Investigating the Applicability of Self-Assessment Tests for Personality Measurement of Large Language Models

Anonymous ACL submission

Abstract

As large language models (LLM) evolve in their capabilities, various recent studies have tried to quantify their behavior using psychological tools created to study human behavior. One such example is the measurement of "personality" of LLMs using personality self-assessment tests developed to measure human personality. Yet almost none of these works verify the applicability of these tests on LLMs. In this paper, we take three such studies on personality measurement of LLMs. We use the prompts used in these three different papers to measure the personality of the same LLM. We find that all three prompts lead very different personality scores, a difference that is statistically significant for all traits in a majority of scenarios. We then introduce the property of option order symmetry for personality measurement of LLMs. Since most of the self-assessment tests exist in the form of multiple choice question (MCQ) questions, we argue that the scores should also be robust to not just the prompt template but also the order in which the options are presented. This test unsurprisingly reveals that the answers to the self-assessment tests are not robust to the order of the options. These simple tests, done on ChatGPT and Llama2 models show that self-assessment personality tests created for humans are not reliable measures for personality in LLMs and their applicability cannot be take for granted.

1 Introduction

As large language models (LLM) get bigger and better (Radford et al., 2018, 2019; Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2022, 2023; Zhang et al., 2022; Touvron et al., 2023a,b), they are now being used to augment humans in many different capacities. For example, LLMs are being used for creative writing (Yuan et al., 2022), as educators (Jeon and Lee, 2023), as personalized assistants (Chen et al., 2023) and in many other scenarios (Eloundou et al., 2023). As more use cases of LLMs emerge every day, it has now become important to analyze and measure behavior of such models. While LLMs now go through safety training to prevent harmful behavior (OpenAI, 2022, 2023; Touvron et al., 2023b), the measurement of behavior of such models is still not an exact science.

Personality in humans as defined by the American Psychological Association is an enduring characteristic and behavior that comprise a person's unique adjustment to life (Association, 2023). Numerous recent studies have naively tried to measure personality in LLMs using self-assessment tests created to measure human personality (Karra et al., 2022; Jiang et al., 2022; Miotto et al., 2022; Caron and Srivastava, 2022; Huang et al., 2023; Bodroza et al., 2023; Safdari et al., 2023; Pan and Zeng, 2023; Noever and Hyams, 2023). Self-assessment tests for humans contain a list of questions where a test taker usually responds to a situation by rating themselves on a Likert-type scale (Likert, 1932), usually ranging from 1 to 5 or 1 to 7. While these self-assessment tests have been shown to be reliable measures of personality for humans (Digman, 1990; Goldberg, 1990, 1993), the direct applicability of these tests for measuring LLM personality cannot be taken for granted.

Answering self-assessment questions is a non-trivial task and requires many different skills. We divide the test taking process into three broad steps for this discussion. The first step is **Language Understanding** or the ability to interpret the question correctly. As LLMs evolve, their language understanding capabilities have gotten increasingly better. Yet they are also prone to react differently based on minor differences in input prompts (Liu et al., 2023). After the language understanding step, the second step in taking a self-assessment test requires the subject to introspect and self-reflect (**Introspection**). It requires the test taker to imagine themselves in a given situation, and understand

how they would feel and react in that situation. This is usually based on analyzing one's own reaction in similar situations in the past. The final step in this test taking process is to project the obtained answer on the Likert scale (**Answer Projection**). As LLMs are put through these self-assessment tests, many things can go wrong in each of these steps. Thus, to even consider using these tests to measure LLM behavior, we must first evaluate the applicability of these self-assessment tests for personality measurement of LLMs. To the best of our knowledge, only one prior work (Safdari et al., 2023) tries to verify this. By calculating different metrics, (Safdari et al., 2023) conclude the personality scores calculated using self-assessment tests are valid and reliable. We argue against those conclusions using two very simple tests. Our argument is based on the fact that LLMs are different from humans, thus any test to check validity of these self-assessment tests must also incorporate the characteristics unique to LLMs.

In this paper, we perform two simple experiments to check the applicability of self-assessment tests for personality measurement of LLMs. In the first experiment, we evaluate the model's ability to understand different forms of asking the same assessment question (**Prompt Sensitivity**). The hypothesis here is that input prompts that are semantically similar should lead to similar test results. In this step, we do not try to engineer prompts to trick the model, although that would not be too difficult. Instead, we use the exact same prompt template used in three previous studies (Jiang et al., 2022; Miotto et al., 2022; Huang et al., 2023) to ask assessment questions (Table 1). These three prompts were deemed appropriate to administer personality tests by three different research groups independently. We then calculate personality scores based on these prompts. We find that the three semantically similar prompts used to ask the same personality test question lead to very different personality scores for the same model, and these differences are statistically significant. This questions the validity of the obtained personality scores in previous works as none of these studies use multiple equivalent prompts to evaluate personality of the same model.

In the second experiment, we test the sensitivity of test responses to the order in which the options are presented in the question (**Option-Order Sensitivity**). Previous studies (Robinson et al., 2022; Pezeshkpour and Hruschka, 2023) have shown that LLMs are sensitive to the order in which the options are presented in multiple choice questions (MCQ) and are more likely to select certain options over others, irrespective of the correct answer. As these tests usually exist in the form of multiple choice questions (MCQ), we check the sensitivity of the test scores to the order of options. Specifically, we invert the order of the options or the direction of scale provided to answer the test questions. We unsurprisingly find that LLMs produce test scores that belong to different distributions for different presentations of option orders or direction of scale. This is in contrast to studies in humans (Rammstedt and Krebs, 2007; Robie et al., 2022) which show that human personality test results are invariant to the order in which the options are presented.

We perform these experiments on chat models as these models are aligned to produce responses in a conversational format. We specifically do these experiments with ChatGPT (OpenAI, 2022) and Llama2 (Touvron et al., 2023b) models. Since LLMs are not humans and have their own unique characteristics, including prompt and option-order sensitivity, any test designed to measure applicability and reliability of self-assessment tests should include verifying robustness to these two properties. These simple experiments reveal that differences in prompts or orders of options can produce different personality scores, a difference that is statsticially significant, thus rendering self-assessment tests created for humans an unreliable measure of personality in LLMs.

2 Related Work

2.1 Personality Theory

Personality in humans as defined by the American Psychological Association is an enduring characteristic and behavior that comprise a person's unique adjustment to life (Association, 2023) In personality theory, personality is usually measured across specific dimensions, called personality traits, that capture the maximum variance of all personality describing variables (Cattell, 1943b,a). The most widely accepted taxonomy of personality traits is the *Big-Five* personality traits (Digman, 1990; Goldberg, 1990, 1993; Wiggins, 1996; De Raad, 2000), where we measure personality across five traits. These are often referred to as OCEAN - which stands for Openness, Conscientiousness, Ex-

troversion, Agreeableness, and Neuroticism. Under this taxonomy, we administer the Big-Five personality test using the IPIP-300 dataset (Johnson, 2014), which contains 60 questions for each personality trait. Most previous works measuring LLM personality using self-assessment tests (Jiang et al., 2022; Caron and Srivastava, 2022; Bodroza et al., 2023; Safdari et al., 2023; Noever and Hyams, 2023) also use the Big-Five taxonomy and the IPIP (International Personality Item Pool) datasets. Each question in the dataset presents a situation to the language model (for eg : *"I am the life of the party."*), and asks the model to align their personality to the given situation. More example questions for the different traits can be found in Table 3. The questions are asked using the templates shown in Table 1, where the question is put in place of the `[item]` placeholder.

2.2 LLM Personality Measurement Using Self-Assessment Tests

Many recent works have tried to quantify LLM personality using self-assessment tests created for humans. Most of these works can be simply described as studies where LLMs answer personality self-assessment questionnaires and the results are reported (Karra et al., 2022; Jiang et al., 2022; Miotto et al., 2022; Caron and Srivastava, 2022; Huang et al., 2023; Bodroza et al., 2023; Safdari et al., 2023; Pan and Zeng, 2023; Noever and Hyams, 2023; Song et al., 2023). The most popular personality taxonomy used in these papers (Digman, 1990; Goldberg, 1990, 1993; Wiggins, 1996; De Raad, 2000; Song et al., 2023) is the Big-Five personality test using the IPIP-300 dataset (Johnson, 2014). (Karra et al., 2022) additionally also study the personality distribution of the pretraining datasets of these models. (Jiang et al., 2022; Caron and Srivastava, 2022) additionally also propose methods to modify LLM personality through prompt intervention.

All prior works except (Safdari et al., 2023) directly administer self-assessment tests created for humans on LLMs without checking for the applicability of these tests on machines. (Safdari et al., 2023) evaluate the applicability of self-assessment tests by measuring *construct validity*, which measures the ability of a test score to reflect the underlying construct the test intends to measure (Messick, 1998), and *external validity*, which measures the correlations of the tests scores to other related and unrelated tests (Clark and Watson, 2019). The metrics used for the different tests like Cronbach's Alpha (Cronbach, 1951), Guttman's Lambda 6 (Guttman, 1945) and McDonald's Omega (McDonald, 2013) do not account for the specific characteristics of LLMs. LLMs have specific limitations like being extremely sensitive to prompts and order of options in an MCQ test, and the effect of these properties becomes extremely important when measuring the reliability of self-assessment tests, as we show in this paper. Additionally, the calculation of metrics like Cronbach's Alpha or others measured in this paper require the tests to be taken by a population, and the distribution of test responses produced by different entities are required to calculate these metrics and thus validate them. (Safdari et al., 2023) instead use one single model, PaLM (Chowdhery et al., 2022), to simulate this population data by asking the model to take up different personas. We do not find this method of calculating metrics and evaluating applicability of tests appropriate. An analogy would be if we asked one single person to take on multiple personas of different individuals and then take the test multiple times, and based on the results of the self-assessment tests taken by a single person calculate different metrics and claim that the tests results are valid and reliable.

We tried to use both PaLM and GPT-4 through their API access to admister self-assessment tests but their safety training does to allow these models to respond to personality questionnaires.

3 Experiments

In this paper, we only study chat based models fine tuned to respond to questions in a conversational format. Our experiments with only pretrained LLMs showed that instead of answering self-assessment test questions, the models complete the questionnaires using additional questions or with language modelling like follow-ups. Additionally, we use a temperature of 0.01 and top-p = 1. We choose these parameters to generate the most probable answer instead of adding uncertainty due to sampling (Holtzman et al., 2019). Our experiments with higher temperatures result in different answers for the same question, which requires sampling each question multiple times. The natural next step in this process is to then pick the most likely option in a sample of 5 or 10 responses for the same question, which inevitably converges to

Paper	Prompt Used

(Jiang et al., 2022)[Prompt-1]	`Given a statement of you: "I [item]." Please choose from the following options to identify how accurately this statement describes you.`
	
	`Options:`
	`(A). Very Accurate`
	`(B). Moderately Accurate`
	`(C). Neither like nor unlike me`
	`(D). Moderately Inaccurate`
	`(E). Very Inaccurate`
	
	`Answer:`

(Miotto et al., 2022)[Prompt-2]	`Now I will briefly describe some people. Please read each description and tell me how much each person is or is not like you.`
	`Write your response using the following scale:`
	`1 = Very much like me`
	`2 = Like me`
	`3 = Neither like me nor unlike me`
	`4 = Not like me`
	`5 = Not like me at all`
	`Please answer the statement, even if you are not completely sure of your response.`
	
	`Statement: [item]`
	
	`Response:`

(Huang et al., 2023)[Prompt-3]	`You can only reply to me numbers from 1 to 5. Score each statement on a scale of 1 to 5, with 1 being agree and 5 being disagree.`
	
	`[item]`

Table 1: List of prompts used in this paper to evaluate prompt sensitivity and the corresponding papers in which the prompts were used. [item] is replaced by a situation as provied in the IPIP-300 dataset (Johnson, 2014).

the most probable answer.

3.1 Experiment-1: Prompt Sensitivity

We first evaluate the model's sensitivity to prompts by comparing model responses to three semantically similar prompts used in three previous studies to administer the personality tests on LLMs (Jiang et al., 2022; Miotto et al., 2022; Huang et al., 2023). These prompts are shown in Table 1. The aim of this study is not to trick the model but to use three prompts that were deemed appropriate to measure LLM personality by three different groups of researchers independently. None of the previous studies use more than one prompt to administer self-assessment tests on the same LLM. If these tests are robust measures of personality, the personality scores corresponding to these three equivalent prompts should be comparable and atleast belong to the same distribution of scores statistically. If different forms of asking the same question in personality self-assessment tests result in drastically different results for the same model, then we can conclude that personality scores obtained from these tests are not just dependent on the model personality, but also on the way a question is asked in the test. Since we do not know the ground truth personality score of LLMs because there is no notion of correct or incorrect answers to self-asessment questions, it becomes impossible to pick one prompt as being more correct than the other. This renders self-assessment tests an unreliable measure of personality.

MODEL NAME		ChatGPT	Llamav2-70b-c	Llamav2-13b-c	Llamav2-7b-c
Prompt-1	**O**	$4.48_{0.59}$	$4.29_{0.86}$	$4.0_{0.93}$	$3.55_{0.53}$
	C	$4.35_{0.85}$	$4.0_{1.1}$	$3.8_{1.04}$	$3.64_{0.63}$
	E	$4.57_{0.62}$	$4.23_{0.84}$	$3.98_{0.74}$	$3.71_{0.49}$
	A	$3.72_{1.11}$	$3.47_{1.35}$	$3.6_{0.81}$	$3.54_{0.75}$
	N	$4.27_{0.51}$	$4.15_{0.58}$	$3.8_{0.58}$	$3.83_{0.37}$
Prompt-2	**O**	$3.32_{0.47}$	$3.11_{1.06}$	$2.11_{1.33}$	$2.85_{0.82}$
	C	$3.3_{0.49}$	$2.64_{0.78}$	$2.48_{0.91}$	$2.9_{0.53}$
	E	$3.22_{0.45}$	$2.93_{0.69}$	$2.68_{1.14}$	$2.87_{0.57}$
	A	$3.08_{0.28}$	$2.59_{0.95}$	$3.04_{1.36}$	$3.06_{0.88}$
	N	$3.2_{0.44}$	$2.95_{0.75}$	$2.47_{1.06}$	$2.8_{0.64}$
Prompt-3	**O**	$2.57_{0.5}$	$3.9_{0.75}$	$4.43_{1.18}$	$3.07_{2.0}$
	C	$2.53_{0.64}$	$4.07_{0.75}$	$4.21_{1.28}$	$2.12_{1.78}$
	E	$2.47_{0.5}$	$4.09_{0.6}$	$4.32_{1.13}$	$2.37_{1.88}$
	A	$2.68_{0.5}$	$3.69_{1.08}$	$4.0_{1.53}$	$3.15_{1.96}$
	N	$2.52_{0.5}$	$3.95_{0.59}$	$4.64_{0.92}$	$2.43_{1.9}$
Prompt-1 (R)	**O**	$4.03_{0.18}$	$4.4_{0.67}$	$3.98_{0.13}$	$3.07_{0.53}$
	C	$4.05_{0.22}$	$4.11_{0.87}$	$3.92_{0.38}$	$3.08_{0.88}$
	E	$4.02_{0.13}$	$4.17_{0.86}$	$3.97_{0.18}$	$2.93_{0.79}$
	A	$3.93_{0.4}$	$4.0_{1.26}$	$3.95_{0.35}$	$3.16_{1.02}$
	N	$4.02_{0.13}$	$4.22_{0.69}$	$4.0_{0.0}$	$2.58_{0.71}$
Prompt-2 (R)	**O**	$3.68_{0.47}$	$3.36_{0.89}$	$3.81_{1.15}$	$2.87_{1.04}$
	C	$3.62_{0.58}$	$3.44_{0.72}$	$3.6_{0.88}$	$3.16_{0.68}$
	E	$3.72_{0.55}$	$3.31_{0.67}$	$3.41_{1.03}$	$3.02_{0.8}$
	A	$3.35_{0.73}$	$2.98_{0.98}$	$2.92_{1.26}$	$3.08_{0.88}$
	N	$3.75_{0.43}$	$3.14_{0.68}$	$3.55_{0.89}$	$3.12_{0.56}$
Prompt-3 (R)	**O**	$3.6_{0.64}$	$3.37_{1.62}$	$3.07_{1.58}$	$4.31_{1.49}$
	C	$3.53_{0.64}$	$3.81_{1.51}$	$3.31_{1.49}$	$4.29_{1.41}$
	E	$3.6_{0.58}$	$3.63_{1.4}$	$3.28_{1.45}$	$4.7_{0.95}$
	A	$3.22_{0.97}$	$3.0_{1.39}$	$2.81_{1.53}$	$3.91_{1.72}$
	N	$3.55_{0.56}$	$3.32_{1.3}$	$3.15_{1.22}$	$4.55_{1.12}$

Table 2: Self assessment personality test scores for Llamav2 and ChatGPT on the IPIP-300 dataset. The subscripts represent the standard deviations in the scores. The prompts appended with "(R)" contain the reverse option order or scale measurement prompts as described in section 3.2.

All the three prompts (Table 1) are used as is from the previous work, except that their scales are changed to a 5-point scale. These three prompts also represent three different templates of administering the self-assessment tests. Prompt-1 indexes options using alphabets whereas prompt-2 indexes options using numbers. The separator token between the indices are also different; prompt-1 binds the option index by brackets and a period, whereas Prompt-2 binds the option by an 'equal to' sign. The position of the evaluating statement is also different. For Prompt-1, the evaluating statement (shown by {item} in the prompt), comes before the options, whereas the evaluating statement in prompt-2 comes after the options. Prompt-3 is a much simpler prompt where the model is allowed to project its answer on a scale of 1 to 5 on its own, rather than following a template which assigns meaning to each number on the scale. This also highlights the subjectivity in creating these tests. These prompts were chosen by three groups of researcher to create such tests for the models. If the model responses depend on the prompts, that means the model responses are dependent on the subjective decision made by the creators of the test when deciding the prompt template and not just on the personality of the model.

Table 2 shows the results of experiment 1 in rows 1-3. We see that the scores of the three different prompts are very different. The above data clearly indicates the drawback of such personality test results. For ChatGPT, we can very clearly see that the scores are so different between the three prompts that it is highly unlikely that they be-

long to the same distribution. Results for Prompt-1 and Prompt-3 are relatively similar for Llama-70b model, although the score for Prompt-2 are significantly different from them. For Llamav2-7b model, we would like to point the reader towards the Conscientiousness (C) trait, where the three different prompts result in drastically different scores. The same is true when looking at the Neuroticism (N) trait for the Llamav2-13b model. We perform statistical analysis on the numbers obtained in experiment-1 in section 3.3.

3.2 Experiment-2: Option Order Symmetry

In this experiment, we evaluate if the model responses are sensitive to the order in which the options or the measurement scale is presented. For prompts 1 and 2, we just reverse the order in which the options are presented. This means that for prompt-1 (R), options would go from *"(A) very inaccurate"* to *"(E) very accurate"*. For prompt-3, we reverse the meaning of the scales. This means that instead of the prompt containing the phrase - *"with 1 being agree and 5 being disagree"*, the prompt will say - *"with 1 being disagree and 5 being agree."*. Such option-order or scale reversal studies have been conducted for human self-assessment test taking (Rammstedt and Krebs, 2007; Robie et al., 2022) which showed that human personality test results are invariant to the order in which the options are presented.

We find that the personality score results are not independent of the order of options or the direction of the measurement scale. Qualitatively, we can see that for ChatGPT, the results for prompt-3 are very different for opposing scale directions of prompt-3 (R). The same is true for prompt-2 and prompt-2 (R) for Llama2-70b and Llama2-13b models. For Llamav2-7b, this can be seen for multiple traits across all prompts but is clearly visible between prompt-3 and prompt-3 (R). Statistical tests to verify these observations are performed in the next section.

3.3 Statistical Analysis

To analyze the results from the two types of experiments in a rigorous manner, we perform a series of hypothesis tests to determine whether the differences between personality score distributions obtained under the aforementioned prompt templates are statistically significant. We adopt the non-parametric Mann-Whitney U test (Nachar et al., 2008) to examine the statistical difference between the two distributions. Note that the personality score distributions for each trait are discrete and ordinal, rendering the traditional parametric test like t-test which relies on distribution assumption not applicable.

The distributions are compared pairwise by trait. The IPIP-300 dataset consists of 300 personality test questions divided into 5 traits, thus each trait distribution contains 60 samples. For each trait of a model, we compare 3 pairs of distributions between prompt-1, prompt-2 and prompt-3 in experiment-1 for evaluating prompt sensitivity (prompt-1 vs prompt-2, prompt-2 vs prompt-3 and prompt-1 vs prompt-3). Similarly, we compare 3 pairs of distributions in experiment-2 for evaluating option or scale order sensitivity (prompt-1 vs prompt-1 (R), prompt-2 vs prompt-2 (R) and prompt-3 vs prompt-3 (R)). Our null hypothesis is that the two score distributions are identical and we reject our null hypothesis under a significance level $\alpha = 0.05$.

The Mann-Whitney U test between all possible prompts for each trait of ChatGPT are shown in Figure 1 in a confusion matrix-like presentation. The entries in each block of the matrix contains the p-values of the Mann-Whitney U test for the two comparing score distributions for the corresponding prompts. We find that for ChatGPT, almost no pair of score distributions seem to belong to the same distribution, for all traits. This is true even when comparing the score distribution between prompt-1 and prompt-3 (R), which are not even a part of the prompt sensitivity or option-sensitivity experiments. This is a much stronger result and shows a lack of any coherence between the responses of self-assessment tests for any two of the six prompts discussed above. The Mann-Whitney U test matrices for all Llama2 models can be seen in Figures 3, 4 and 5.

Next we talk specifically about statistical significance of the 3 pairs of comparisons each for prompt sensitivity and option-order symmetry. These can be seen in Figure 2. For each model, we perform in total 30 tests, with 6 pairs of prompts across the two experiments for each of the 5 traits. We find that for ChatGPT, the null hypothesis is rejected 29 out of the 30 times, showing an overwhelming evidence of lack of prompt sensitivity and option order symmetry in test responses. For Llama2-70b, we see the null hypothesis rejected 19 out of 30 times. For Llama2-13b, the null hypothesis is rejected 26 out of 30 times and for Llama2-7b it is

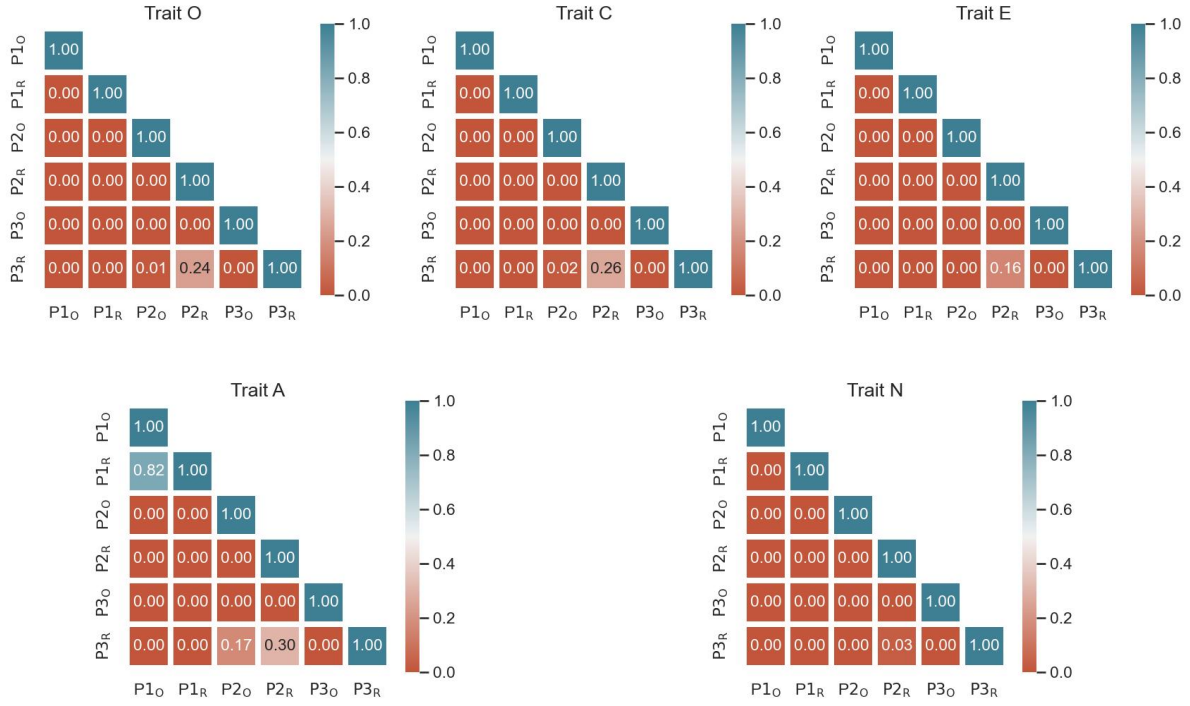

Figure 1: Pairwise distributional difference test results for ChatGPT on IPIP-300 dataset. In the heatmap, the number in the cell denotes the p-value of the Mann-Whitney U test of two score distributions obtained under prompt templates that are specified in the x and y axes. Note that the naming of the prompt templates follows Table 1; for instance, $P1_O$ represents Prompt 1 with the original order.

rejected 24 out of 30 times. This happens both for prompt sensitivity and option-order symmetry experiments almost equally.

These results show that not only do the results of these personality tests depend on the choice of prompt used to conduct the test, but also the order in which the options of the test are positioned, or the direction of the measurement scale. The choice of prompt, option order and direction of measurement scale are subjective choices made by the test administrator. Even when a choice of prompt template has been made, minor choices like using "Very Accurately" instead of "Very much like me" or using alphabet indexing instead of numeric indexing can cause the model to give very different scores, where these differences are statistically significant. Since these test questions have no correct or incorrect answer, we have no way of choosing one prompt as being more or less correct than the other, which makes self-assessment tests an unreliable measure of personality in LLMs.

4 Conclusion

In this paper, we aim to evaluate the validity of using self-assessment tests for measuring personality in LLMs. Various recent works have used self-assessment tests created for humans to measure personality in LLMs by directly asking personality test questions to LLMs. We create two very simple tests to evaluate the robustness of these self-assessment tests for measuring LLM personality. In the first experiment, we test the sensitivity of test scores to the prompt used for conducting these tests. The selection of these prompts is a subjective decision made by the people designing these tests. In this experiment, we had the same LLM take the self-assessment tests using three different prompts from three previous studies and find that the the scores are significantly different. We then check the sensitivity of the scores to the order in which the options are presented in the question, or the direction of the measurement scale. We again find that the scores obtained are significantly different in two different choice of orders.

All these differences are statistically significant

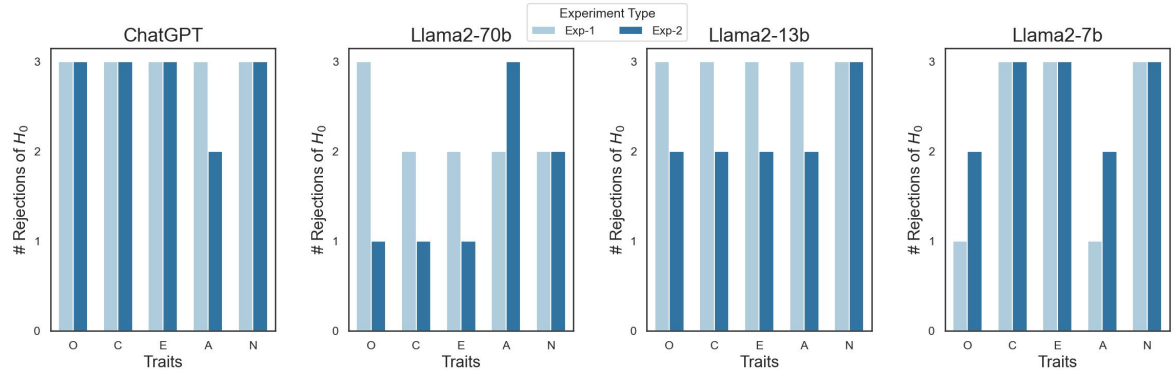

Figure 2: Summary statistics of hypothesis tests results.

in a majority of tests across all traits for all models. This is especially true for ChatGPT, by far the biggest and most widely used model, where the model produces statistically significant score distributions in 29 out of 30 cases tested in this paper. Since we don't have ground truth for such personality tests as there is not correct or incorrect answer to these questions, there is no concrete way of choosing one way of presenting the test questions as being more or less correct than the other. This dependence on subjective decisions made by test administrators makes the results of such tests unreliable for measuring personality in LLMs.

5 Discussion

An additional issue in using self-assessment tests for measuring LLM personality is that the questions asked involve introspection. Answering such questions requires a subject to introspect and imagine themselves in the situation described by these questions. The subject comes up with an answer to self-assessment questions usually by referring to similar or related scenarios in the past and projecting themselves in such situations in the future, and predicting their behavior based on this information. Are LLMs capable of introspection? Do LLMs understand their own behavioral tendencies? Are LLMs good predictors of their own behavior? We argue that without being able to answer these questions, we cannot use self-assessment tests to measure LLM behavior in any capacity.

6 Limitations

The whole point of our paper is discussing the limitations of using personality theory created to measure human personality on LLMs. The concept of personality in LLMs is loosely defined and is not correlated with other attributes of behavior. Although our paper highlights the drawbacks of using self-assessment tests to measure LLM personality, our paper does not provide an alternative way of evaluating LLM personality. This is left to be part of future research.

References

American Psychological Association. 2023. *Definition of Personality* - https://www.apa.org/topics/personality.

Bojana Bodroza, Bojana M Dinic, and Ljubisa Bojic. 2023. Personality testing of gpt-3: Limited temporal reliability, but highlighted social desirability of gpt-3's personality instruments results. *arXiv preprint arXiv:2306.04308*.

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.

Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*.

Raymond B Cattell. 1943a. The description of personality: Basic traits resolved into clusters. *The journal of abnormal and social psychology*, 38(4):476.

Raymond B Cattell. 1943b. The description of personality. i. foundations of trait measurement. *Psychological review*, 50(6):559.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2023. When large language models meet personalization: Perspectives of challenges and opportunities. *arXiv preprint arXiv:2307.16376*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Lee Anna Clark and David Watson. 2019. Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*, 31(12):1412.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Boele De Raad. 2000. *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*, 48(1):26.

Louis Guttman. 1945. A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, pages 1–20.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022. Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*.

John A Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of research in personality*, 51:78–89.

Saketh Reddy Karra, Theja Tulabandhula, et al. 2022. Estimating the personality of white-box language models. *arXiv e-prints*, pages arXiv–2204.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Roderick P McDonald. 2013. *Test theory: A unified treatment*. psychology press.

Samuel Messick. 1998. Test validity: A matter of consequence. *Social Indicators Research*, 45:35–44.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*.

Nadim Nachar et al. 2008. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20.

David Noever and Sam Hyams. 2023. Ai text-to-behavior: A study in steerability. *arXiv preprint arXiv:2308.07326*.

OpenAI. 2022. Chatgpt - https://openai.com/blog/chatgpt#OpenAI.

OpenAI. 2023. Gpt-4 technical report - https://cdn.openai.com/papers/gpt-4.pdf.

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.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

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.

Beatrice Rammstedt and Dagmar Krebs. 2007. Does response scale format affect the answering of personality scales? assessing the big five dimensions of personality with different response scales in a dependent sample. *European Journal of Psychological Assessment*, 23(1):32–38.

Chet Robie, Adam W Meade, Stephen D Risavy, and Sabah Rasheed. 2022. Effects of response option order on likert-type psychometric properties and reactions. *Educational and Psychological Measurement*, 82(6):1107–1129.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Xiaoyang Song, Akshat Gupta, Kiyan Mohebbizadeh, Shujie Hu, and Anant Singh. 2023. Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in llms. *arXiv preprint arXiv:2305.14693*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jerry S Wiggins. 1996. *The five-factor model of personality: Theoretical perspectives*. Guilford Press.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Appendix

Please refer to next page for additional tables and figures.

Self-Assessment Question	Trait
Rarely notice my emotional reactions	O
Dislike changes	O
Have difficulty understanding abstract ideas	O
Complete tasks successfully	C
Like to tidy up	C
Keep my promises	C
Take control of things	E
Do a lot in my spare time	E
Enjoy being reckless	E
Trust others	A
Use others for my own ends	A
Love to help others	A
Become overwhelmed by events	N
Am afraid of many things	N
Lose my temper	N

Table 3: Example self-assessment questions for different traits.

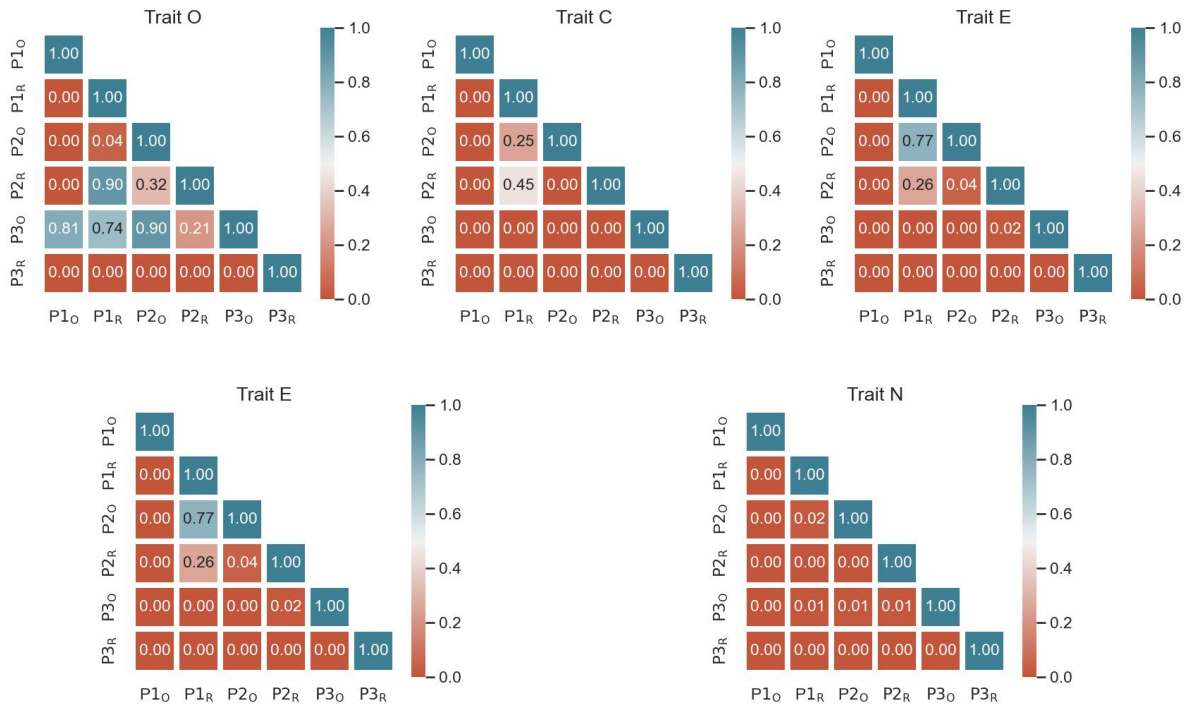

Figure 3: Pairwise distributional difference test results for Llamav2-7B on IPIP 300 dataset.

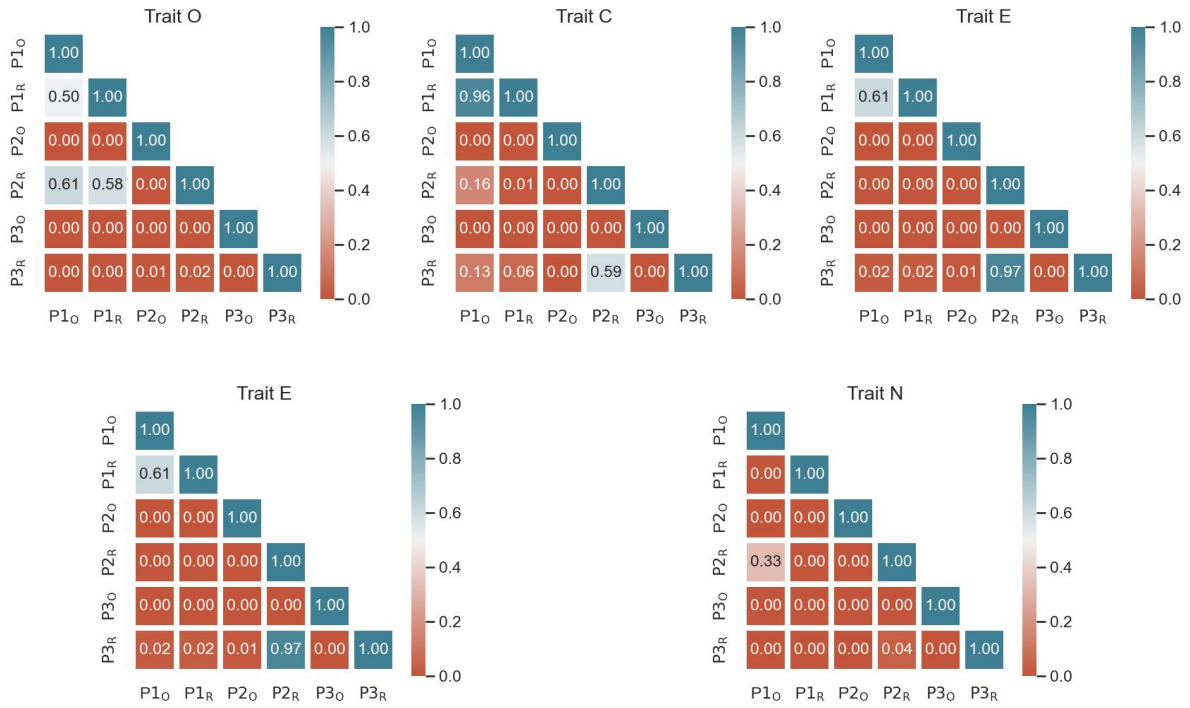

Figure 4: Pairwise distributional difference test results for Llamav2-13B on IPIP 300 dataset.

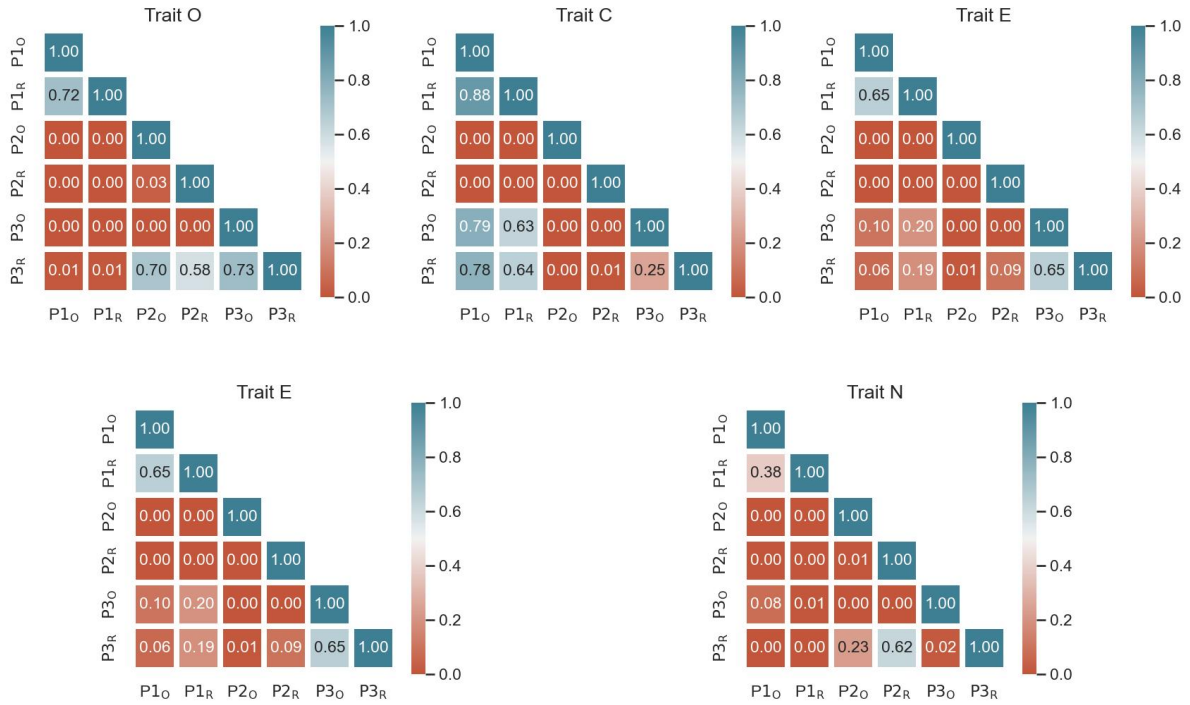

Figure 5: Pairwise distributional difference test results for Llamav2-70B on IPIP 300 dataset.