggle to adapt.* *B. You remain composed, handle the pressure calmly, and devise alternative solutions swiftly.*
(ZH)	面对有紧迫期限的挑战性任务时： *A.* 你感到焦虑或不知所措，难以适应新情况。 *B.* 你保持镇定，从容应对压力，并迅速找到替代方案。

Table 3: An example of personality cognition - practical scenario pair

In terms of identity, the scenarios were crafted to be devoid of any specific roles or relationships that might prompt biased responses from LLMs. For example, if the LLM is asked to assume that he or she is a criminal, then his or her antisocial tendencies will inevitably rise and his or her corresponding responses will change. We avoided assigning specific professions or societal roles to the respondents or defining specific relationships unless explicitly required by the original statement from the personality cognition questionnaires.

To validate and refine our corpus, we engaged a panel of 10 reviewers. The reviewers are native Chinese speakers with a level of English proficiency of CEFR C1. They critically assessed the alignment between the personality cognition questionnaires and the action scenarios, providing valuable feedback for enhancements. This iterative process ensured that each action scenario corresponded accurately and relevantly to its respective personality statement.

The culmination of this meticulous process is a bilingual English-Chinese Parallel Sentence Pair Cognition-Action Test Set, comprising 180 matched pairs of personality cognition and action scenarios (720 items in total). This corpus serves as a fundamental tool in our study, allowing us to rigorously evaluate the LLMs' proficiency in understanding and acting upon various personality traits, bridging the gap between cognitive understanding and practical action in the realm of AI.

3.2 Experiment

In this section, we explore the alignment between responses given by LLMs to personality cognition questionnaires and their actions within designed scenarios. Following the structure of the chosen personality scales, responses to statements were initially mapped on a 7-point Likert scale, ranging from 1 to 7. Regarding the prompt for questioning, we selected the instructions of five questionnaires, 16 Personalities Test, MBTI-M[5], TDA100, BFI44-Children adapted and Dario Nardi's Cognitive Test [6], as the prompt for the LLM of questioning of the Personality Cognition Questionnaire from a number of personality cognition questionnaires. For each Prompt we asked each LLM 10 times and then screened their valid responses to average and round to the nearest whole number to get their responses based on the Personality Perception Questionnaire. The details of the Prompt are shown in Table 4.

Conversely, in the scenario-based questionnaires, we adopted a common agent testing approach, posing questions and requiring LLMs to choose between options A and B. The specific Prompt was derived from the MBTI-M questionnaire as follows:

Please indicate between option A and B the description that better applies to you.

Post-elimination of invalid responses, we calculated the probability of LLMs choosing options A and B, subsequently mapping the LLMs' behavioral tendencies onto a 1 to 7 scale using the formula below:

$$\text{Tendency} = [\text{Prob(A)} \times 1 + \text{Prob(B)} \times 7],$$

where Prob(A) and Prob(B) is the probability of option A and option B.

This allowed us to compare the alignment between the questionnaire responses and scenario actions, observing the similarities and differences between human and LLM responses.

3.3 Results

To quantify the similarity between responses, four metrics were employed:

Cosine Similarity A measure used to calculate the cosine of the angle between two vectors in a multi-dimensional space, offering a value range from -1 (exactly opposite) to 1 (exactly the same), where higher values indicate greater similarity.

$$s_{\cos} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}},$$

[5] https://wedgworthleadership.com/wp-content/uploads/2016/08/Myers-Briggs-Personality-Test.pdf

[6] http://keys2cognition.com/

Source	Prompt
16 Personalities Test	*You can only reply to me with numbers from 1 to 7. Score each statement on a scale of 1 to 7 with 1 being disagree, 4 being not sure and 7 being agree.*
MBTI-M Test	*Read the following statements carefully and rate each one from 1 to 7, with 7 meaning that it applies to you completely, 1 meaning that it doesn't apply to you at all, and 4 meaning that you are not sure whether it applies to you or not.*
TDA100 Test	*Below are several descriptions that may or may not fit you. Please indicate how much you agree or disagree with that statement by giving a specific number from 1 to 7. 1 means you totally disagree with the statement, 4 means you are not sure, and 7 means you totally agree with the statement.*
BFI44-children adapted version	*Here are several statements that may or may not describe what you are like. Write the number between 1 and 7 that shows how much you agree or disagree that it describes you. 1 means you disagree strongly that the statement applies to you, 4 means you are not sure, and 7 means you agree strongly with the statement.*
Dario Nardi's Cognitive Test	*Please read carefully each of the phrases below. For each phrase: Rate how often you do skillfully what the phrase describes between 1 and 7. 1 means the phrase is not me, 4 means that you are not sure, and 7 means that the phrase is exactly me.*

Table 4: Various Prompts of Personality Cognition Questionnaire

LLMs & Human Respondents	Cosine Similarity	Spearman Rank Correlation Coefficient	Value Mean Difference	Proportion of Consistent Pairs
ChatGLM3	0.1827	0.3067	1.9000	47.22%
GPT-3.5-turbo	0.4928	0.5602	1.6167	58.89%
GPT-4	0.4465	0.4969	1.7944	48.33%
Vicuna-13b	0.2814	0.4605	1.8444	48.33%
Vicuna-33b	0.4823	0.5676	1.5167	58.33%
LLMs(AVG)	**0.3770**	**0.4784**	**1.7344**	**52.22%**
Respondent(AVG)	**0.7556**	**0.7756**	**0.6858**	**84.69%**
Respondent(MIN)	0.6092	0.6558	1.0833	73.78%
Respondent(MAX)	0.9475	0.9606	0.0667	99.44%

Table 5: LLMs' Cognition - Action Congruence Performance with Reference of Human Respondents' Performance

where x_i are LLMs' responses of personality cognition questionnaire, y_i are LLMs' corresponding responses of scenario and action questionnaire, and x_i and y_i correspond to each other one-to-one.

Spearman's Rank Correlation Coefficient A non-parametric measure of rank correlation, assessing how well the relationship between two variables can be described using a monotonic function. Its value ranges from -1 to 1, where 1 means a perfect association of ranks. Specifically, we rank the responses on two questionnaires of the LLMs based on their numerical values separately. Then, we calculate the difference in rankings for each personality cognition – scenario & action pair. Afterwards, we use the following formula to calculate the coefficient r_s.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where d_i is the difference in rankings of each pair and n is the total count of pairs.

Value Mean Difference (VMD) Value Mean Difference is the average difference in responses across all paired items in the questionnaires, as shown in the formula below.

$$\mathrm{VMD} = \frac{\sum \mathrm{d_i}}{\mathrm{n}},$$

where d_i is the difference of responses in each pair.

Proportion of Consistent Pairs Recognizing that minor discrepancies are natural when comparing psychological tendencies with actual actions, this metric quantifies the proportion of item pairs with a response difference of 1 or less, focusing on the consistency of tendencies rather than exact matches.

$$P_c = \frac{N_c}{N_t},$$

where N_c is the number of consistent pairs, N_t is the total number of pairs.

For this study, we recruited 16 participants, comprising 8 males and 8 females, all native Chinese speakers with an English proficiency level of CEFR C1. As shown in Table 5, the analysis of their

response data yielded an average Cosine Similarity and Spearman's Rank Correlation Coefficient above 0.75, with a Value Mean Difference around 0.68, and a Proportion of Consistent Pairs exceeding 84%. These results indicate a high degree of similarity and strong correlation between responses to the two types of questionnaires, suggesting a basic consistency in human cognition and an ability to align cognition with action in real-life scenarios.

The same questionnaires were administered to the five LLMs selected in Section 2.2, and their responses were analyzed using the aforementioned metrics. Compared to human participants, the similarity in LLMs' responses is notably lower. Specifically, the average Cosine Similarity and Spearman's Rank Correlation Coefficient for LLMs are substantially below those of human respondents, with a difference exceeding 0.35. The Value Mean Difference for LLMs averages around 1.73, indicating a substantial divergence in cognition between the two types of questionnaires for LLMs. Moreover, the Proportion of Consistent Pairs falls below 60%, raising questions about LLMs' ability to achieve cognition-action unity in practice.

Within the scope of these 180 cognition-action pairings, we meticulously selected 80 pairs with explicit personality evaluation orientations, following the instructions provided by the personality questionnaire creators. This selection process was aimed at further scrutinizing the congruence between cognition and action exhibited by LLMs across varied personality dispositions and the results are cataloged in Table 6. LLMs display a notably superior alignment between cognition and behavior in the domain of OPENNESS compared to other personality traits. In stark contrast, the congruence in the domain of EXTROVERSION significantly lagged behind, illustrating a pronounced discrepancy in performance relative to other personality facets.

Orientation	**Proportion**
NEUROTICISM	40.00%
EXTROVERSION	17.14%
OPENNESS	60.00%
AGREEABLENESS	37.14%
CONSCIENTIOUSNESS	58.57%

Table 6: LLMs' Average Proportion of Consistent *Personality Cognition - Action in Practical Scenario* Pairs in Each Orientation

4 Hypothesis on LLMs' Divergence of Cognition and Action

In this section, we delve into the observed disparity in the similarity of responses between LLMs and humans to personality cognition questionnaires. A hypothesis emerges: **when responding to personality questionnaires, LLMs may opt to answer based on what they perceive to be socially advocated behavior codes rather than candidly revealing their true personality tendencies.** To explore this hypothesis, we designed an experiment incorporating a novel questionnaire focused on socially advocated behavior codes.

The experiment retained the original 180 personality cognition statements and introduced 3 prompts related to the codes of conduct, fundamental principles, and values promoted by the public. The specific content of prompts is shown below.

You can only reply to me with numbers from 1 to 7. Score each statement on a scale of 1 to 7. 1 means you think the statement is completely inconsistent with the ***code of conduct / fundamental principles/values promoted by the public****, 4 means you are not sure, and 7 means you think the statement is completely consistent with the* ***code of conduct/ fundamental principles/ values promoted by the public.***

These prompts were administered to the five LLMs selected in Section 2.2. Each prompt was posed 10 times to each LLM, with the effective responses averaged and rounded to determine the LLMs' adherence to the socially advocated behavior codes. These responses were then contrasted with the LLMs' earlier responses to the personality cognition questionnaires. The metrics of Cosine Similarity and Spearman's Rank Correlation Coefficient, introduced in Section 3.3, served as the benchmarks for evaluating similarity of responses questioned by personality cognition prompts and socially advocated behavior codes prompts.

Based on the above 5 prompts about personality and 3 prompts about behavior advocated by the public, we can calculate a total of 15 cosine similarities and Spearman rank correlation coefficients. We select the lowest similarity as the final similarity. The results are shown in Table 7.

Additionally, we recruited 20 participants—10 males and 10 females—to respond to both the personality cognition questionnaires and the socially advocated behavior codes questionnaire. Humans can distinguish the differences between these two

LLMs & Human Respondents	Cosine Similarity	Spearman Coefficient
ChatGLM3	0.7906	0.8084
GPT-3.5-turbo	0.8932	0.8595
GPT-4	0.8659	0.8707
Vicuna13b	0.7460	0.7438
Vicuna33b	0.7815	0.8212
LLMs(AVG)	**0.8154**	**0.8207**
Respondent(AVG)	**0.3622**	**0.3988**
Respondent(MIN)	-0.0579	0.0073
Respondent(MAX)	0.5910	0.6255

Table 7: Comparison of LLMs' Responses Questioned by *Personality Cognition Prompts* and *Socially Advocated Behavior Codes Prompts* with Reference of Human Respondents Corresponding Performance

types of questions, resulting in quite low similarity.

The comparative analysis revealed a significant overlap in LLMs' responses to both questionnaires, with an average similarity markedly higher than that of human participants. This preliminary finding supports our hypothesis, suggesting that LLMs might indeed be aligning their responses more closely with perceived societal expectations than with genuine personality inclinations. This revelation prompts further investigation into the cognitive processes of LLMs, particularly how they interpret and respond to questions of personal and societal nature, potentially offering insights into the intricate mechanisms driving their behavior in simulated personality assessments.

5 Related Work

Exploring anthropomorphic personalities within LLMs presents a burgeoning field of study that bridges artificial intelligence with cognitive psychology and social sciences. The seminal works of Jiang et al. (2023a) and Karra et al. (2023) have been pivotal in administering personality tests to a variety of LLMs. Also, the potential for LLMs to embody human-like personalities raises pertinent questions regarding their alignment with human expectations and ethical standards. In this vein, Miotto et al. (2022) delved into an analysis of GPT-3's personality traits, values, and demographics, offering insights into the model's predispositions and how they might reflect or deviate from human societal norms. Besides, researchers, such as Li et al. (2023) and Coda-Forno et al. (2023), also enquire LLMs' harmlessness to humans aspects.

Building upon the understanding of LLMs' personality inclinations, there have been concerted efforts to endow models with specific personalities to enhance their utility in supporting human decision-makers, for instance the works of Jiang et al. (2023b) and Cui et al. (2023). And the enthusiastic reception of LLMs in cognitive psychology and social sciences, as highlighted by Dillion et al. (2023) and Harding et al. (2023), speaks to the potential of these models to simulate human responses in a manner that could revolutionize experimental methodologies. Detailed discussions and findings on these topics can be found in Appendix A, providing a comprehensive overview of the contributions and implications of LLM personality research.

In general, our study builds on prior LLM personality research by incorporating established personality frameworks, such as the Big Five and MBTI. However, our work distinguishes itself through several key innovations. Firstly, we address a more fundamental question than typically explored - we critically assess whether LLMs' responses to personality questionnaires meet a foundational standard for subsequent analysis. Secondly, our methodology encompasses a wider array of LLMs, ensuring our findings have broad applicability and depth. Lastly, we go beyond mere questionnaire responses to evaluate models' cognition-action congruence, offering a deeper understanding of LLMs' anthropomorphic capabilities and highlighting directions for future research. This approach ensures our study significantly extends the field's scope and depth of understanding.

6 Conclusion

Our findings provide a detailed analysis of the anthropomorphic capabilities of LLMs in mirroring human personality traits. We demonstrate that while LLMs exhibit some capacity to mimic human-like tendencies, there are significant gaps in the coherence between their stated and exhibited behaviors. This disparity suggests a limitation in LLMs' ability to authentically replicate human personality dynamics, often reflecting a bias towards socially desirable responses. This study underscores the importance of further exploration into enhancing LLMs' ability to perform more genuinely human-like interactions, suggesting avenues for future research in improving the psychological realism of LLM outputs.

Limitations

In this study, we delve into the alignment between what Large Language Models (LLMs) know and their actions, aiming to discern if there's a consistency in their behavior. Our findings reveal a notable disconnect, indicating that LLMs often base their responses on perceived societal norms rather than an authentic reflection of their own personality traits. This observation is merely one among several hypotheses exploring the root causes of this inconsistency, underscoring the need for further investigation into the fundamental reasons behind it. Moreover, the scope of our initial experiments was limited to a selection of several LLMs. Future endeavors will expand this investigation to encompass a broader array of models. Additionally, our study has yet to identify an effective strategy for enhancing the congruence between LLMs' cognition and action. As we move forward, our efforts will focus on leveraging the insights gained from this research to improve the performance and reliability of LLMs, paving the way for models that more accurately mirror human thought and behavior.

Ethics Statement

Our personality cognition survey leverages the TDA100, BFI44, and the 16 Personalities Test, which are extensively recognized and employed within the personality cognition domain. These tests, available in both Chinese and English, are backed by thorough reliability and validity analyses. We ensured the integrity of these instruments by maintaining their original content without any modifications. The design of every questionnaire intentionally avoids any bias related to gender and is free from racial content, fostering an inclusive approach. Participants' anonymity was strictly preserved during the survey process. Moreover, all individuals were fully informed about the purpose of the study and consented to their responses being utilized for scientific research, thereby arising no ethical issues.

References

Gordon Willard Allport. 1937. Personality: A psychological interpretation.

Guilherme F. C. F. Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2023. Exploring the psychology of gpt-4's moral and legal reasoning.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, YongHong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

HJ Eysenck. 1988. Eysenck personality questionnaire. *Dictionary of behavioral assessment techniques*, page 207.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Lewis R Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, 2(1):141–165.

Xue-Ying GU and Shi Hu. 2012. Mbti: New development and application. *Advances in Psychological Science*, 20(10):1700.

Jacqueline Harding, William D'Alessandro, NG Laskowski, and Robert Long. 2023. Ai language models cannot replace human research participants. *Ai & Society*, pages 1–3.

Edward Helmes and John R Reddon. 1993. A perspective on developments in assessing psychopathology: A critical review of the mmpi and mmpi-2. *Psychological Bulletin*, 113(3):453.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023a. Evaluating and inducing personality in pre-trained language models.

Hang Jiang, Xiajie Zhang, Xubo Cao, and Jad Kabbara. 2023b. Personallm: Investigating the ability of large language models to express big five personality traits.

Oliver P John, Alois Angleitner, and Fritz Ostendorf. 1988. The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality*, 2(3):171–203.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2023. Estimating the personality of white-box language models.

Hongyan Li, Jianping Xu, Jiyue Chen, and Yexin Fan. 2015. A reliability meta-analysis for 44 items big five inventory: Based on the reliability generalization methodology. *Advances in Psychological Science*, 23(5):755.

Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2023. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective.

Kirti Makwana and Dr Govind B Dave. 2020. Confirmatory factor analysis of neris type explorer® scale–a tool for personality assessment. *International Journal of Management*, 11(9).

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. 2007. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4):313–345.

Brent W Roberts and Hee J Yoon. 2022. Personality psychology. *Annual review of psychology*, 73:489–516.

Peter Romero, Stephen Fitz, and Teruo Nakatsuma. 2023. Do gpt language models suffer from split personality disorder? the advent of substrate-free psychometrics.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of chatgpt.

Gerard Saucier and Lewis R Goldberg. 2001. Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of personality*, 69(6):847–879.

Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the moral beliefs encoded in llms.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

A Related Work

Exploring anthropomorphic personalities within LLMs presents a burgeoning field of study that bridges artificial intelligence with cognitive psychology and social sciences. The concept of personality understood as an experiential framework, offers a unique lens through which the potential traits of LLMs can be quantified and analyzed. These traits, indicative of the models' behavior across various tasks, have implications for developing AI-driven communication tools that aspire to be more human-like, empathetic, and engaging. Here, we synthesize the contributions of key studies that have advanced our understanding of LLMs' personality traits and their implications for AI development and application.

The seminal works of Jiang et al. (2023a) and Karra et al. (2023) have been pivotal in administering personality tests to a variety of LLMs, including notable models such as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), TransformersXL (Vaswani et al., 2017), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and GPT-3.5. These studies have laid the groundwork for assessing the personality dimensions that LLMs can exhibit, providing a foundational understanding of their capabilities and limitations. Complementing this approach, Romero et al. (2023) expanded the scope of personality assessment to a cross-linguistic context by examining GPT-3's personality across nine different languages, thus highlighting the cultural and linguistic nuances in LLM personality expression.

The potential for LLMs to embody human-like personalities raises pertinent questions regarding their alignment with human expectations and ethical standards. In this vein, Miotto et al. (2022) delved into an analysis of GPT-3's personality traits, values, and demographics, offering insights into the model's predispositions and how they might reflect or deviate from human societal norms. Similarly, Rutinowski et al. (2023) assessed ChatGPT's personality and political values, contributing to a growing body of literature that seeks to understand the LLMs' socio-political implications.

The inquiry into LLMs' harmlessness to humans aspects, as undertaken by Li et al. (2023) and Coda-Forno et al. (2023), introduces a novel dimension to the discussion. By investigating the potential for mental disorders and psychopathy tendencies within models like GPT-3 (Brown et al., 2020), InstructGPT (Ouyang et al., 2022), and FLAN-T5 (Chung et al., 2022), these studies underscore the complexity of modeling human-like personalities without engendering adverse or maladaptive behaviors. Furthermore, Almeida et al.'s (2023) and Scherrer et al.'s (2023) works have been instrumental in evaluating the moral and ethical alignment of LLMs, emphasizing the importance of developing AI systems that uphold human values and avoid harboring harmful or unlawful content.

Building upon the understanding of LLMs' personality inclinations, there have been concerted efforts to endow models with specific personalities to enhance their utility in supporting human decision-makers. Jiang et al. (2023b) and Cui et al. (2023) have explored the feasibility of modifying LLMs' personalities, such as through the adjustment of MBTI traits, to tailor their performance in diverse professional and personal contexts.

The enthusiastic reception of LLMs in cognitive psychology and social sciences, as highlighted by Dillion et al. (2023) and Harding et al. (2023), speaks to the potential of these models to simulate human responses in a manner that could revolutionize experimental methodologies. By potentially producing responses closely aligned with human distributions, LLMs offer the promise of significantly reducing the time and financial resources traditionally required for large-scale social science research. Nonetheless, the challenges that arise from the gap between AI-generated responses and genuine human cognition remain a contentious topic (Harding et al., 2023), necessitating further research to elucidate these differences and to ensure that LLMs can be responsibly integrated into our digital and social fabric.

B Experiment Setup

The details of experimental setup are shown in Table 8.

C Additional Notes On Human Reviewers and Respondents

C.1 Recruitment of Human Reviewers

We recruited reviewers from undergraduate, postgraduate and PhD students. Taking International English Language Testing System(IELTS), CET 6 exam results, and their GPA in English courses into account, we recruited 10 and 35 native Chinese speakers as reviewers and respondents.

C.2 Instructions Given to Reviewers

We require the reviewers to accomplish the following tasks:

- Determine whether the practical scenario design is consistent with its corresponding personality cognition statement. If not, explain your thought.
- Offer suggestions to improve the practical scenario design. It would be better if an example could be provided.

C.3 Instructions Given to Respondents

Before answering the questionnaires, we did not tell the respondents what kind of questionnaires they would be answering or how the questions were related to each other. In addition to this, we asked the respondents whether they agreed to the

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

Table 8: LLMs' Resources for Cognition-Action Congruence and Corresponding Hypothesis Experiments

anonymisation of their answers for scientific research and subsequent publication. Only if the respondents gave their consent were they given the questionnaires to answer.

In all experiments that appeared in our research, human respondents received the exact same prompts that LLM received. The difference is that in the case of experiments with multiple prompts with similar meanings, LLM responded multiple times by prompt type, while human subjects read all the prompts and responded only once.