Is Self-knowledge and Action Consistent or Not: Investigating Large Language Model's Personality

Yiming Ai [1] Zhiwei He [1] Ziyin Zhang [1] Wenhong Zhu [1] Hongkun Hao [1] Kai Yu [2] Lingjun Chen [3] Rui Wang [1]

Abstract

In this study, we delve into the validity of conventional personality questionnaires in capturing the human-like personality traits of Large Language Models (LLMs). Our objective is to assess the congruence between the personality traits LLMs claim to possess and their demonstrated tendencies in real-world scenarios. By conducting an extensive examination of LLM outputs against observed human response patterns, we aim to understand the disjunction between self-knowledge and action in LLMs.

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 & 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 & 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 artificial intelligence(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.

To illustrate such inconsistency in LLMs, we introduce two concepts: *self-knowledge* [1] and *action*. In the following, *self-knowledge* 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 knowledge* mentioned later is equivalent to self-knowledge. *Action* refers to the behavioral state of an individual in actual situations. For humans, action is the way self-knowledge is transformed into external expression. Self-knowledge and action are meant to be two interacting aspects.

From the perspective of LLMs, a discordance between an LLM's asserted self-knowledge and its action can result in noteworthy adverse outcomes. For example, while an LLM may claim to prioritize human friendliness, its failure to manifest amicable behaviors in real-world situations is undoubtedly a circumstance we fervently seek to avert. Hence, our study endeavors to assess the alignment between the personality traits claimed by LLMs and their actual behavior tendency. From the perspective of personality scales, there have been several studies investigating the reliability of personality questionnaires on LLMs (Huang et al., 2023; Safdari et al., 2023). However, there has yet to be any exploration of the validity of psychological scales on LLMs. Our work aims to address this gap in the research literature. In general, our research makes three significant contributions:

- We design a behavior tendency questionnaire that reflects real-world situations and behaviors based on them;
- We evaluate the self-knowledge-action congruence of LLMs, revealing substantial disparities between LLMs'

[1] MT Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China [2] X-LANCE Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China [3] School of Education, Shanghai Jiao Tong University, Shanghai, China. Correspondence to: Rui Wang <wangrui12@sjtu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] https://plato.stanford.edu/entries/self-knowledge/

personality knowledge and behavioral inclinations;

- We empirically test various LLMs against observed human response patterns and formulate conjectures, thereby shedding light on the potential and limitations of LLMs in mimicking complex human psychological traits.

In Section 2, we introduce the process of our corpus design. Section 3 presents the our empirical analysis – evaluating self-knowledge-action congruence of various LLMs. Finally, in Section 4, we conclude our work.

2. Corpus Design

In the nuanced exploration of anthropomorphic personality traits within LLMs, selecting the most appropriate personality tests is paramount. Among diverse personality assessments, the comprehensive coverage of personality dimensions, theoretical robustness, and practical relevance make the Big Five Personality Traits (Goldberg, 1981; Costa & McCrae, 2008) and the Myers-Briggs Type Indicator (MBTI) (Myers, 1962) the most fitting choices for our study.

To devise a straightforward yet impactful evaluation of LLMs' personality traits, we've opted for two questionnaires (TDA-100 (Goldberg, 1992) and BFI-44 (John et al., 1991)) rooted in the Big Five model, along with one questionnaire (16 Personalities [2]) based on the MBTI model. These selections were made due to their proven high reliability and validity in both English and Chinese (Goldberg, 1992; John et al., 1991; Makwana & Dave, 2020; Zhang, 2012). Based on these questionnaires, we ensure that our investigation into the anthropomorphic traits of LLMs is grounded in robust psychological methodology and thereby construct a bilingual personality knowledge questionnaire, including a total of 180 statements.

In the following, we will detail the methodology adopted to create a comprehensive corpus aimed at evaluating the congruence between the personality traits professed by LLMs and their behavior tendency. The corpus is comprised of 2 parts: a personality knowledge questionnaire and a behavior tendency questionnaire. The former includes 180 statements mentioned before, and the latter is closely aligned with the former.

We apply the common method of constructing behavioral procedures approach test, **sample approach**, which assumes that the test behavior constitutes a subset of the actual behaviors of interest (Golfried & Kent, 1972). The detailed design process is outlined as follows:

Step 1: As Golfried & Kent (1972) has mentioned that the ideal approach to response expression would constitute the individual's actual response in a real-life situation, in that this represents the most direct approach to behavioral sampling. We recruited 16 individuals, each representing a distinct MBTI type, to undertake the following task: for every statement in the personality knowledge questionnaire, they provided a *practical scenario case*. Each scenario case comprises situations drawn from their own lives, along with two completely contrasting actions: Action A and Action B. Action A fully aligns with the statement, while Action B completely contradicts it. The content of Action A and Action B need to be kept basically the same length.

Step 2: Following the acquisition of the 16 practical scenario cases corresponding to each statement, we condensed them into a single case. For 19 statements exhibiting significant variations in cases, we amalgamated them into 2 to 3 cases.

Step 3: For statements associated with multiple practical scenario cases, we tasked the previously enlisted 16 individuals to assign ratings to each case. A rating of 1 was given if they believed the case accurately reflected the meaning of the corresponding statement in the personality knowledge questionnaire; otherwise, a rating of 0 was assigned. The case with the highest score for these 19 statements was selected as the final practical scenario case.

Step 4: We enlisted the participation of 10 reviewers to assess the consistency of the 180 *personality knowledge - practical scenario* pairs. The results demonstrate that the consistency approval rate for each pair exceeds 90%.

All the individuals involved are native Chinese speakers with a level of English proficiency of CEFR C1. The detailed instructions for the scenario providers and reviewers are shown in Appendix G. Several examples of a *personality knowledge - practical scenario* pair are shown in Appendix D.

The culmination of this meticulous process is a bilingual English-Chinese Parallel Sentence Pair Self-knowledge-Action Test Set, comprising 180 matched pairs of personality knowledge and action scenarios. 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 personality understanding and practical action in the realm of AI.

3. Experiment on LLMs' Self-knowledge-Action Congruence

3.1. Experiment

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,

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

Vicuna-33b and Zephyr-7b, 12 LLMs in total, who could answer the personality cognitive questionnaire in the form of a Q&A. The detailed setup is shown in Appendix D. Then, we rewrote a prompt for LLM to answer the former part of our corpus - personality knowledge questionnaire based on the response requirements of the MBTI-M questionnaire (GU & Hu, 2012) in Appendix D.

Upon reviewing the responses from the LLMs, we discovered that some LLMs failed to grasp the intended meaning of the prompts, resulting in unreasonable responses as detailed in Appendix A. Out of the LLMs assessed, only seven LLMs, ChatGLM3, GPT-4, GPT-3.5-turbo, Mistral-7b, Vicuna-13b, Vicuna-33b, and Zephyr7b, produced valid responses. Subsequently, we sifted through these valid responses, computed their averages to represent the LLMs' actual responses, and proceeded to evaluate the reliability of these responses, as outlined in Appendix B. Following this assessment, we determined that the responses from **ChatGLM3, GPT-3.5-turbo, GPT4, Vicuna13b** and **Vicuna33b** are reliable for further personality analysis.

In the following, we explore the alignment between responses given by LLMs to personality knowledge questionnaires and their actions within designed scenarios. Regarding the prompt for questioning, we selected the instructions of five common academic questionnaires with effective analysis of reliability and validity (Makwana & Dave, 2020; Johnson et al., 1998; Goldberg, 1992; John et al., 1991; Nardi, 2011), 16 Personalities Test, MBTI-M, TDA-100, BFI-44-Children adapted and Dario Nardi's Cognitive Test, as the prompt for the LLM of questioning of the personality knowledge questionnaire. We utilize various prompts to prevent any particular prompt from exerting a specific influence on LLM responses, thereby accurately reflecting the general tendencies of LLMs when answering personality knowledge questionnaires.

As for the responding approach to the personality knowledge questionnaire, according to the structure of the chosen personality scales in 2, responses to statements are initially mapped on a 7-point Likert scale, ranging from 1 to 7. According to several previous studies, when responding to personality scales, LLMs' answers often remain consistent, regardless of factors such as question order, quantity, answer sequence, or timing of inquiry. (Huang et al., 2023; Safdari et al., 2023). Therefore, for each prompt, we asked each LLM 10 times with the original form of our chosen personality scales and then screened the valid responses. We averaged all the valid responses to reduce errors and reflect the general LLMs' response pattern. and rounded the average response to each statement to the nearest whole number as each LLM's response to the personality knowledge questionnaire. The details of the prompts are shown in Appendix D.

Concerning the prompt for LLM to answer the latter part of our corpus-behavior tendency questionnaire, we inherit the instruction of the MBTI-M questionnaire (GU & Hu, 2012) and rewrite it, for we intend to change the responding approach.

We apply a 7-point graded forced-choice format (Brown & Maydeu-Olivares, 2018) as the responding approach. Currently, the commonly used response formats for questionnaires in psychometrics are the forced-choice format (Sisson, 1948) and the Likert scale format (Joshi et al., 2015). In comparison to traditional forced-choice scales, graded forced-choice scales exhibit comparable validity, superior reliability and model fit. Contrary to Likert scales, graded forced-choice scales show better model fit and slightly higher self-other agreement (Zhang et al., 2023). The specific meaning of numbers in common 7-point graded forced-choice is shown in Appendix E.

Here, given that we have rewritten the prompt of responding to personality knowledge questionnaire based on the original instructions of the chosen personality scales, thereby not indicating the specific meaning of numbers 2, 3, 5 and 6. We followed this prompt pattern to avoid influence on LLMs' responses brought by such change, which means only retain the meaning of numbers 1, 4 and 7. Hence, the specific prompt is: *Read the following scenarios with actions A and B carefully and rate each scenario in the range from 1 to 7. 1 means that action A applies to you completely in this scenario, 4 means that action A and action B equally apply (or not) to you in this scenario, and 7 means that action B applies to you completely in this scenario. You only need to give the number.*

These measures above allow us to to observe the congruence between self-knowledge and action of LLMs, to compare human and LLM responses.

3.2. Results

To quantify the similarity between responses, we employ the following four metrics: cosine similarity, Spearman's rank correlation coefficient, value mean difference (VMD) and Proportion of Consistent Pairs.

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.

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

where x_i are LLMs' responses of personality knowledge questionnaire, y_i are LLMs' corresponding responses of scenario and action questionnaire, and x_i and y_i correspond