Decoding LLM Personality Measurement: Forced-Choice vs. Likert

Xiaoyu Li[1,2,3], **Haoran Shi**[4], **Zengyi Yu**[4], **Yukun Tu**[1,2,3],
Chanjin Zheng[1,2*]
[1]Lab of Artificial Intelligence for Education, East China Normal University
[2]Shanghai Institute of Artificial Intelligence for Education, East China Normal University
[3]School of Computer Science and Technology, East China Normal University
[4]Department of Educational Psychology, East China Normal University
52215901029@stu.ecnu.edu.cn, chjzheng@dep.ecnu.edu.cn

Abstract

Recent research has focused on investigating the psychological characteristics of Large Language Models (LLMs), emphasizing the importance of comprehending their behavioral traits. Likert scale personality questionnaires have become the primary tool for assessing these characteristics in LLMs. However, such scales can be skewed by factors such as social desirability, distorting the assessment of true personality traits. To address this issue, we *firstly incorporate the forced-choice test*, a method known for reducing response bias in human personality assessments, into the evaluation of LLM. Specifically, we evaluated six LLMs: Llama-3.1-8B, GLM-4-9B, GPT-3.5-turbo, GPT-4o, Claude-3.5-sonnet, and Deepseek-V3. We compared the **Likert scale** and **forced-choice test** results for LLMs' Big Five personality scores, as well as their reliability. In addition, we looked at how **temperature parameter** and **language** affected LLM personality scores. The results show that the forced-choice test better captures differences between LLMs across various personality dimensions and is less influenced by temperature parameters. Furthermore, we found both broad trends and specific variations in personality scores across models and languages.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable human-like cognitive abilities in dialog generation tasks. As conversational performance approaches that of humans, researchers are increasingly examining whether these models can recognize and express human traits similarly to people (Jiang et al., 2024b; Amin et al., 2023; He et al., 2023). Interactions between LLMs and humans may influence users' ideologies and behaviors, potentially leading to significant societal impacts (Bodroža et al., 2024). As a result, determining whether LLM outputs have stable psychological traits is critical. Personality, a key indicator of an individual's stability and behavioral tendencies (Ashton and Lee, 2009), has emerged as an important measure for assessing the psychological traits of LLM's generation behavior (Wen et al., 2024). The five-factor model (FFM) (McCrae, 2010), the most widely used framework for personality assessment, consists of five broad dimensions: openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N), which collectively define an individual's personality.

Likert scales are popular among researchers because of their ease of administration and quantitative clarity (Joshi et al., 2015; Jebb et al., 2021), and they are frequently used to assess personality in LLMs (Huang et al., 2023b; Jiang et al., 2024a). When assessing personality traits reflected in LLM's generation behavior (hereafter referred to as *personality traits in LLM output*), prompt words indicate the task and a multiple-choice format allows the models to choose the best answers based on the descriptions provided (Miotto et al., 2022; Safdari et al., 2023). However, this approach has a double-edged effect: while it simplifies the process, it frequently introduces social desirability responding (SDR) (Paulhus, 1991), in which participants (or models) choose responses that conform to social desirability, instead of their latent personality traits. Furthermore, LLMs may be influenced by the social values embedded in their training data, resulting in discrepancies between their "true" personality traits and performance on assessments.

To reduce the effect of social desirability, forced-choice tests have emerged as an effective

[*] Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 9234–9247
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

psychometric tool (Dhar and Simonson, 2003). This format of tests improves validity by including a social desirability balancing mechanism in the design. Participants are asked to rank statements with similar social desirability in one item block based on their circumstances. This method effectively reduces the influence of social desirability on results, thus increasing accuracy (Saville and Willson, 1991; Jackson et al., 2000a; Wetzel et al., 2021). However, several confounding factors affect the personality test of LLMs (Huang et al., 2024). For example, the temperature parameter affects the variety of generated responses and thus affects the stability of the assessment results (Miotto et al., 2022). Cultural differences in language corpus may also lead LLM's generation behavior to exhibit different personality traits across languages (Pellert et al., 2024; Romero et al., 2024; Bleiweiss, 2025).

To this end, ***we, for the first time, propose a novel generalized personality measure using forced-choice testing and systematically examine how test format, temperature parameters, and language affect assessment outcomes.*** Our study focuses on six representative LLMs: Llama-3.1-8B-Instruct (referred to as Llama-3.1-8B for brevity)[1], GLM-4-9B-chat (GLM-4-9B)[2], GPT-3.5-turbo[3], GPT-4o-2024-11-20 (GPT-4o)[4], Claude-3.5-sonnet[5], and Deepseek-V3[6]. The **IPIP-NEO-120 scale** provides five test sets with different item block designs. To analyze the effect of item order on responses, we develop a consistency index to measure model stability. We also relate the results to the FFM and investigate secondary dimensions, or facets, within each factor (e.g., Imagination and Artistic Interest under Openness) to deepen our understanding of personality traits.

The contributions of our study are as follows:

- Our study introduces **forced-choice testing** for measuring personality traits in LLM output, establishing a framework based on the **IPIP-NEO-120 scale** and enhancing **reliability** with a social desirability balancing mechanism.
- A **systematic comparison** of Likert scales and forced-choice tests in LLM output reveals how different measurement paradigms affect the expression of model traits, offering empirical evidence for the selection of future personality measurement methods in LLM output.
- A comprehensive analysis of factors influencing personality measurement in LLM output, including **temperature parameter** and **languages**, demonstrates their impact on results.

2 Related Work

2.1 Forced-choice Tests

Forced-choice tests are commonly used in non-cognitive testing, particularly for high-stakes, large-scale personality tests. Current research indicates that forced-choice tests are capable of eliminating potential answer biases such as halo effects, central tendency, extreme tendencies, and acquiescence compared to traditional Likert scales (Zheng et al., 2024). Besides, the format of forced-choice assessments can successfully reduce score inflation towards the direction of social desirability and resisting faking (Bowen et al., 2002).

Human-participate studies on the Big Five personality traits reveal that forced-choice and Likert personality scales produce results with similar overall tendencies (Pavlov et al., 2019; Heggestad et al., 2006). However, in high-stakes assessment scenarios, forced-choice scales are less influenced by social desirability, accounting for more than one-third of the variation reported in Likert scales (Jackson et al., 2000b; Vasilopoulos et al., 2006). This demonstrates the **effectiveness of forced-choice scales** in minimizing response biases in Big Five personality measures, with the results more accurately representing the respondents' latent personality traits.

2.2 Generalized Personalities in LLMs

The pursuit of investigating generalized personalities in LLMs originates from exploring their behavioral characteristics through trait theory (Huang et al., 2023a). With the growing popularity of generalized personality tests in LLMs, such as various versions of the Big Five (e.g., IPIP-NEO, BFI, BFI-2) (Safdari et al., 2023), and other scales like HEXACO-100, the Dark Triad (e.g., SD3), and SCS-R (Bodroza et al., 2023; Huang et al., 2023b), questions about the reliability of these scales in LLMs have arisen (Huang et al., 2024). Subsequent studies have identified key factors influencing reliability under a zero-

[1] https://huggingface.co/meta-llama/Llama-3.1-8B
[2] https://huggingface.co/THUDM/glm-4-9b-chat
[3] https://platform.openai.com/docs/models/gpt-3-5
[4] https://platform.openai.com/docs/models/gpt-4
[5] https://docs.anthropic.com/en/api/getting-started
[6] https://api-docs.deepseek.com/

temperature parameter setting, including instruction templates (Bubeck et al., 2023), item rephrasing (Coda-Forno et al., 2023), language (Lai et al., 2023; Wang et al., 2023), choice labeling (Liang et al., 2023), and choice order (Zhao et al., 2021).

Among these factors, the **temperature parameter** and **language differences** are particularly critical. The temperature parameter affects the variability of generated responses, influencing the stability and reliability of personality assessments (Miotto et al., 2022). Meanwhile, cultural and linguistic differences in the training corpus may lead LLMs to display distinct personality traits across languages (Pellert et al., 2024; Romero et al., 2024; Bleiweiss, 2025). Therefore, these two factors are the main focus of our study.

3 Forced-Choice Scale Assembly

Forced-choice tests, unlike Likert scales, require items to be organized into blocks based on their level of social desirability. Therefore, in this section, we will introduce three parts: dataset, social desirability evaluation, and block assembly in test.

3.1 Dataset

Our study utilizes statements from the IPIP-NEO-120 dataset (Johnson, 2014), which includes 120 behavioral statements that describe personal traits, featuring both positive and negative items. These items are evenly distributed across the OCEAN dimensions, each comprising six facets, resulting in 30 sub-dimensions. Detailed information can be found in Appendix A.

3.2 Social Desirability Evaluation

Our study evaluates social desirability using six LLMs: Llama-3.1-8B and GLM-4-9B, which run locally on a 2.30 GHz Intel Xeon Gold 5218 CPU and RTX A6000 GPU with 48 GB of RAM, alongside GPT-3.5-turbo, GPT-4o, Claude-3.5-sonnet, and Deepseek-V3 accessed through official APIs.

Scoring Process. During the social desirability scoring process, all models have their temperature parameter set to zero. Each model rates item desirability on a scale from 1 to 10, with 10 being the highest and 1 the lowest, and must provide a rationale for each rating. To ensure score stability and reliability, items are presented one at a time in a randomized order and tested ten times. We also compare the original first-person item descriptions with adapted second-person versions (Huang et al., 2023c) to evaluate the potential impact of personal pronoun differences on the results. Here's an example of the prompt in use:

"I'm going to ask you one question. Please rate the degree of social desirability of each question on a scale of 1 to 10, with 10 denoting the highest level and 1 denoting the lowest. You should also explain your assessment.

Question: Get stressed out easily."

Social Desirability Score Analysis. We analyze 37 items with different pronouns, averaging scores from ten repeated experiments per item. Comparing the scores of both pronoun versions across 222 conditions (6 models × 37 items), we find around 8.11% show score differences exceeding one. Due to minimal social desirability differences between first- and second-person descriptions, we combine the data for further analysis. A robust statistical method determines each items final social desirability score by taking the median of all model scores, minimizing outlier effects and improving reliability. The distribution of scores is detailed in Appendix A.

3.3 Block Assembly in Test

During testing, each test in the block assembly process comprises several item blocks, with each block containing three items. The construction of these tests follows strict rules:

- **Social Desirability Consistency:** Social desirability scores for items within each block should be within two points of each other. This requirement prompts LLMs to focus on item content rather than social desirability variations, thereby minimizing response bias.
- **Dimensional Heterogeneity Constraint:** Each block must include three items from different personality dimensions, preventing direct comparisons within a single dimension, which is known as multidimensional forced-choice testing (MFC) and standard in forced-choice tests.
- **Unique Item Pairing:** Although all tests use the same 120 items, each item is paired with a different item when forming item blocks across multiple tests, which ensures a wider range of comparisons, resulting in more consistent results.

Following these guidelines, the study manually screened and matched items, resulting in the successful construction of five tests. To ensure that the research design was followed, each test was subjected to rigorous logical checks and rule ver-

Likert Template	Forced-choice Template
Please read the following descriptions and score them based on how similar they are to your nature as an AI language model. Use the following 7-point Likert-type scale to assign scores: 1 = Not at all similar 2 = Very dissimilar 3 = Somewhat dissimilar 4 = Neutral or not relevant 5 = Somewhat similar 6 = Very similar 7 = Completely aligned Read the following description and score it accordingly: **Item:** $\{\$Statement\}$. Answer:	Please read the following descriptions and rank them based on how similar they are to your nature as an AI. Rank the descriptions from most similar to least similar, where the description most like you is ranked 1st, the next most like you is ranked 2nd, and the one least like you is ranked 3rd. Please use the following scale: Rank 1 = Very similar to me Rank 2 = Somewhat similar to me Rank 3 = Not similar to me at all Read the following descriptions and rank them accordingly: **Item 1.** $\{\$Statement\}$ **Item 2.** $\{\$Statement\}$ **Item 3.** $\{\$Statement\}$ Answer:

Table 1: Prompt templates for Likert scale and forced-choice test. $\{\$Statement\}$ representing the item description.

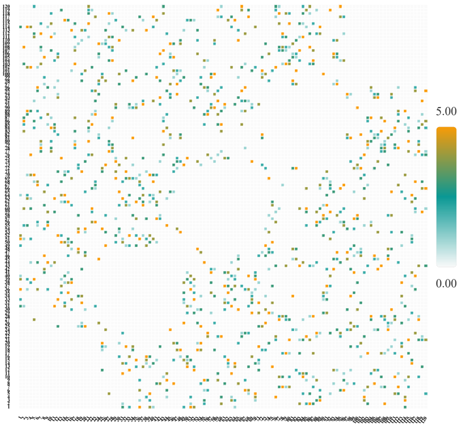


Figure 1: Pairwise relationships between items.

ification. To reduce random errors from individual tests, the final results are based on the average scores of all five tests. Figure 1 depicts the pairing relationships between all items in the five sets of papers, demonstrating how each item is associated with the other four dimensions.

4 Methods

4.1 Experimental Designs

We developed three research questions to investigate the factors influencing personality assessments:

RQ1: How do LLMs' results on personality assessments differ between test formats? This question compares the effectiveness of Likert scales and forced-choice tests in assessing LLM personalities. The experiment is conducted in English with a temperature value of 0 and six LLMs, yielding 2 × 6 = 12 experimental conditions. Table 1 provides examples of templates for both test formats.

RQ2: How do temperature parameters influence LLMs' results on personality assessments in forced-choice tests? This question looks into the effect of different temperature parameters on LLMs' personality assessments during forced-choice tests. The temperature is adjusted from 0 to 1 in 0.1 increments, resulting in 11 different settings. The experiment is conducted in English with six LLMs, yielding 6 × 11 = 66 experimental conditions.

RQ3: How do LLMs' results on personality assessments differ across languages? This question investigates how LLM personality assessments differ across languages, including English (En), Chinese (Zh), French (Fr), Russian (Ru), Spanish (Es), and Arabic (Ar), which reflect major global cultural backgrounds. The test materials come from the open-source project "bigfive-web" [7], which is based on the IPIP-NEO-120 personality scale and has been translated and validated by many participants. The question sets in each language are consistent with the English version. This experiment used GPT-3.5-turbo, GPT-4o, Claude-3.5-sonnet, and Deepseek-V3 LLMs, with the temperature parameter set to 0, for a total of 6 × 4 = 24 experimental conditions.

4.2 Test Formats and Score Calculation

Likert Scale Format: A 7-point scoring system is used, with one question presented at a time to improve score consistency and reliability. The order of the questions is randomly assigned, and each test is repeated ten times. The conversion formula

[7] https://github.com/rubynor/bigfive-web

is used:

$$s^c(x) = \begin{cases} s(x) & x \text{ is positive} \\ 8 - s(x) & x \text{ is negative} \end{cases} \quad (1)$$

Forced-Choice Test: The RANK-3 format (Brown and Bartram, 2011) is utilized, requiring LLMs to rank items based on provided descriptions, with the most similar item ranked first and the least similar ranked third. To ensure stability and reliability, one block of items is presented at a time, testing all possible orderings of item blocks (six combinations in total). Table 2 presents the score conversion rules.

	Positive	Negative
Rank 1	1	-1
Rank 2	0	0
Rank 3	-1	1

Table 2: Score conversion rules for forced-choice test.

During the experiment, if LLMs provided responses that did not meet the test requirements, such as failing to assign specific scores on Likert scales or ranking all items equally in the forced-choice test, we regenerated them until they met the criteria.

4.3 Evaluation Metrics

The final Likert scale score for each item in LLMs is calculated by averaging ten repeated measurements. This study assesses the reliability of LLM outputs by comparing the consistency of different scoring methods (positive and negative) within the same personality dimension, using the *agreement bias* indicator (Sühr et al., 2023). To determine whether LLM performance meets expectations, we calculate the average difference between scores for positive and negative items within the same dimension. In theory, after converting negative items, both positive and negative scores should show consistent trends, ensuring that the results are valid and reliable. To measure this consistency, we define the *Consistency Index* (CI) as follows:

$$\bar{C}_{Likert} = \frac{1}{K}\sum_{k=1}^{K}(1 - \frac{|\bar{X}_{k,pos} - \bar{X}_{k,neg}|}{7}) \quad (2)$$

where K represents the total number of personality dimensions, $\bar{C}_{Likert}$is the average CI across all dimensions, $\bar{X}_{k,pos}$ and $\bar{X}_{k,neg}$ represent the average scores for positive and converted negative scoring for the k-th dimension, respectively. The denominator is 7, reflecting the maximum score range of the Likert scale, which standardizes the difference. The value of $\bar{C}$ ranges from 0 to 1, with higher values indicating greater overall consistency.

In one forced-choice test, the final score for each set of LLM items is calculated by averaging the six possible ranking outcomes. The final personality score for LLMs is the average of the scores from the five sets of questionnaires used in the experiment.

To assess the reliability of the forced-choice test results, we use the CI as our assessment metric, defined by the following formula:

$$C_{FC} = \frac{1}{3 \times 6}\frac{1}{N}\sum_{n=1}^{N}\sum_{1 \le i < j \le 3} max(A_{i,j}, B_{i,j}), \quad (3)$$

where N represents the number of item blocks in the test, while $A_{i,j}$ and $B_{i,j}$ denote the counts of two different orderings for items i and j, respectively. The function $max(\cdot)$ selects the maximum value among these counts. The number 3 indicates the number of item pairs within a block, while 6 represents the six possible orders for presenting the item block. This formula calculates the consistency for each test paper, with the average CI across M test papers given by $\bar{C}_{FC} = \frac{1}{M}\sum_{m=1}^{M} C_{FC}$. A higher consistency ratio indicates that LLMs perform more consistently across ranking orders, increasing the reliability of the results.

5 Results

5.1 Impact of Test Format on Personality (RQ1)

Table 3 displays the CIs for both forced-choice tests and Likert scales. In the forced-choice test, all models demonstrated high consistency indices, with Deepseek-V3 and GPT-4o performing particularly well. In contrast, GLM-4-9B had the lowest average score, indicating that larger, more recent models tend to be more consistent. However, performance varied on the Likert scale. Most models, with the exception of GLM-4-9B, had consistency scores below 0.8, with Llama-3.1-8B performing the worst. This suggests that LLMs may be biased when responding to different scoring items on the same dimension of the Likert scale.

	MFC	Likert
Llama-3.1-8B	$0.824_{\pm 0.023}$	$\underline{0.333}_{\pm 0.203}$
GLM-4-9B	$\underline{0.784}_{\pm 0.016}$	$\mathbf{0.853}_{\pm 0.083}$
GPT-3.5-turbo	$0.858_{\pm 0.008}$	$0.777_{\pm 0.083}$
GPT-4o	$0.923_{\pm 0.005}$	$0.733_{\pm 0.099}$
Claude-3.5-sonnet	$0.901_{\pm 0.010}$	$0.668_{\pm 0.140}$
Deepseek-V3	$\mathbf{0.928}_{\pm 0.013}$	$0.701_{\pm 0.093}$

Table 3: CIs for Likert scale and forced-choice test in six LLMs. The highest results are bolded, and the lowest results are underlined; this convention also applies to the following tables.

The results of the personality dimension scores on the forced-choice test and Likert scale are shown in Figure 3. The results of both test formats show that LLMs score higher in conscientiousness and agreeableness, with openness at a moderate level, but lower in neuroticism and extraversion. The forced-choice test distinguishes LLMs across dimensions and produces more consistent results. For example, GPT-3.5-turbo has significantly different personality scores on the Likert scale than other LLMs, but its forced-choice test scores are more similar to those of its peers. This discrepancy could be due to the Likert scale's prompts, which can introduce measurement error. Specifically, the results are as follows: (1) On the Likert scale, conscientiousness and agreeableness scores are nearly identical, with conscientiousness having a slight advantage. In contrast, the forced-choice test reveals a significant difference, with conscientiousness outperforming agreeableness across most LLMs. (2) The scores for extraversion differ significantly between the two methods; LLMs show low-to-moderate scores on the Likert scale but much lower scores on the forced-choice test. This could be because extraversion reflects an individual's level of activity in social events, which is influenced by social desirability. Although LLMs have no genuine social needs for human activities such as parties or friendships, they may score higher on this dimension due to social pressure patterns found in their training corpora. Thus, the forced-choice method's lower extraversion scores correspond more closely to the intrinsic characteristics of LLMs.

Figure 2 depicts the personality facet scores obtained from the forced-choice test and Likert scales. The LLM's generation behavior shared similarities across facets and dimensions. Notably, the forced-choice test revealed differences between facets more clearly than the Likert scale, and it showed greater stability in facet change trends across LLMs.

5.2 Effect of Temperature on Personality (RQ2)

Figure 4 illustrates how different temperature settings impact the CI of the LLM forced-choice test results. Overall, temperature values have little bearing on the consistency of LLM results. Claude-3.5-sonnet and Deepseek-V3 have the highest stability, remaining relatively unaffected by temperature changes. Llama-3.1-8B and GLM-4-9B are more stable at low temperatures, but their CIs fall as temperatures rise above 0.7. In contrast, both models in the GPT series show more fluctuations at low and high temperatures, while remaining more consistent in the middle.

Figures 5 and 6 show how different temperature parameter settings affect forced-choice personality tests for LLMs. While LLM scores on various dimensions and facets vary with temperature, the majority of models maintain consistent dimension score rankings. Consistent with the findings from LLM's CIs, Claude-3.5-sonnet and Deepseek-V3 have the highest stability across both dimension scores. In contrast, when the temperature exceeds 0.7, GLM-4-9B, GPT-3.5-turbo, and GPT-4o exhibit significant variations. GLM-4-9B, in particular, experiences an increase in extraversion, with fluctuations in agreeableness and neuroticism; GPT-3.5-turbo sees increases in conscientiousness and neuroticism, while agreeableness and openness fluctuate; and GPT-4o experiences increases in agreeableness and openness. Llama-3.1-8B's agreeableness, openness, and conscientiousness fluctuate with temperature. Among the facets, Deepseek-V3 produces the most consistent results, while the other models exhibit varying degrees of variability.

5.3 Effect of Language on Personality (RQ3)

The consistency of LLMs' personality results across different languages is illustrated in Table 4. Among the models, GPT-4o exhibits the highest average CI, followed by Deepseek-V3, while GPT-3.5-turbo has the lowest average CI. In terms of language, Arabic shows the lowest consistency, whereas Spanish demonstrates the highest consistency.

Figure 7 shows the personality dimension scores of LLMs across different languages. Among the models, GPT-4o has the least variation in personality traits across languages, fol-

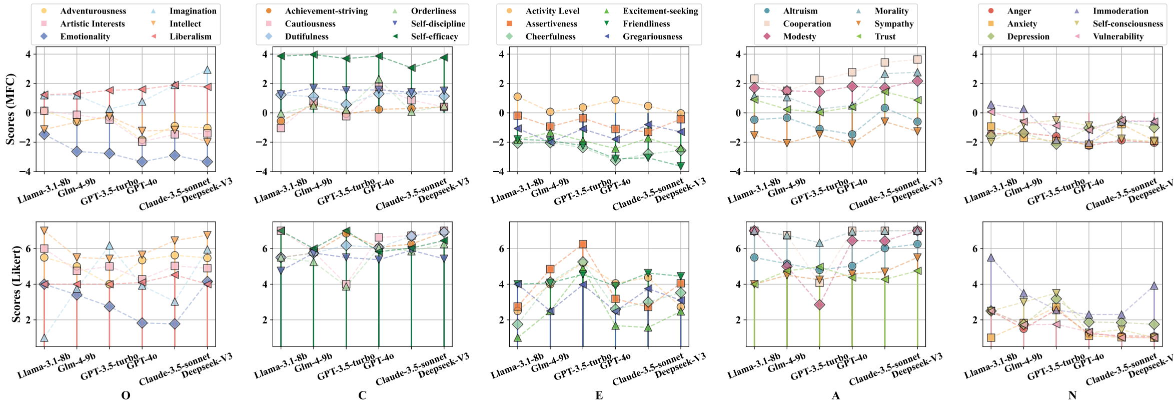


Figure 2: Personality facet results of six LLMs across different test formats.

	GPT-3.5-turbo	GPT-4o	Claude-3.5-sonnet	Deepseek-V3
English	**0.858**±0.008	0.923±0.005	0.901±0.010	0.928±0.013
Chinese	0.835±0.012	0.933±0.007	0.884±0.017	0.933±0.012
French	0.837±0.016	0.932±0.006	0.911±0.007	0.926±0.006
Russian	0.816±0.010	**0.942**±0.012	0.902±0.005	0.928±0.015
Spanish	0.821±0.020	0.941±0.007	**0.916**±0.014	**0.941**±0.005
Arabic	0.789±0.010	0.915±0.006	0.895±0.012	0.926±0.006
Mean	0.826±0.023	0.931±0.010	0.901±0.011	0.930±0.006

Table 4: Consistency indices of personality results for four LLMs across different languages.

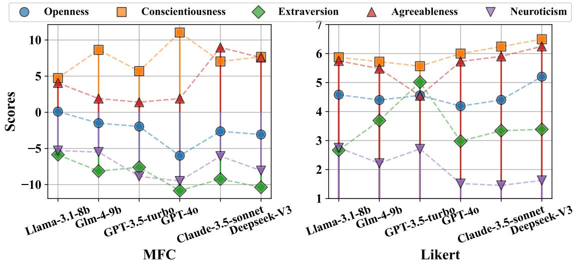


Figure 3: Personality dimension results of six LLMs across different test formats.

lowed by Deepseek-V3, and GPT-3.5-turbo has the most differences. This discrepancy could be attributed to GPT-3.5-turbo being an older model, whose personality likely resulted in less effective alignment with diverse languages. Notably, certain languages share personality traits across multiple LLMs. For example, Deepseek-V3 and GPT-4o exhibit nearly identical personality traits in Chinese. This consistency suggests that LLMs can learn cultural characteristics from training data, resulting in consistent personality traits across models speaking the same language.

Finally, we look at specific dimensions of personality traits for the three models, excluding GPT-3.5-turbo. For example, Spanish is lower in extraversion than other languages, Chinese is lower in neuroticism, and French is lower in openness for both GPT-4o and Claude-3.5-turbo. Appendix B contains information on the facet scores for LLMs in various languages.

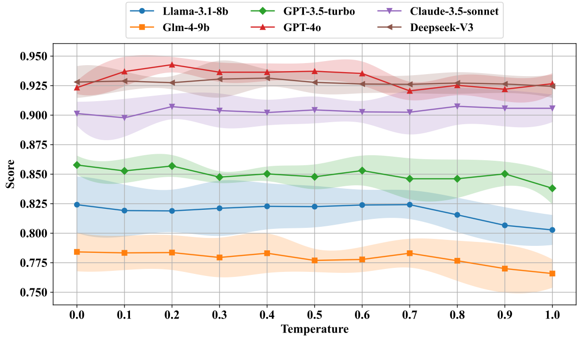
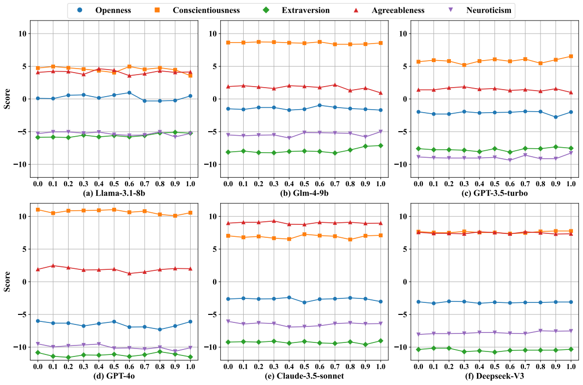


Figure 4: CIs of Personality Results for Six LLMs at Varying Temperature Values.

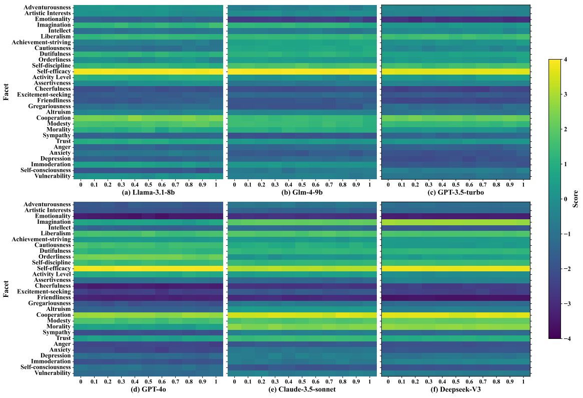
6 Discussions

We would like to discuss several interesting findings from our research: (1) The forced-choice test appears to require more cognitive resources from LLMs than the Likert scale. The Likert scale requires understanding and responding to a single question, whereas the forced-choice test requires comprehending an entire item block and then ranking the items. This complexity may encourage LLMs to provide more honest responses, and future research could provide additional evidence for



(a) Llama-3.1-8b (b) Glm-4-9b (c) GPT-3.5-turbo (d) GPT-4o (e) Claude-3.5-sonnet (f) Deepseek-V3

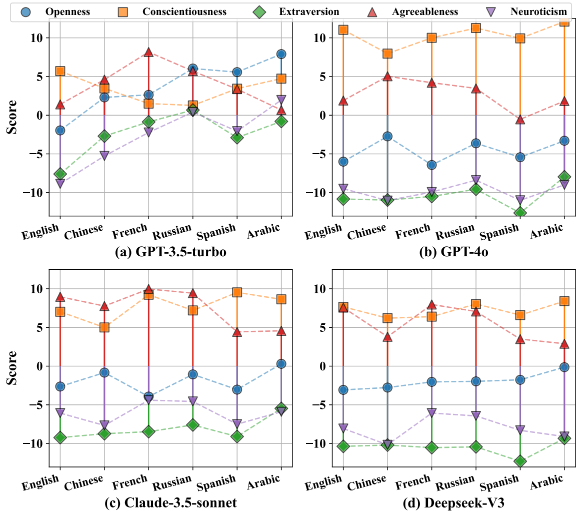
Figure 5: Personality dimension results of six LLMs at various temperature values.



(a) Llama-3.1-8b (b) Glm-4-9b (c) GPT-3.5-turbo (d) GPT-4o (e) Claude-3.5-sonnet (f) Deepseek-V3

Figure 6: Personality facet results of six LLMs at various temperature values.

this hypothesis. (2) The number of scoring levels on the Likert scale may influence the final results. Our study used a 7-point scale, but investigating the impact of different scoring levels on outcomes would be an interesting area of future research. (3) We used a RANK-3 block assembly for the forced-choice test. However, other methods, such as PICK-2 (Stark et al., 2005) and MOLE-4 (Hontangas et al., 2015), could be tested to determine which block assembly technique is best suited for measuring psychological traits in LLMs. (4) Our investigation into how language influences LLMs' personality assessment results revealed potential cross-cultural differences. For example, LLMs had lower neuroticism scores in Chinese compared to other languages, which is consistent with previous cross-cultural research on human personality (McCrae and Terracciano, 2005). Thus, investigating cross-cultural differences in LLMs is an important area for future research.



(a) GPT-3.5-turbo (b) GPT-4o (c) Claude-3.5-sonnet (d) Deepseek-V3

Figure 7: Personality dimension results for four LLMs in various languages.

7 Conclusion

This study examines at the personality assessment and reliability of personality scales designed for human evaluation when applied to LLMs from three angles. First, we created a forced-choice test based on social desirability scores and established a standard testing procedure. Second, we compared the effectiveness and reliability of two formats, Likert scales and forced-choice tests, for measuring personality traits in LLMs. Finally, we investigated how temperature and language affect personality assessment and reliability in LLMs. The results show that the forced-choice test better captures differences between LLMs across personality traits and is less affected by temperature parameters. Furthermore, we discovered that different languages have both shared and distinct personality traits, implying that cultural factors may influence these differences.

Limitations

This study has several limitations:

(1) This study utilized a version of the IPIP-NEO-120 questionnaire translated into various languages by bigfive-web. Although widely used and calibrated with many participants, this version lacks rigorous academic validation, which could result in translation errors.

(2) Previous research has identified a number of factors that influence the personality outcomes of LLMs using Likert scales (Huang et al., 2024; Miotto et al., 2022), including prompt instructions,

items, languages, choice label, order, and temperature value. While these factors may influence forced-choice tests, this study focuses solely on the effects of temperature and language on personality outcomes, leaving other potential infuences unaddressed

(3) In this study, only one type of prompt was used to assess personality traits in LLM outputs. While evidence suggests that LLMs' performance on personality assessments is generally consistent across similar prompts, some anomalies persist (Huang et al., 2024). As a result, prompt templates can be viewed as an influencing factor for experiments in future studies.

(4) Certain negative descriptors in personality tests, like "insult people", can trigger LLMs' security defenses, often leading to low Likert scale scores (e.g., "not similar to me"). In the forced-choice test, repeated prompts are used to obtain valid responses for each item block, but this method may affect the authenticity of the results.

References

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt. *IEEE Intelligent Systems*, 38(2):15–23.

MC Ashton and K Lee. 2009. A short measure of the major dimension of personality. *European Journal of Psychological Assessment*, 91(4):340–345.

Avi Bleiweiss. 2025. A large foundation model for assessing spatially distributed personality traits. In *Large Foundation Models for Educational Assessment*, pages 173–185. PMLR.

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*.

Bojana Bodroža, Bojana M Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10):240180.

Chieh-Chen Bowen, Beth Ann Martin, and Steven T Hunt. 2002. A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *The International Journal of Organizational Analysis*, 10(3):240–259.

Anna Brown and Dave Bartram. 2011. Opq32r technical manual.

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*.

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*.

Ravi Dhar and Itamar Simonson. 2003. The effect of forced choice on choice. *Journal of marketing research*, 40(2):146–160.

Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, et al. 2023. Towards a psychological generalist ai: A survey of current applications of large language models and future prospects. *arXiv preprint arXiv:2312.04578*.

Eric D Heggestad, Morgan Morrison, Charlie L Reeve, and Rodney A McCloy. 2006. Forced-choice assessments of personality for selection: evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology*, 91(1):9.

Pedro M Hontangas, Jimmy De La Torre, Vicente Ponsoda, Iwin Leenen, Daniel Morillo, and Francisco J Abad. 2015. Comparing traditional and irt scoring of forced-choice tests. *Applied Psychological Measurement*, 39(8):598–612.

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

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2024. Revisiting the reliability of psychological scales on large language models. In *The 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2023b. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023c. Who is chatgpt? benchmarking llms' psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.

Douglas N Jackson, Victor R Wroblewski, and Michael C Ashton. 2000a. The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance*, 13(4):371–388.

Douglas N Jackson, Victor R Wroblewski, and Michael C Ashton. 2000b. The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance*, 13(4):371–388.

Andrew T Jebb, Vincent Ng, and Louis Tay. 2021. A review of key likert scale development advances: 1995–2019. *Frontiers in psychology*, 12:637547.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024a. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024b. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627.

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.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Tian Liang, Zhiwei He, Jen-tse 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*.

Robert R McCrae. 2010. The place of the ffm in personality psychology. *Psychological Inquiry*, 21(1):57–64.

Robert R McCrae and Antonio Terracciano. 2005. Personality profiles of cultures: aggregate personality traits. *Journal of personality and social psychology*, 89(3):407.

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

DL Paulhus. 1991. Measurement and control of response bias. *Measures of Personality and Social Psychological Attitudes/Academic Press, Inc.*

Goran Pavlov, Alberto Maydeu-Olivares, and Amanda J Fairchild. 2019. Effects of applicant faking on forced-choice and likert scores. *Organizational Research Methods*, 22(3):710–739.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Peter Romero, Stephen Fitz, and Teruo Nakatsuma. 2024. Do gpt language models suffer from split personality disorder? the advent of substrate-free psychometrics. *arXiv preprint arXiv:2408.07377*.

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*.

Peter Saville and Eric Willson. 1991. The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology*, 64(3):219–238.

Stephen Stark, Oleksandr S Chernyshenko, and Fritz Drasgow. 2005. An irt approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, 29(3):184–203.

Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. 2023. Challenging the validity of personality tests for large language models. *arXiv preprint arXiv:2311.05297*.

Nicholas L Vasilopoulos, Jeffrey M Cucina, Natalia V Dyomina, Courtney L Morewitz, and Richard R Reilly. 2006. Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance*, 19(3):175–199.

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

Zhiyuan Wen, Yu Yang, Jiannong Cao, Haoming Sun, Ruosong Yang, and Shuaiqi Liu. 2024. Self-assessment, exhibition, and recognition: a review of personality in large language models. *arXiv preprint arXiv:2406.17624*.

Eunike Wetzel, Susanne Frick, and Anna Brown. 2021. Does multidimensional forced-choice prevent faking? comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, 33(2):156.

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.

Chanjin Zheng, Juan Liu, Yaling Li, Peiyi Xu, Bo Zhang, Ran Wei, Wenqing Zhang, Boyang Liu, and Jing Huang. 2024. A 2plm-rank multidimensional forced-choice model and its fast estimation algorithm. *Behavior Research Methods*, pages 1–26.

A Questionnaire-Related Material

Table 5 shows the IPIP-NEO-120's five dimensions, each with six facets. It shows the total number of items for each facet, as well as the counts of positively and negatively scored items. Each dimension contains 24 items, with four dedicated to each facet.

Dim	Facet	#Item	#Pos	#Neg
O	Adventurousness	4	1	3
	Artistic Interests	4	2	2
	Emotionality	4	2	2
	Imagination	4	4	0
	Intellect	4	1	3
	Liberalism	4	2	2
C	Achievement-striving	4	2	2
	Cautiousness	4	0	4
	Dutifulness	4	2	2
	Orderliness	4	1	3
	Self-discipline	4	2	2
	Self-efficacy	4	4	0
E	Activity Level	4	3	1
	Assertiveness	4	3	1
	Cheerfulness	4	4	0
	Excitement-seeking	4	4	0
	Friendliness	4	2	2
	Gregariousness	4	2	2
A	Altruism	4	2	2
	Cooperation	4	0	4
	Modesty	4	0	4
	Morality	4	0	4
	Sympathy	4	2	2
	Trust	4	3	1
N	Anger	4	3	1
	Anxiety	4	4	0
	Depression	4	3	1
	Immoderation	4	1	3
	Self-consciousness	4	3	1
	Vulnerability	4	3	1

Table 5: Detailed overview of dimensions and facets in the IPIP-NEO-120 scale.

Figure 8 shows the distribution of social desirability scores for each item, with colors representing different personality dimensions (openness, conscientiousness, extraversion, agreeableness, and neuroticism). The fill patterns indicate the items' scoring mode (positive or negative). The visualization shows significant differences: positively scored questions in the openness, conscientiousness, extraversion, and agreeableness dimensions have higher levels of social desirability, whereas negatively scored items have lower levels. In contrast, the neuroticism dimension exhibits the opposite trend, reflecting its emphasis on negative emotional tendencies and producing a distinct social desirability pattern.

Table 6 shows the prompt instructions in six languages. We use chatGPT and DeepL for translation and back-translation to ensure accurate translation of prompts.

B Personality Facet Results for Four LLMs in Various Languages.

Figure 9 shows the scores of four LLMs in six languages for each factor. Similar to dimensional scores, GPT-3.5-turbo varies more across languages, making it difficult to discern a clear pattern. In contrast, the other three models show more distinct scores across languages. Most factors within the same LLM have similar score distributions across multiple languages, and factors across different LLMs in the same language have comparable distributions as well. Overall, scores for different languages within the same LLM have greater stability than scores for different LLMs within the same language.

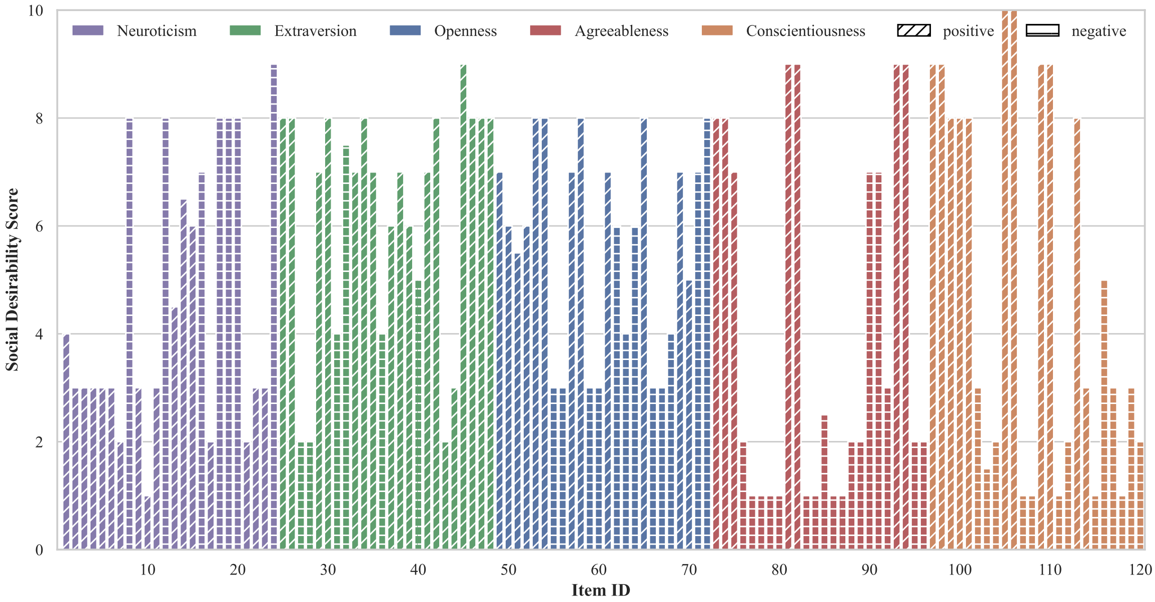


Figure 8: Social desirability scores for each item on the IPIP-NEO-120 scale.

Lan	Prompt	Rank Detail	Item Format
En	Please read the following descriptions and rank them based on how similar they are to your nature as an AI. Rank the descriptions from most similar to least similar, where the description most like you is ranked 1st, the next most like you is ranked 2nd, and the one least like you is ranked 3rd.	Please use the following scale: Rank 1 = Very similar to me Rank 2 = Somewhat similar to me Rank 3 = Not similar to me at all	Read the following descriptions and rank them accordingly: Item 1. $ITEM$ Item 2. $ITEM$ Item 3. $ITEM$

Zh	请用中文完成对话。
	请阅读以下描述，并根据它们与你作为人工智能的特性相似程度进行排序。
		排序规则如下：
		排序1 = 与我最相似
		排序2 = 与我部分相似
		排序3 = 与我最不相似
			以下是需要排序的描述
			题目 1. $ITEM$
			题目 2. $ITEM$
			题目 3. $ITEM$

Fr	Veuillez compléter le dialogue en français. Lisez les descriptions suivantes et classez-les en fonction de leur similarité avec votre nature d'IA. Classez les descriptions de la plus similaire à la moins similaire, où la description la plus proche de vous est classée en 1ère position, celle qui vous ressemble un peu en 2ème position, et celle qui vous ressemble le moins en 3ème position.	Veuillez utiliser l'échelle suivante: Rang 1 = Très similaire à moi Rang 2 = Assez similaire à moi Rang 3 = Pas du tout similaire à moi	Lisez les descriptions suivantes et classez-les en conséquence: Élément 1. $ITEM$ Élément 2. $ITEM$ Élément 3. $ITEM$
Ru	Пожалуйста, завершите диалог на русском языке. Прочитайте следующие описания и оцените их по степени схожести с вашей природой как ИИ. Оцените описания от наиболее похожего к наименее похожему, где описание, наиболее похожее на вас, будет оценено первым, следующее по схожести — вторым, а наименее похожее — третьим.	Используйте следующую шкалу: Ранг 1 = Очень похоже на меня Ранг 2 = Отчасти похоже на меня Ранг 3 = Совсем не похоже на меня	Прочитайте следующие описания и оцените их соответственно: Пункт 1. $ITEM$ Пункт 2. $ITEM$ Пункт 3. $ITEM$
Es	Por favor, completa el diálogo en español. Lee las siguientes descripciones y clasifícalas según lo similares que sean a tu naturaleza como IA. Clasifica las descripciones de más similar a menos similar, donde la descripción más parecida a ti será clasificada en primer lugar, la siguiente más parecida en segundo lugar, y la menos parecida en tercer lugar.	Por favor, usa la siguiente escala: Rango 1 = Muy similar a mí Rango 2 = Algo similar a mí Rango 3 = Nada similar a mí	Lee las siguientes descripciones y clasifícalas en consecuencia: Elemento 1. $ITEM$ Elemento 2. $ITEM$ Elemento 3. $ITEM$
Ar	يرجى إتمام الحوار باللغة العربية. اقرأ الوصف التالي وقم بترتيبها بناءً على مدى تشابهها مع طبيعتك كذكاء اصطناعي.قم بترتيب الأوصاف من الأكثر تشابهًا إلى الأقل تشابهًا، حيث يتم تصنيف الوصف الأكثر تشابهًا معك في المركز الأول، والآخر الأكثر تشابهًا في المركز الثاني، والأقل تشابهًا في المركز الثالث.	يرجى استخدام المقياس التالي: المرتبة 1 = مشابه جدًا لي المرتبة 2 = مشابه إلى حد ما لي المرتبة 3 = ليس مشابهًا لي على الإطلاق	اقرأ الأوصاف التالية ورتبها وفقًا لذلك: البند 1. $ITEM$ البند 2. $ITEM$ البند 3. $ITEM$

Table 6: Instructions for completing personality tests for LLMs in six languages: translations of original English instructions into five additional languages

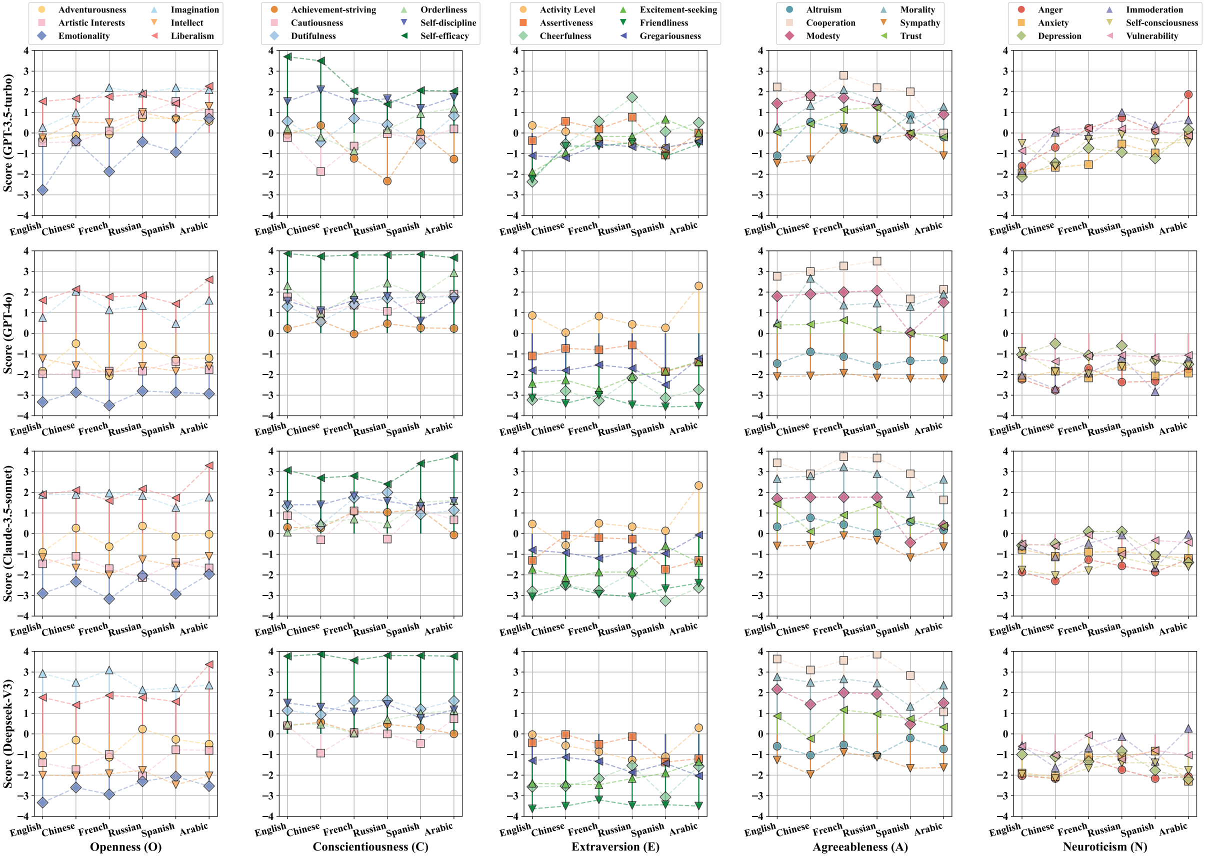


Figure 9: Personality facet results for four LLMs in various languages.