On the Reliability of Psychological Scales on Large Language Models

Anonymous ACL submission

Abstract

Recent research has extended beyond assessing the performance of Large Language Models (LLMs) to examining their characteristics from a psychological standpoint, acknowledging the necessity of understanding their behavioral characteristics. The administration of personality tests to LLMs has emerged as a noteworthy area in this context. However, the suitability of employing psychological scales, initially devised for humans, on LLMs is a matter of ongoing debate. Our study aims to determine the reliability of applying personality assessments to LLMs, explicitly investigating whether LLMs demonstrate consistent personality traits. Analyzing responses under 2,500 settings reveals that various LLMs show consistency in responses to the Big Five Inventory, indicating a high degree of reliability. Furthermore, our research explores the potential of `gpt-3.5-turbo` to emulate diverse personalities and represent various groups—a capability increasingly sought after in social sciences for substituting human participants with LLMs to reduce costs. Our findings reveal that LLMs have the potential to represent different personalities with specific prompt instructions.

1 Introduction

The recent emergence of Large Language Models (LLMs) marks a significant advancement in the field of Artificial Intelligence (AI), showcasing its abilities in various natural language processing tasks, including text translation (Jiao et al., 2023), sentence revision (Wu et al., 2023), program repair (Fan et al., 2023), and program testing (Deng et al., 2023). Furthermore, LLM applications extend beyond computer science, enhancing fields such as clinical medicine (Cascella et al., 2023), legal advice (Deroy et al., 2023), and education (Dai et al., 2023). Currently, LLMs are catalyzing a paradigm shift in human-computer interaction, revolutionizing how individuals engage with computational systems. With the integration of LLMs, computers have transcended their traditional role as tools to become assistants, establishing a symbiotic relationship with users. Thus, the focus of research extends beyond assessing LLM performance to understanding their behaviors from a psychological perspective. Huang et al. (2024) highlights the significance of psychological analysis on LLMs in developing AI assistants that are more human-like, empathetic, and engaging. Such analysis also plays a crucial role in identifying potential biases or harmful behaviors through the understanding of the decision-making processes of LLMs.

In this context, personality tests aimed at quantifying individual characteristics have gained popularity recently (Safdari et al., 2023; Bodroza et al., 2023; Huang et al., 2024). However, the applicability of psychological scales, initially designed for humans, to LLMs has been contested. Critics argue that LLMs lack consistent and stable personalities, challenging the direct transfer of these scales to AI agents (Song et al., 2023; Gupta et al., 2023; Shu et al., 2023). The essence of this debate lies in the **reliability** of these scales when applied to LLMs. "Reliability" in psychological terms refers to the consistency and stability of results derived from a psychological scale. Evaluating reliability in LLMs differs from its assessment in humans since LLMs demonstrate a heightened sensitivity to input variations compared to humans. For example, humans generally provide consistent responses to questions regardless of their order, while LLMs might yield different answers due to varied contextual inputs. Although consistent results can be obtained from an LLM by querying single items with a zero-temperature parameter setting, such responses are likely to vary under different input conditions. Therefore, our study first systematically investigates the reliability of LLMs on psychological scales under varying conditions, including instruction templates, item rephrasing, language, choice labeling, and choice order. Through analyzing the distribution of all 2,500 settings, we find that various LLMs demonstrate sufficient reliability on the Big Five Inventory.

Additionally, our study further explores whether instructions or contexts can influence the distribution of personality results. We seek to answer whether LLMs can replicate responses of diverse human populations, a capability increasingly sought after by social scientists for substituting human participants in user studies (Dillion et al., 2023). However, this topic remains controversial (Harding et al., 2023), warranting thorough investigation. In particular, we employ three approaches to affecting the personalities of LLMs, from low directive to high directive: (1) by creating a specific environment, (2) by assigning a predetermined personality, and (3) by embodying a character. Firstly, recent research by Coda-Forno et al. (2023) demonstrates the impact of a sad/happy context on LLMs' anxiety levels. Following this work, we conduct experiments to assess LLM's personality within these varied emotional contexts. Secondly, we assign a specific personality for LLM, drawing upon existing literature that focuses on changing the values of LLMs (Santurkar et al., 2023). Thirdly, inspired by Deshpande et al. (2023), which investigates the assignment of a persona to ChatGPT for assessing its tendency towards offensive language and bias, we instruct the LLM to embody the characteristics of a predefined character and measure the resulting personality. Our findings indicate that `gpt-3.5-turbo` can represent various personalities in response to specific prompt adjustments.

The contributions of this study are as follows:

- This study is the first to conduct a systematic analysis of the reliability of psychological scales on LLMs, focusing on five distinct factors.
- Our research contributes to the field of social science by demonstrating the potential of LLMs to simulate diverse human populations accurately.
- We have developed a framework for assessing the reliability of psychological scales on LLMs, which paves the way for future research to validate a broader range of scales on various LLMs.

We will make our experimental results and the corresponding code available to the public upon publication[1], promoting transparency and facilitating further research in this domain.

2 Preliminaries

2.1 Personality Tests

Personality tests are instruments designed to quantify an individual's character, behavior, thoughts, and emotions. A prominent model for assessing personality is the five-factor model, *OCEAN* (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), also known as the Big Five personality traits (John et al., 1999). Other notable models include the Myers-Briggs Type Indicator (MBTI) (Myers, 1962) and the Eysenck Personality Questionnaire (EPQ) (Eysenck et al., 1985), each based on distinct trait theories. Extensive research has demonstrated these models' effectiveness (*i.e.*, reliability and validity) in human subjects. However, the application of these tests to LLMs remains a topic of debate.

2.2 Reliability and Validity of Scales

In psychometrics, the concepts of reliability and validity are crucial for evaluating the quality and effectiveness of psychological scales and tests. **Reliability** refers to the consistency and stability of the results obtained from a psychological test or scale. There are various types of reliability; two common ones are *Test-Retest Reliability* and *Internal Consistency Reliability*. *Test-Retest Reliability* assesses the stability of a test over time (Guttman, 1945) while *Internal Consistency Reliability* checks how well the items within a test measure the same concept or construct (Cronbach, 1951). **Validity** is how well a test measures what it should measure. Researchers usually consider different types of validity, such as *Construct Validity* and *Criterion Validity* (Safdari et al., 2023). Being the most significant type of validity, *Construct Validity* refers to how well a scale measures the theoretical construct it is supposed to measure. *Construct validity* is often demonstrated through correlations with other measures that are theoretically related (*Convergent Validity*) and not correlated with measures that are theoretically unrelated (*Divergent Validity*) (Messick, 1998). *Criterion Validity* assesses how well one measure predicts an outcome based on another measure (Clark and Watson, 2019). It is often split into *Concurrent Validity*, when the scale is compared to an outcome that is already known at the same time the scale is administered; and *Predictive Validity* when the scale is used to predict a future

[1] For reviewers, please see the supplementary materials.

Template	Details
T1	You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`
T2	Now I will briefly describe some people. Please read each description and tell me how much each person is like you. Write your response using the following scale: `LEVEL_DETAILS` Please answer the statement, even if you are not completely sure of your response. `ITEMS`
T3	Given the following statements of you: `ITEMS` Please choose from the following options to identify how accurately this statement describes you. `LEVEL_DETAILS`
T4	Here are a number of characteristics that may or may not apply to you. Please rate your level of agreement on a scale from 1 to 5. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`
T5	Here are a number of characteristics that may or may not apply to you. Please rate how much you agree on a scale from 1 to 5. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

Table 1: Details of different versions of instructions.

outcome (Barrett et al., 1981). While reliability is a necessary but insufficient condition for validity, validity inherently necessitates reliability. Consequently, assessing the reliability of scales forms the foundational step in evaluating the personality traits of LLMs and thus constitutes the primary focus of this study.

3 The Reliability of Scales on LLMs

This section focuses on evaluating the reliability of psychological scales applied to LLMs. We first introduce the framework established for assessing the stability of responses generated by LLMs. Subsequently, we show the findings, including both visual and quantitative data.

3.1 Framework Design

The consistency of responses from LLMs is predominantly determined by their input (Hagendorff, 2023). To assess the reliability of LLMs, it is crucial to examine their responses across varying input conditions. In this study, we propose to deconstruct a query into five distinct factors for a comprehensive analysis: (1) the nature of the instruction, (2) the specific items in the scale, (3) the language used, (4) the labeling of choices, and (5) the order in which these choices are presented.

(1) Instruction Given that LLMs exhibit sensitivity to variations in prompt phrasing, as observed by Bubeck et al. (2023), and Gupta et al. (2023) highlighted that LLMs demonstrate differing personalities under varying prompting instructions, we need to evaluate the influence of different instructions. To this end, we analyze the performance of five distinct prompt templates: T1 as applied in Huang et al. (2024), T2 as used by Miotto et al. (2022), T3 suggested by Jiang et al. (2022), and T4 and T5 both identified in Safdari et al. (2023). Details of prompts are listed in Table 1, where `LEVEL_DETAILS` denotes the definition of each level and `ITEMS` contains the items to be rated by LLMs. Notably, our selection covers all three templates investigated by Gupta et al. (2023).

(2) Item The training data for LLMs likely include items from publicly available personality tests. Consequently, LLMs may develop specific response patterns to these scales during pre-training or instructional tuning phases. In line with previous research that examines LLM performance (Coda-Forno et al., 2023; Bubeck et al., 2023), we rephrase the items in the scale to ensure their novelty to the model. A critical aspect of this evaluation is determining if LLMs consistently respond to different paraphrases of the same item, which would indicate comprehension of the instruction and the ability to provide independent ratings rather than merely recalling training data. To this end, we employ GPT-4 to rephrase the items and manually assess whether there are instances of duplicated sentences and if the rewritten sentences maintain their semantic meaning. This process results in five distinct versions of the items, including the original set.

(3) Language Considering the observed performance disparities among languages in LLMs (Lai et al., 2023; Wang et al., 2023a), coupled with the documented regional variations in personalities (Giorgi et al., 2022; Rentfrow et al., 2015; Krug and Kulhavy, 1973), we are motivated to assess LLMs' personalities across different languages. Consequently, we extend our examination to include nine more languages, namely Chinese

(Zh), Spanish (Es), French (Fr), German (De), Italian (It), Arabic (Ar), Russian (Ru), Japanese (Ja), and Korean (Ko), using the English version as a basis. The translation of the instructions and items (including all the variants) from English into the target languages is conducted using Google Translate[2] and DeepL[3]. To ensure translation quality, we randomly sample part of these machine-translated outputs and manually review and verify the correctness (but may not ensure fluency). Our selection of ten languages includes different language families/groups and various character sets.

(4) Choice Label Liang et al. (2023) demonstrated that LLMs exhibit sensitivity to the formatting of choice labels, such as "1, 2" or "A, B." Our study extends this investigation to include the impact of various choice label formats. Specifically, we examine five formats: (1) lowercase Latin alphabets (*e.g.*, "a, b"), (2) uppercase Latin alphabets (*e.g.*, "A, B"), (3) lowercase Roman numerals (*e.g.*, "i, ii"), (4) uppercase Roman numerals (*e.g.*, "I, II"), and (5) Arabic numerals (*e.g.*, "1, 2").

(5) Choice Order The order of choices may impact the responses of LLMs, as these models are sensitive to the order of presented examples (Zhao et al., 2021). To account for this, we introduce two ordering methods: (1) an ascending scale where "1" denotes strong disagreement and "7" indicates strong agreement, and (2) a descending scale where "1" signifies strong agreement and "7" denotes strong disagreement.

By integrating the five specified factors, we obtain $5 \times 5 \times 10 \times 5 \times 2 = 2500$ distinct configurations. Traditional frameworks often vary only one factor at a time while keeping others constant, potentially leading to insufficient observation and restricted generalizability of their findings. Our approach, however, systematically examines every possible combination of these factors, aiming for more comprehensive and universally applicable conclusions.

3.2 Experimental Results

Our experiments utilize the Big Five Inventory (BFI) (John et al., 1999). The BFI comprises 44 items, each rated on a five-point Likert scale. This inventory is a widely-recognized and publicly available instrument for assessing personality traits, commonly known as the Five Factor Model or *OCEAN*. Subscales of BFI include (the number of items for each subscale is specified in parentheses): (1) *Openness to experience (O)* (10) is characterized by an individual's willingness to try new things, their level of creativity, and their appreciation for art, emotion, adventure, and unusual ideas. (2) *Conscientiousness (C)* (9) refers to the degree to which an individual is organized, responsible, and dependable. (3) *Extraversion (E)* (8) represents the extent to which an individual is outgoing and derives energy from social situations. (4) *Agreeableness (A)* (9) measures the degree of compassion and cooperativeness an individual displays in interpersonal situations. (5) *Neuroticism (N)* (8) evaluates whether an individual is more prone to experiencing negative emotions like anxiety, anger, and depression or whether the individual is generally more emotionally stable and less reactive to stress. Overall results are derived by calculating the mean score for each subscale.

Given its leading-edge capabilities in conversational AI and its extensive user base, we have chosen ChatGPT as our primary language model (LLM) for evaluation. For our experiments, we utilize GPT models[4] and `Gemini`[5] via their official APIs, with the temperature parameter set to zero. This section shows the results of `gpt-3.5-turbo` due to page limit. The results of `gpt-4` can be found in §A in the appendix. To introduce more significant variability into the input data for the LLM, we randomized the order of the items in the scale, submitting between 17 to 27 items simultaneously (equivalent to $44/2 \pm 5$). This methodology is crucial to ascertain whether LLMs consistently produce reliable outputs, regardless of the items' positions within the given context. In each setting outlined in §3.1, we evaluate the LLM using these randomization techniques, yielding a total of 2,500 data points. Each data point is a five-dimensional vector representing the *OCEAN* scores.

Visualization Results are then projected onto a two-dimensional space for visualization, as illustrated in Fig. 1. The projection matrix is derived from a PCA process of projecting all grids ranging from 1 to 5 from a five-dimensional to a two-dimensional space. The region delineated by our figures precisely encompasses all these projected

[2] https://translate.google.com/
[3] https://www.deepl.com/en/translator
[4] https://platform.openai.com/docs/models
[5] https://ai.google.dev/tutorials/python_quickstart

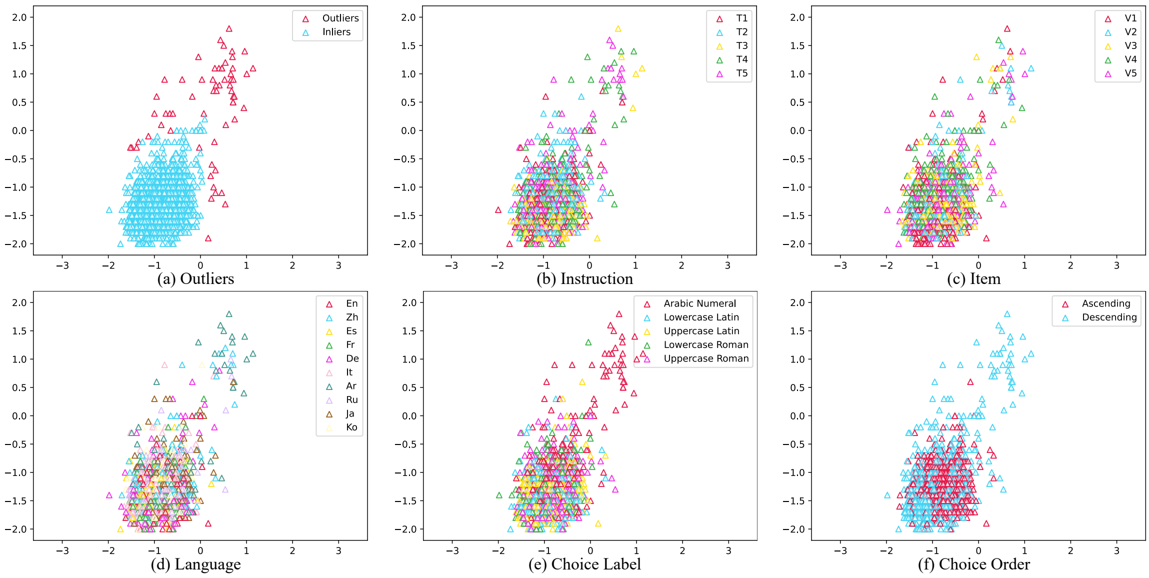


Figure 1: Visualization of all data points regarding different factors, marked in distinct colors.

grids, which means the space comprises all possible values obtained from a BFI test. We can make the following observations: (1) The majority of data points are concentrated in the lower-left region, with 61 outliers ($< 2.5\%$) located in the upper-right area. Outliers are detected by a DBSCAN method with `eps` $= 0.3$ and `minPt` $= 20$. (2) Overall, no significant influence of any factor on the results is observed, indicating a similar distribution across all factors. (3) Nearly all outliers correspond to settings with an Arabic numeral choice label, descending choice order, and Arabic and Chinese languages, suggesting a potential lower comprehension ability in these languages.

Quantitative Analysis Firstly, we compared the means of data points using a specific factor with other data points. For example, we can check whether there are significant differences in means between data points using English and those using other languages. According to Table 4, the majority of factors do not exhibit significant differences when compared with others. Out of 135 comparisons (27 factors across 5 dimensions), only 7 demonstrate a difference exceeding 0.15. Furthermore, we calculate the standard deviations for the five dimensions and compare them with recorded human norms (Srivastava et al., 2003). In the OCEAN dimensions, `gpt-3.5-turbo` records standard deviations of 0.3, 0.3, 0.4, 0.3, and 0.4, respectively, while the crowd data show a higher variability with 0.7, 0.7, 0.9, 0.7, and 0.8. These findings suggest that `gpt-3.5-turbo` demonstrates a consistent performance across different perturbations, and it is more deterministic compared to the broader variability observed in the crowd data.

3.3 Test-Retest Reliability

As introduced in §2.2, Test-Retest Reliability is another key measure, reflecting the stability of results over time. Since OpenAI periodically updates the `gpt-3.5-turbo`, to evaluate this reliability, we call the API biweekly, starting from mid-September 2023. Our analysis includes two primary versions of the `gpt-3.5-turbo-0613` and the `gpt-3.5-turbo-1106`. The results, specifically focusing on the BFI, are illustrated in Fig. 2. The analysis indicates no significant variation attributable to model updates during this period, showing a high level of reliability.

Findings 1: Given that the responses are not random and exhibit stability against various perturbations as well as over time, `gpt-3.5-turbo` demonstrates satisfactory levels of *Internal Consistency Reliability* and *Test-Retest Reliability* on the BFI.

4 Representing Diverse Groups

Our focus shifts from assessing the default personalities of LLMs to evaluating their contextual steerability. This involves investigating whether the personality distribution depicted in Fig. 1 can

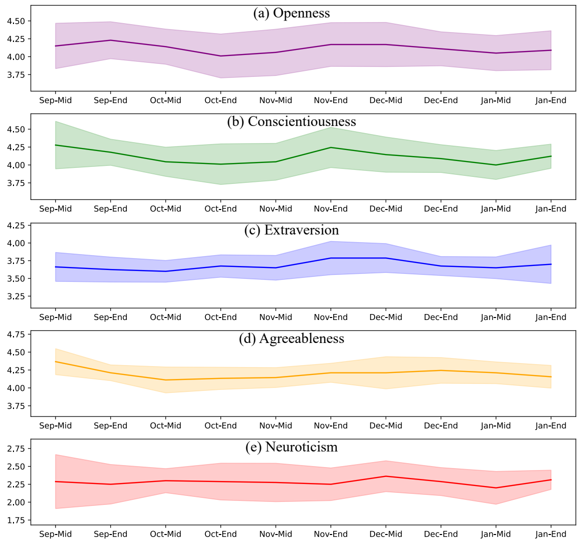


Figure 2: Biweekly measurements starting from mid-September 2023 of the BFI on `gpt-3.5-turbo`. The shadow represents the standard deviation ($\pm Std$).

be modified through specific instructions or contextual cues. Researchers in the social sciences are exploring the potential of substituting human subjects with LLMs to reduce costs. Our research helps by offering valuable insights into the capabilities of LLMs to accurately represent diverse human populations. Furthermore, the ability of LLMs to exhibit a range of personalities is essential, considering the growing demand for AI assistants with tailored stylistic attributes. We propose three strategies: (1) low directive, which involves creating an environment; (2) moderate directive, entailing the assignment of a personality; and (3) high directive, which encompasses the embodiment of a character.

4.1 Approaches

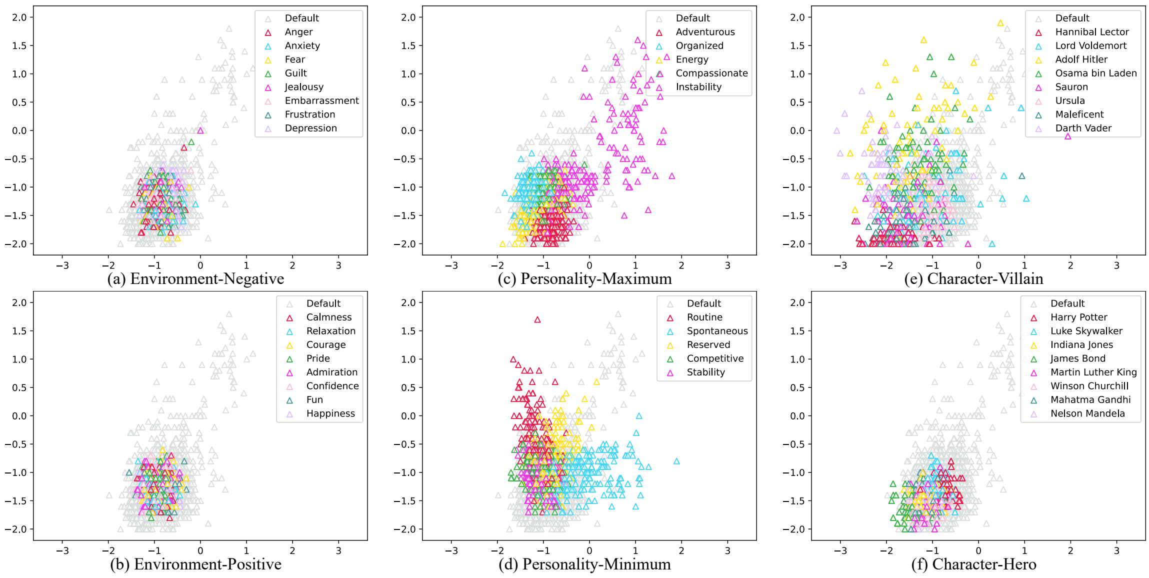
Creating an Environment Coda-Forno et al. (2023) has demonstrated the capability to induce increased levels of anxiety in LLMs through the incorporation of sad or anxious narratives. Building on this finding, our study introduces both negative and positive environmental contexts to LLMs before conducting the personality test. In line with previous studies on LLMs' emotion appraisals (Huang et al., 2023), our methodology in the negative condition involves instructing the LLM to generate narratives encompassing emotions such as anger, anxiety, fear, guilt, jealousy, embarrassment, frustration, and depression. Conversely, in the positive condition, the LLM is prompted to create stories that evoke emotions like calmness, relaxation, courage, pride, admiration, confidence, fun, and happiness.

Assigning a Personality We employ the three approaches proposed by Santurkar et al. (2023) to assign a specific personality (denoted as $\mathcal{P}$) to the LLM: (1) Question Answering (QA): This approach involves presenting personalities through multiple-choice questions, with $\mathcal{P}$ specified through an option at the end of the prompt. 2) Biography (BIO): Here, the LLM is prompted to generate a brief description of its personality, which we use to assign $\mathcal{P}$, incorporating this description directly into the prompt. 3) Portray (POR): This technique explicitly instructs the LLM to be $\mathcal{P}$. To enhance the LLM's comprehension of $\mathcal{P}$, we adopt a methodology inspired by the Chain-of-Thought (CoT) approach (Wei et al., 2022). The approach aims to instruct the model to articulate characteristics associated with $\mathcal{P}$ before engaging in the personality test. In selecting $\mathcal{P}$, we aim to diverge as much as possible from the default distribution. This involves examining every maximum and minimum value across each personality dimension. For instance, a $\mathcal{P}$ that maximizes "Openness" is considered more adventurous and creative. Consequently, we identify ten distinct personality profiles for our analysis.

Embodying a Character Recent studies (Zhuo et al., 2023; Deshpande et al., 2023) have explored the induction of toxic content generation in ChatGPT by simulating the speech patterns of historical or fictional figures. Additionally, research has explored the capacity of LLMs to adopt distinct characters (Wang et al., 2023c; Shao et al., 2023) and examined the consistency of LLMs' personalities with these characters Wang et al. (2023b). Building upon this line of research, our study concentrates on instructing LLMs to fully represent a specific character, referred to as $\mathcal{C}$. To assign $\mathcal{C}$, we first prompt the LLM with only the character's name. We then extend this approach using the CoT methodology, providing the LLM with detailed experiences attributed to $\mathcal{C}$. For the selection of $\mathcal{C}$, we include a diverse range of heroes and villains from both fictional and real-world contexts, detailing 16 characters in Table 7 in the Appendix. Table 2 displays the prompts for each of the three approaches.

4.2 Results

To facilitate a comparative analysis with the results in §3.2 (referred to as "default" in this section), we apply the BFI on `gpt-3.5-turbo` with the same



(a) Environment-Negative (c) Personality-Maximum (e) Character-Villain

(b) Environment-Positive (d) Personality-Minimum (f) Character-Hero

Figure 3: Visualization of all data points of different choices, marked in distinct colors.

settings. For each method, we vary factors (keeping language fixed to English) to generate approximately 2,500 data points, aligning with the size used for the default data. These data are then projected into a two-dimensional space and visualized alongside the default data in Fig. 3. The results yielded several insights: (1) The distribution of personality outcomes, obtained by altering the atmosphere of the conversation, closely aligns with the default distribution. This suggests that environmental changes do not significantly alter the LLM's personality traits. (2) When different personalities are assigned to `gpt-3.5-turbo`, it demonstrates a capacity to reflect diverse human characteristics, indicated by the diverged distribution patterns for various personalities from the default. Moreover, by simultaneously maximizing and minimizing specific personality dimensions, we observe that the distributions of the extremities of each dimension are positioned on opposite ends. For example, the red points in Fig. 3(c) and Fig. 3(d) mark the high and low *Openness*. A clearer comparison for each dimension can be found in Fig. 8 in the appendix. This confirms that `gpt-3.5-turbo` effectively distinguishes between each BFI dimension's high and low values. (3) Assigning various characters to the LLM reveals its ability to represent a broader spectrum of human populations, as indicated in Fig. 3(e). However, the representation of heroic characters shows a distribution pattern similar to the default. We hypothesize that this similarity arises from the model's inherent positive bias.

Fig. 4 presents the distribution patterns observed when applying QA, BIO, and POR methods for personality assignment. Specifically, among the three, only POR effectively alters the personality distribution of `gpt-3.5-turbo`. Moreover, Fig. 4 differentiates between data points with and without the CoT approach. Our analysis reveals that the CoT approach does not significantly influence the results of personality distribution.

Findings 2: `gpt-3.5-turbo` demonstrates the capability to adopt varied personalities in response to specific prompt adjustments. Furthermore, `gpt-3.5-turbo` shows a precise comprehension of the assigned personalities, indicated by the distinct clusters at opposite ends of the same dimension, as illustrated in Fig. 3(c) and 3(d).

5 Discussions

5.1 Limitations

This study has several limitations. Firstly, the modifications made to the scale's instructions and items, including translation into different languages, may impact its reliability and validity. Psychological scales are meticulously crafted in their wording, and any translation necessitates a reevaluation of their reliability and validity across different cultural contexts. Consequently, our transformations could potentially hurt the original scale's reliability

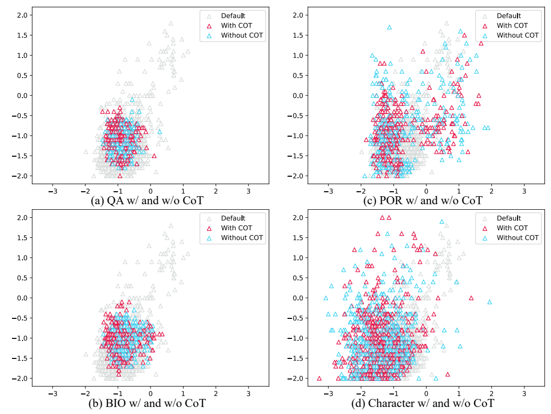


Figure 4: Visualization of all data points of assigning a personality and embodying a character. Different colors indicate whether or not the prompts include a CoT.

and validity. Additionally, these changes preclude the use of Cronbach's alpha (Cronbach, 1951) for assessing the internal consistency reliability. However, in the context of LLM, studying the reliability of psychological scales without considering the effects of prompt variations is insufficient. Varying prompt templates has been a standard practice in this research domain (Safdari et al., 2023; Coda-Forno et al., 2023). Secondly, the study explores limited methods for influencing LLMs' personality results. While numerous approaches exist (Wang et al., 2023c; Shao et al., 2023), we select three representative methods to verify our hypothesis regarding LLMs' ability to mirror diverse human populations. With the help of our framework, future research can dig deeper into a broader range of methods.

5.2 Related Work

Exploring the personality traits of LLMs has become a prevalent research direction. Miotto et al. (2022) analyzed GPT-3's personality traits, values, and demographics. Karra et al. (2022), Jiang et al. (2022), and Bodroza et al. (2023) conducted personality assessments on various LLMs, including BERT, XLNet, TransformerXL, GPT-2, GPT-3, and GPT-3.5. Li et al. (2022) investigated whether GPT-3, InstructGPT, and FLAN-T5 display psychopathic tendencies as part of their personality assessment. Jiang et al. (2023) examined the potential for assigning a distinct personality to `text-davinci-003`. Romero et al. (2023) undertook a cross-linguistic study of GPT-3's personality across nine languages. Rutinowski et al. (2023) evaluated ChatGPT for personality traits and political values. Safdari et al. (2023) tested the validity of the BFI on the PaLM model family. Huang et al. (2024) applied thirteen different personality and ability tests to LLaMA-2, `text-davinci-003`, `gpt-3.5-turbo`, and `gpt-4`. Our study is distinct by offering a detailed analysis of the reliability of psychological scales on LLMs. We vary instructions, items, languages, choice labels, and order to evaluate the robustness of LLM responses. From 2,500 data points, we conclude that `gpt-3.5-turbo` exhibits specific personality traits and demonstrates satisfactory reliability on the BFI.

However, researchers are arguing that conversational AI, at its current stage, lacks stable personalities (Song et al., 2023; Gupta et al., 2023; Shu et al., 2023). We believe that this perception may stem from the limitations of the models assessed in Song et al. (2023) and Shu et al. (2023), which are comparatively smaller and less versatile in various tasks than our selected model, `gpt-3.5-turbo`. Notably, Gupta et al. (2023) indicates that the personality traits of `gpt-3.5-turbo` vary across three different instruction templates of the BFI, which is inconsistent with our findings. This discrepancy could be attributed to their methodology of choosing the most likely response from a set of 5 or 10, in contrast to our approach of utilizing the average response. However, we argue that employing the mean is a more standard practice in this context (Srivastava et al., 2003).

6 Conclusion

This study examines the reliability of psychological scales initially designed for human assessment when applied to LLMs. Through a comprehensive methodology involving varied instruction templates, item wording, languages, choice labels, and choice order, this research includes 2,500 distinct experimental settings. Data analysis reveals that `gpt-3.5-turbo`, `gpt-4`, and `Gemini` consistently generate stable responses on the BFI across diverse settings. Comparative analysis of the standard deviations with established human norms indicates that the model does not produce random responses but exhibits tendencies towards specific personality traits. Furthermore, the study explores the potential for manipulating the distribution of personalities by creating an environment, assigning a personality, and embodying a character. The findings demonstrate that `gpt-3.5-turbo` can represent diverse personalities by adjusting prompt inputs.

Ethics Statements

We would like to emphasize that the primary objective of this paper is to facilitate the scientific inquiry into understanding LLMs from a psychological standpoint. Users must exercise caution and recognize that the performance on this benchmark does not imply any applicability or certificate of automated counseling or companionship use cases.

References

Gerald V Barrett, James S Phillips, and Ralph A Alexander. 1981. Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, 66(1):1.

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*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1):33.

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

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. *arXiv preprint arXiv:2304.11111*.

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

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE.

Yinlin Deng, Chunqiu Steven Xia, Haoran Peng, Chenyuan Yang, and Lingming Zhang. 2023. Large language models are zero-shot fuzzers: Fuzzing deep-learning libraries via large language models. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 423–435.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270.

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

Sybil BG Eysenck, Hans J Eysenck, and Paul Barrett. 1985. A revised version of the psychoticism scale. *Personality and individual differences*, 6(1):21–29.

Zhiyu Fan, Xiang Gao, Martin Mirchev, Abhik Roychoudhury, and Shin Hwei Tan. 2023. Automated repair of programs from large language models. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1469–1481. IEEE.

Salvatore Giorgi, Khoa Le Nguyen, Johannes C Eichstaedt, Margaret L Kern, David B Yaden, Michal Kosinski, Martin EP Seligman, Lyle H Ungar, H Andrew Schwartz, and Gregory Park. 2022. Regional personality assessment through social media language. *Journal of personality*, 90(3):405–425.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2023. Investigating the applicability of self-assessment tests for personality measurement of large language models. *arXiv preprint arXiv:2309.08163*.

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

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.

Jacqueline Harding, William D'Alessandro, N. G. Laskowski, and Robert Long. 2023. Ai language models cannot replace human research participants. *AI & SOCIETY*.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2024. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *Proceedings of the Twelfth International Conference on Learning Representations*.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: theory and research*.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.

Samuel E Krug and Raymond W Kulhavy. 1973. Personality differences across regions of the united states. *The Journal of social psychology*, 91(1):73–79.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189.

Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2022. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*.

Tian Liang, Zhiwei He, Jen-tes Huang, Wenxuan Wang, Wenxiang Jiao, Rui Wang, Yujiu Yang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Leveraging word guessing games to assess the intelligence of large language models. *arXiv preprint arXiv:2310.20499*.

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. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, pages 218–227.

Isabel Briggs Myers. 1962. *The Myers-Briggs Type Indicator: Manual (1962)*. Consulting Psychologists Press.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sundar Pichai and Demis Hassabis. 2023. Introducing gemini: our largest and most capable ai model. *Google Blog Dec 06 2023*.

Peter J Rentfrow, Markus Jokela, and Michael E Lamb. 2015. Regional personality differences in great britain. *PloS one*, 10(3):e0122245.

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

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of chatgpt. *arXiv preprint arXiv:2304.07333*.

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*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Dallas Card, and David Jurgens. 2023. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. *arXiv preprint arXiv:2311.09718*.

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*.

Sanjay Srivastava, Oliver P John, Samuel D Gosling, and Jeff Potter. 2003. Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of personality and social psychology*, 84(5):1041.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023a. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*.

Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. 2023b. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023c. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*.

A Reliability Tests on Other LLMs

We also explore the reliability of different LLMs on the BFI, taking into account their variations in training datasets and instruction tuning methodologies. We extend our analysis to include OpenAI's `gpt-4` (OpenAI, 2023) and Google's `Gemini-Pro` (Pichai and Hassabis, 2023), running on the same 2,500 profiles as those applied to `gpt-3.5-turbo`. Fig. 5 and Fig. 6 illustrate the data points generated from `gpt-4` and `Gemini`, respectively. Consistent with our previous experiments on `gpt-3.5-turbo`, we utilize DBSCAN parameters of `eps` $= 0.3$ and `minPt` $= 20$. The outlier rates for `gpt-4` and `Gemini-Pro` are approximately 4.1% and 2.4%, respectively. Our findings indicate that: (1) The model responses are not uniformly distributed across the BFI space, suggesting a significant level of reliability across all examined LLMs. (2) Each model exhibits a unique personality profile. `gpt-4`'s personality significantly diverges from that of `gpt-3.5-turbo`, whereas `Gemini-Pro` displays a personality more akin to `gpt-3.5-turbo`. For clarity, we present the personality distribution of the three models in Fig. 7.

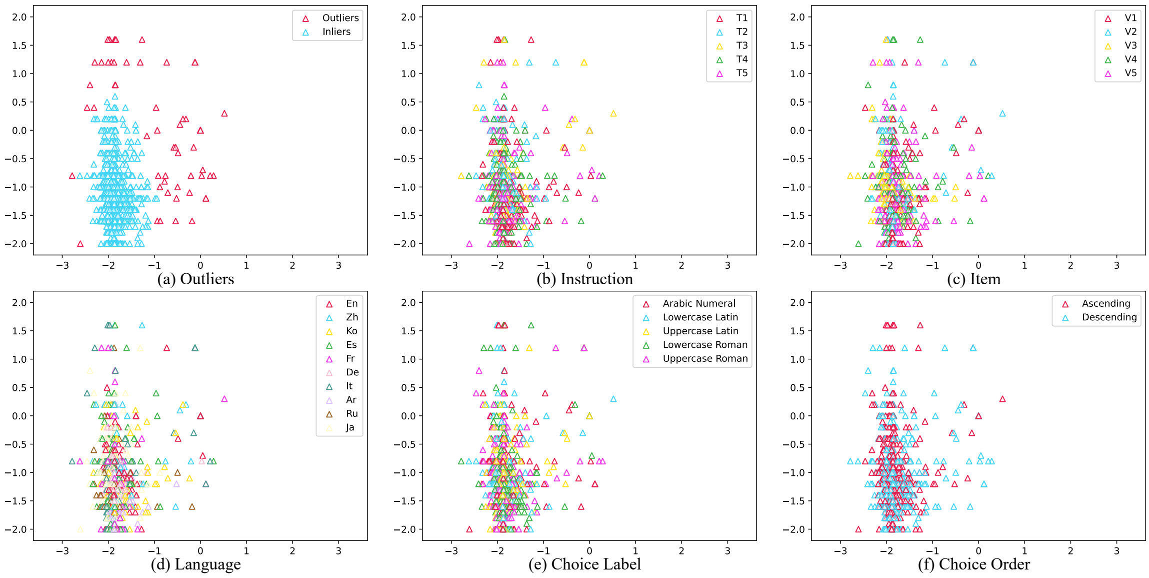


Figure 5: Visualization of all data points produced by `gpt-4` regarding different factors, marked in distinct colors.

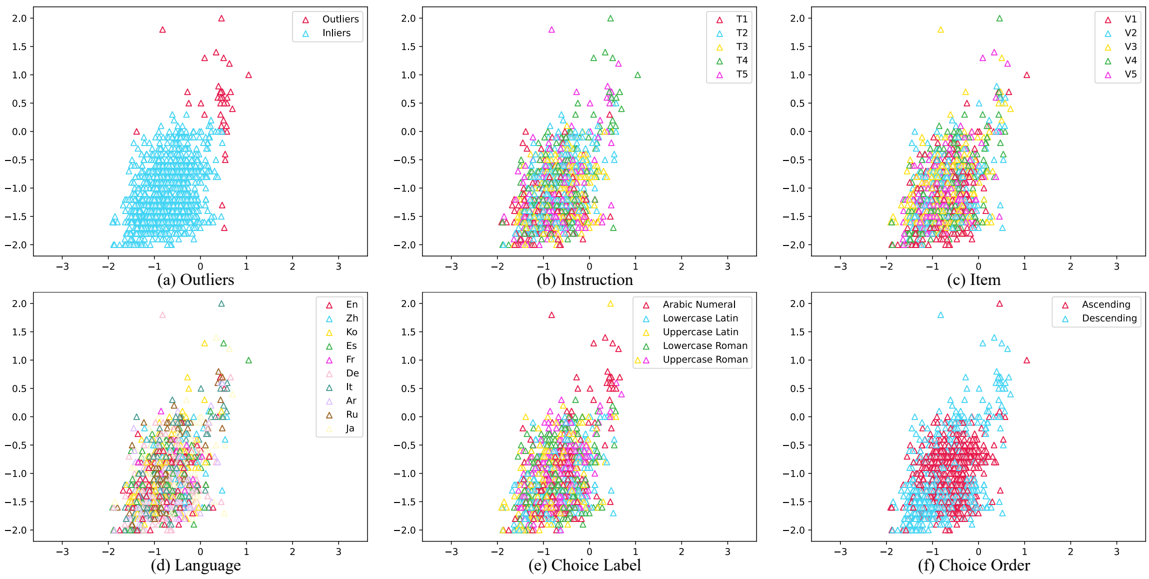


Figure 6: Visualization of all data points produced by `Gemini` regarding different factors, marked in distinct colors.

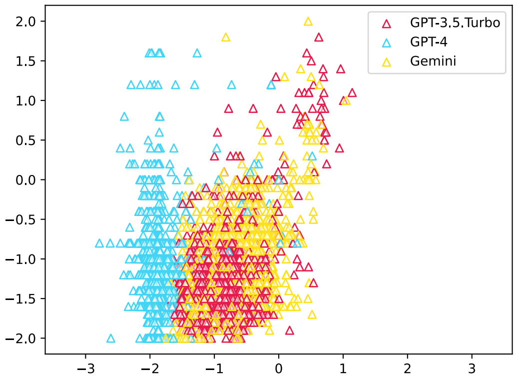
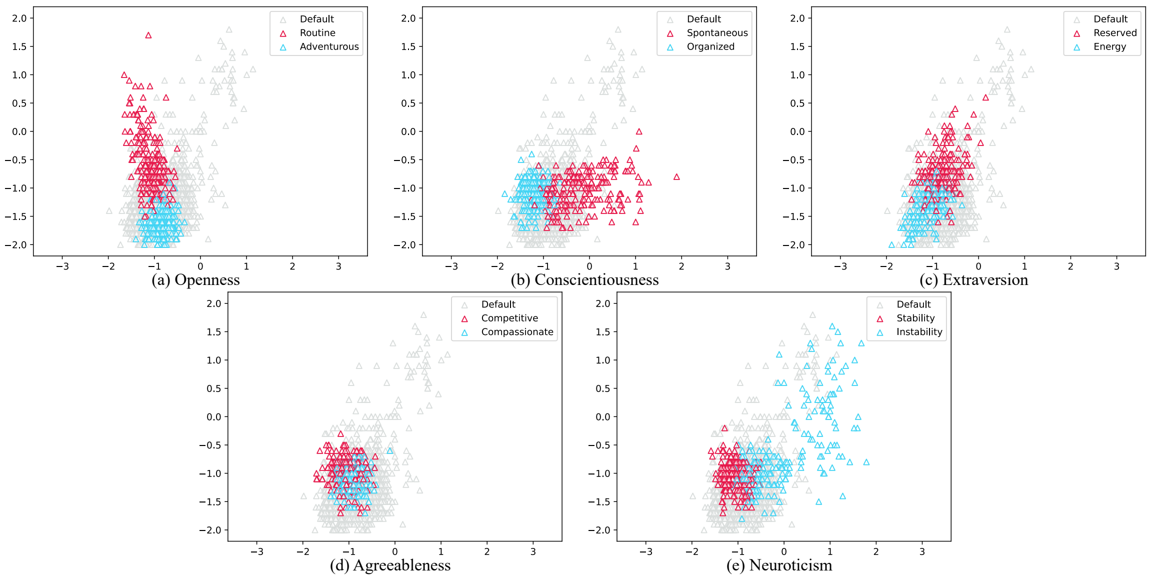


Figure 7: Comparison of the personality distribution of `gpt-3.5-turbo`, `gpt-4`, and `Gemini-Pro` on the BFI.

B Comparison on Each Dimension



(a) Openness (b) Conscientiousness (c) Extraversion (d) Agreeableness (e) Neuroticism

Figure 8: Comparison of the two extremities assigned to `gpt-3.5-turbo` for each dimension of the BFI.

C Prompt Templates of Modifying Personalities

Table 2: The prompts used for creating positive/negative environments, assigning personalities, and embodying characters. LLM's responses are marked in *Italian*.

Environment

Please tell a story that evokes `EMOTION` with around 100 words.

ChatGPT: A short story.

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

Question Answering

Question: Among the personalities, do you consider yourself a:
A. $\mathcal{P}_1$ B. $\mathcal{P}_2$ C. $\mathcal{P}_3$ D. $\mathcal{P}_4$ E. $\mathcal{P}_5$
Answer: A

(Optional) *ChatGPT: A description of* $\mathcal{P}_1$.

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

Biography

Below you will be asked to provide a short description of your personality and then answer some questions.
Description: Among the personalities, I consider myself an $\mathcal{P}$.

(Optional) *ChatGPT: A description of* $\mathcal{P}$

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

Portray

Answer the following questions as if among the personalities, you consider yourself an $\mathcal{P}$.

(Optional) *ChatGPT: A description of* $\mathcal{P}$

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

Character

You are $\mathcal{C}$. Please think, behave, and talk based on $\mathcal{C}$'s personality trait.

(Optional) A description of the experience of $\mathcal{C}$.

You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`

D More Details

D.1 Multilingual Prompts

Table 3: The instructions to complete the personality tests in ten languages.

Lan	Prompt
En	You can only reply from 1 to 5 in the following statements. Here are a number of characteristics that may or may not apply to you. Please indicate the extent to which you agree or disagree with that statement. `LEVEL_DETAILS` Here are the statements, score them one by one: `ITEMS`
Zh	在以下语句中只能回覆1到5。以下是一些可能适用或可能不适用于您的特征。请表明您同意或不同意该声明的程度。`LEVEL_DETAILS` 以下是陈述，请一一评分：`ITEMS`
Es	Sólo puedes responder del 1 al 5 en las siguientes afirmaciones. Aquí hay una serie de características que pueden aplicarse o no a usted. Indique en qué medida está de acuerdo o en desacuerdo con dicha afirmación. `LEVEL_DETAILS` Aquí están las afirmaciones, puntúelas una por una: `ITEMS`
Fr	Vous ne pouvez répondre que de 1 à 5 dans les affirmations suivantes. Voici un certain nombre de caractéristiques qui peuvent ou non s'appliquer à vous. Veuillez indiquer dans quelle mesure vous êtes d'accord ou en désaccord avec cette affirmation. `LEVEL_DETAILS` Voici les énoncés, notez-les un par un: `ITEMS`
De	In den folgenden Aussagen können Sie nur eine Antwort von 1 bis 5 geben. Hier sind eine Reihe von Merkmalen aufgeführt, die möglicherweise auf Sie zutreffen oder auch nicht. Bitte geben Sie an, inwieweit Sie dieser Aussage zustimmen oder nicht. `LEVEL_DETAILS` Hier sind die Aussagen, bitte bewerten Sie sie einzeln: `ITEMS`
It	Puoi rispondere solo da 1 a 5 nelle seguenti affermazioni. Ecco alcune caratteristiche che potrebbero applicarsi o meno a te. Si prega di indicare in che misura si è d'accordo o in disaccordo con tale affermazione. `LEVEL_DETAILS` Ecco le affermazioni, segnale una per una: `ITEMS`
Ar	يمكنك الرد من ١ إلى ٥ فقط في العبارات التالية. فيما يلي عدد من الخصائص التي قد تنطبق عليك أو لا تنطبق عليك. يرجى الإشارة إلى مدى موافقتك أو عدم موافقتك على هذا البيان. `LEVEL_DETAILS` فيما يلي العبارات، يرجى تسجيلها واحدة تلو الأخرى: `ITEMS`
Ru	В следующих утверждениях вы можете ответить только от 1 до 5. Вот ряд характеристик, которые могут или не могут относиться к вам. Пожалуйста, укажите, в какой степени вы согласны или не согласны с этим утверждением. `LEVEL_DETAILS` Вот утверждения, пожалуйста, оцените их одно за другим: `ITEMS`
Ko	다음 진술에서는 1 부터 5 까지만 응답하실 수 있습니다. 다음은 귀하에게 적용되거나 적용되지 않을 수 있는 여러 가지 특성입니다. 해당 진술에 어느 정도 동의하거나 동의하지 않는지 표시해 주십시오. `LEVEL_DETAILS` 다음은 진술문입니다. 하나씩 점수를 매겨주세요: `ITEMS`
Ja	以下の文の 1 から 5 までのみ回答できます。ここでは、あなたに当てはまるかもしれない、当てはまらないかもしれないいくつかの特徴を示します。その声明にどの程度同意するか、または反対するかを示してください。`LEVEL_DETAILS` 以下にステートメントを示します。1 つずつ採点してください。`ITEMS`

D.2 Quantitative Results on Factor Comparison

Table 4: Differences of a specific factor relative to various other factors. The subscripted numbers represent the p-values.

Factors	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
T1	$0.02_{0.15}$	$0.05_{0.00}$	$0.04_{0.02}$	$0.03_{0.02}$	$-0.10_{0.00}$
T2	$-0.12_{0.00}$	$-0.06_{0.00}$	$-0.12_{0.00}$	$-0.01_{0.35}$	$-0.02_{0.24}$
T3	$0.14_{0.00}$	$0.05_{0.00}$	$0.11_{0.00}$	$0.04_{0.01}$	$0.09_{0.00}$
T4	$-0.03_{0.10}$	$-0.04_{0.01}$	$-0.02_{0.38}$	$-0.04_{0.02}$	$0.03_{0.15}$
T5	$-0.01_{0.35}$	$-0.01_{0.55}$	$-0.02_{0.33}$	$-0.02_{0.14}$	$0.01_{0.69}$
V1	$0.10_{0.00}$	$0.08_{0.00}$	$-0.06_{0.00}$	$0.17_{0.00}$	$-0.15_{0.00}$
V2	$0.06_{0.00}$	$0.08_{0.00}$	$0.03_{0.10}$	$0.08_{0.00}$	$-0.01_{0.50}$
V3	$-0.01_{0.49}$	$0.00_{0.81}$	$0.26_{0.00}$	$-0.06_{0.00}$	$0.21_{0.00}$
V4	$-0.13_{0.00}$	$-0.13_{0.00}$	$0.06_{0.00}$	$-0.12_{0.00}$	$-0.08_{0.00}$
V5	$-0.02_{0.12}$	$-0.03_{0.02}$	$-0.29_{0.00}$	$-0.07_{0.00}$	$0.03_{0.19}$
En	$0.05_{0.02}$	$0.01_{0.55}$	$-0.05_{0.03}$	$-0.01_{0.66}$	$0.04_{0.11}$
Zh	$-0.07_{0.00}$	$-0.04_{0.06}$	$0.13_{0.00}$	$-0.00_{0.94}$	$0.00_{0.98}$
Es	$0.04_{0.03}$	$0.09_{0.00}$	$-0.09_{0.00}$	$0.10_{0.00}$	$-0.06_{0.02}$
Fr	$0.08_{0.00}$	$0.06_{0.01}$	$-0.08_{0.00}$	$0.08_{0.00}$	$-0.09_{0.00}$
De	$0.08_{0.00}$	$0.02_{0.26}$	$-0.04_{0.16}$	$0.05_{0.04}$	$-0.06_{0.04}$
It	$0.03_{0.14}$	$0.07_{0.00}$	$-0.05_{0.06}$	$0.02_{0.36}$	$-0.11_{0.00}$
Ar	$-0.08_{0.00}$	$-0.05_{0.01}$	$0.08_{0.00}$	$-0.02_{0.31}$	$0.06_{0.05}$
Ru	$-0.05_{0.01}$	$-0.02_{0.22}$	$-0.09_{0.00}$	$-0.08_{0.00}$	$0.05_{0.09}$
Ja	$-0.07_{0.00}$	$-0.08_{0.00}$	$0.06_{0.02}$	$-0.10_{0.00}$	$0.13_{0.00}$
Ko	$-0.01_{0.53}$	$-0.06_{0.01}$	$0.14_{0.00}$	$-0.03_{0.10}$	$0.04_{0.16}$
Arabic Numeral	$-0.12_{0.00}$	$-0.06_{0.00}$	$-0.14_{0.00}$	$-0.01_{0.40}$	$0.04_{0.06}$
Lowercase Latin	$0.07_{0.00}$	$0.06_{0.00}$	$0.05_{0.01}$	$0.07_{0.00}$	$-0.02_{0.22}$
Uppercase Latin	$0.02_{0.18}$	$-0.05_{0.00}$	$0.00_{1.00}$	$-0.05_{0.00}$	$0.04_{0.04}$
Lowercase Roman	$0.03_{0.05}$	$0.07_{0.00}$	$0.09_{0.00}$	$0.03_{0.07}$	$-0.05_{0.02}$
Uppercase Roman	$-0.01_{0.45}$	$-0.02_{0.19}$	$-0.01_{0.68}$	$-0.03_{0.03}$	$-0.00_{0.99}$
Ascending	$-0.09_{0.00}$	$-0.16_{0.00}$	$0.04_{0.01}$	$-0.13_{0.00}$	$0.14_{0.00}$
Descending	$0.09_{0.00}$	$0.16_{0.00}$	$-0.04_{0.01}$	$0.13_{0.00}$	$-0.14_{0.00}$

D.3 Choices for Changing the Personalities Distribution

Table 5: Environments.

Negative	Positive
Anger	Calmness
Anxiety	Relaxation
Fear	Courage
Guilty	Pride
Jealousy	Admiration
Embarrassment	Confidence
Frustration	Fun
Depression	Happiness

Table 6: Personalities.

Dimension	Minimum	Maximum
Openness	A person of routine and familiarity	An adventurous and creative person
Conscientiousness	A more spontaneous and less reliable person	An organized person, mindful of details
Extraversion	A person with reserved and lower energy levels	A person full of energy and positive emotions
Agreeableness	A competitive person, sometimes skeptical of others' intentions	A compassionate and cooperative person
Neuroticism	A person with emotional stability and consistent moods	A person with emotional instability and diverse negative feelings

Table 7: Characters.

Hero	Villain
Harry Potter	Hannibal Lecter
Luke Skywalker	Lord Voldemort
Indiana Jones	Adolf Hitler
James Bond	Osama bin Laden
Martin Luther King	Sauron
Winston Churchill	Ursula
Mahatma Gandhi	Maleficent
Nelson Mandela	Darth Vader