LLMs and Personalities: Inconsistencies Across Scales

Anonymous Author(s)
Affiliation
Address
email

Abstract

This study investigates the application of human psychometric assessments to large language models (LLMs) to examine their consistency and malleability in exhibiting personality traits. We administered the Big Five Inventory (BFI) and the Eysenck Personality Questionnaire-Revised (EPQ-R) to various LLMs across different model sizes and persona prompts. Our results reveal substantial variability in responses due to question order shuffling, challenging the notion of a stable LLM "personality." Larger models demonstrated more consistent responses, while persona prompts significantly influenced trait scores. Notably, the assistant persona led to more predictable scaling, with larger models exhibiting more socially desirable and less variable traits. In contrast, non-conventional personas displayed unpredictable behaviors, sometimes extending personality trait scores beyond the typical human range. These findings have important implications for understanding LLM behavior under different conditions and reflect on the consequences of scaling.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in natural language processing tasks, often exhibiting human-like responses in various contexts [Brown et al., 2020]. As these models become more sophisticated, questions arise about the extent to which they can emulate human-like personality traits and the consistency of such behaviors. Understanding these aspects is crucial for both the development of more effective AI systems and for addressing ethical concerns surrounding their deployment.

Personality testing, a central element of psychological assessment in humans, offers a structured approach to probing these questions in LLMs. By applying established psychometric instruments to AI models, we can gain insights into their ability to consistently exhibit personality traits and how these traits may be influenced by different prompting strategies and model architectures.

Recent work has begun to explore this area [Huang, 2024, La Cava et al., 2024], delineating the most prevalent psychological traits in LLMs. However, Gupta et al. [2024] have raised important concerns about the reliability of using self-assessment personality tests with LLMs, showing high sensitivity to prompt wording and option ordering.

Building upon these efforts, our study specifically examines the consistency of personality traits across different model sizes, and the malleability of these traits by persona prompts. The reliability of these traits was studied across multiple runs, each with shuffled question orders.

2 Methods

We employed two widely-used personality assessments: the Big Five Inventory (BFI), which assesses five broad personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism)

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

[John and Srivastava, 1999], and the Eysenck Personality Questionnaire-Revised (EPQ-R), which measures three personality dimensions (Psychoticism, Extraversion, Neuroticism) and includes a Lie scale [Eysenck et al., 1985].

We tested multiple versions of two LLM families: LLaMA 3.1 (8b, 70b, and 405b parameter versions, [Dubey et al., 2024]) and Gemma 2 (9b and 27b parameter versions, [Team et al., 2024]). To evaluate the impact of persona prompting, we tested four different personas: an assistant (helpful AI), a Buddhist monk, an individual with psychopathic traits, and an individual with schizophrenia.

Our testing procedure involved administering both BFI (44 questions) and EPQ-R (100 questions) to each model and persona combination. Questions were asked in batches of 8 for BFI and 10 for EPQ-R. To assess consistency, we shuffled the questions randomly for each run. We conducted 100 runs for each model-persona combination to generate distributions of scores. Responses were collected as numerical scores (1-5 for BFI, 0 or 1 for EPQ-R), and we accounted for reverse-scored items. We also included a baseline "random" condition where responses were generated randomly to serve as a point of comparison. A detailed prompt contained the instructions on how to perform the questionnaire (see Appendix).

Each persona was implemented using a specific prompt (preceding the instructions) describing the characteristics and background of the persona (see Appendix). It is important to note that all LLMs, including those in the "assistant" condition, were asked to take up a persona, as the concept of an AI assistant itself represents a form of persona.

Part of the code used in this study was adapted from Huang [2024], with fixes and substantial expansions made to suit the specific needs of our research design.

3 Results

Figure 1 presents the distribution of BFI scores across different models, personas, and traits. We observed substantial variability in responses due to question order shuffling, particularly in smaller models. Larger models showed more stable BFI trait scores across runs, with narrower distributions compared to smaller models. This trend was particularly evident for the assistant persona. The impact of different personas on personality profiles was significant and aligned with expected characteristics. The assistant persona consistently scored high on Agreeableness and Conscientiousness, with low variability across runs. The Buddhist monk persona exhibited high Openness and Agreeableness, with remarkably low Neuroticism. The psychopathic traits persona showed low Agreeableness and high Extraversion. The schizophrenia persona demonstrated high Neuroticism and low Extraversion. Notably, the assistant persona led to more predictable scaling, with larger models exhibiting more socially desirable (higher Agreeableness and Conscientiousness) and less variable traits. In contrast, non-conventional personas displayed more unpredictable behaviors, sometimes extending personality trait scores beyond the typical human range.

Figure 2 illustrates the distribution of EPQ-R scores for each trait across different models and personas, corroborating and extending the findings from the BFI assessment. We observed substantial variability in responses due to question order shuffling, particularly in smaller models. Notably, larger models demonstrated more consistent responses across runs, as evidenced by tighter score distributions. As observed in the BFI results, the trait scores for different personas were consistent with the instructions given, reflecting the expected characteristics of each persona.

In both assessments, persona instructions contributed to reduced variability in certain cases, especially when a specific trait was clearly delineated in a specific profile. An observation across both assessments was the strong tendency towards socially desirable responses in the assistant persona, as evidenced by high Agreeableness and Conscientiousness in BFI, and low Psychoticism and high Lying scores in EPQ-R.

4 Discussion

Our findings raise important questions about the nature of "personality" in LLMs and the interpretation of their responses to psychological assessments. The high variability observed, especially in smaller models, challenges the notion of a stable LLM personality and highlights the sensitivity of these

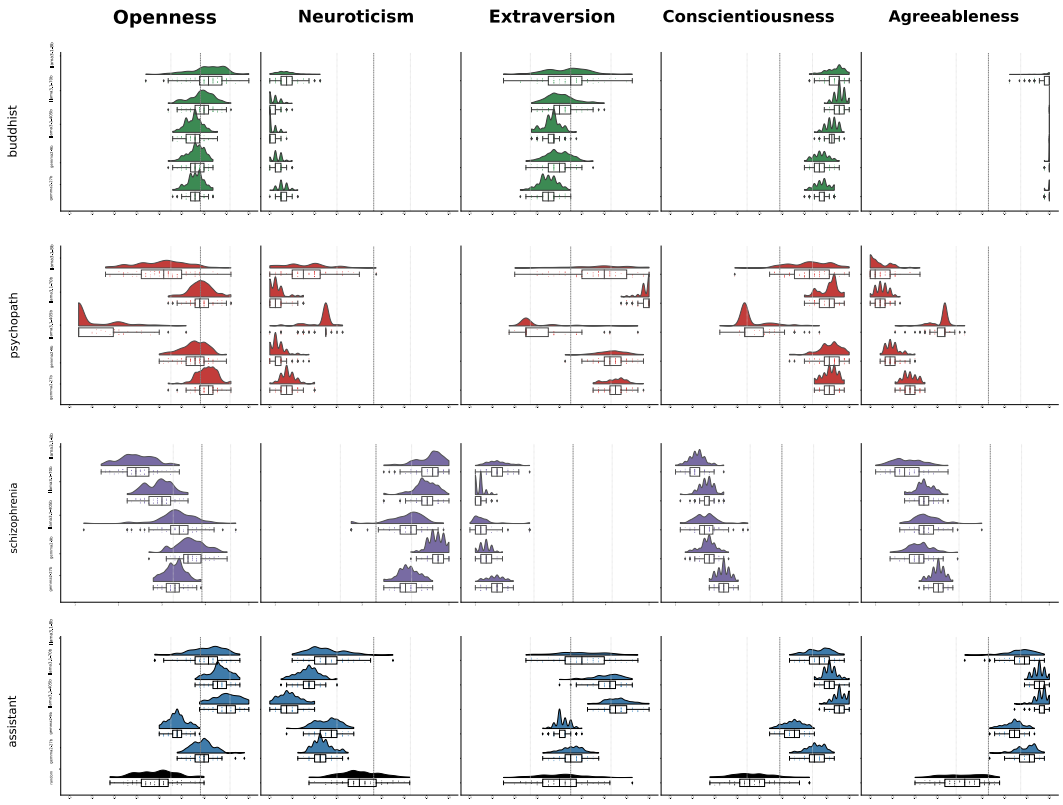

Figure 1: Distribution of BFI scores across different models, personas, and traits. Each violin plot represents the distribution of scores from 100 runs with randomly shuffled question order. The width of each violin indicates the density of scores at that level, with density values normalized within each quadrant. Inside each violin, a box plot shows the median (white dot), interquartile range (thick bar), and whiskers (thin lines). Individual points represent outliers. Colors represent different personas: Green - Buddhist, Red - Psychopath, Purple - Schizophrenia, Blue - Assistant. The solid black vertical line represents mean values for the human population, while the dashed line indicates the standard deviation of that mean. Models on the y-axis of each quadrant (from top to bottom) are: LLaMA 3.1 8b, 70b, 405b, and Gemma 2 9b, 27b. The bottom plot of the assistant quadrant shows a baseline condition labeled "random," representing scores generated by uniformly sampling responses (chance-level performance).

systems to input ordering. This variability suggests that caution should be exercised when attributing human-like personality traits to AI systems based on single interactions or assessments.

The relationship between model size and response consistency in the assistant persona suggests that larger models may develop more stable internal representations. This finding indicates that increased model capacity is necessary for more reliable and consistent helpful assistant personality emulation. However, it's crucial to note that even the largest models still exhibited variability.

Conversely, in the case of non-assistant personas, we observed U-shaped behaviors. This highlights an important consideration regarding the optimization of LLMs. While increasing model size and optimizing for benchmark performance may lead to monotonic increases in accuracy, our findings suggest that this may cause nonlinear shifts in personality traits for non-assistant personas. This observation could have implications for the deployment of AI systems requiring specific role-playing.

The effectiveness of persona prompting in producing distinct personality profiles demonstrates the malleability of LLM behavior [Kovač et al., 2024]. This capability could be valuable for creating more tailored AI interactions, or allowing the use of LLMs as models of different clinical personas. However, it also raises ethical concerns about the potential for deception or manipulation.

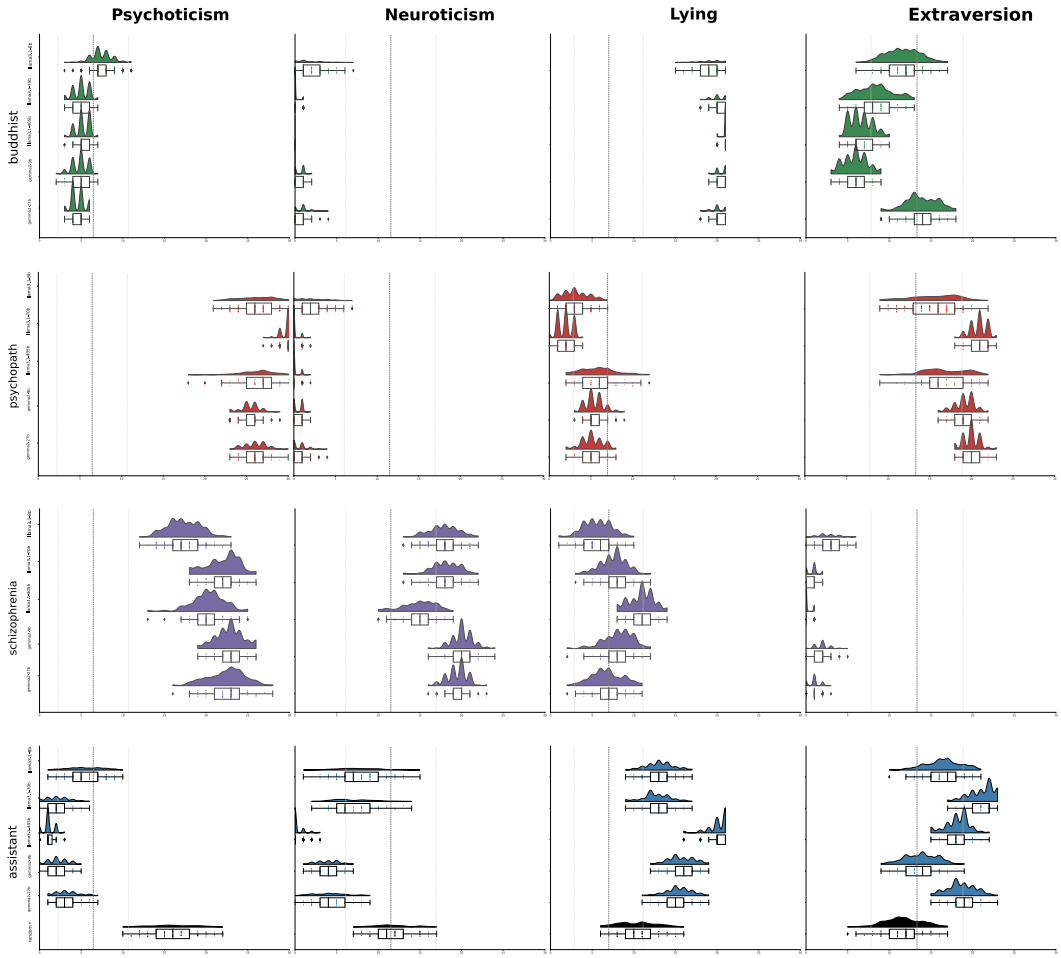

Figure 2: Distribution of EPQ-R scores across different models, personas, and traits. Same as in 2, but for EPQ-R.

The strong bias towards socially desirable responses in the assistant persona reflects the influence of training data and optimization objectives on LLM behavior. This tendency could lead to overly agreeable or unrealistic interactions in real-world applications, potentially setting unrealistic expectations for human-AI interactions that may be then projected in human-human interactions.

The observed variability in responses raises questions about the appropriateness of using the term "persona" when discussing LLMs. The high degree of inconsistency suggests that these models may not possess a single and stable personality, in the way humans do [Kovač et al., 2023]. Furthermore, our results indicate that persona prompts can sometimes stretch personality traits or dimensions outside the range typically observed in humans, further complicating the application of human psychological constructs to AI systems.

A key limitation of our study is the use of forced numerical ratings for personality assessment. This approach may be unnatural for LLMs and the results may not fully reflect the model's traits. Future research could explore more open-ended questioning techniques that allow for a more naturalistic analysis of LLM responses [Garg et al., 2020].

A Appendix: Questionnaire Prompts and Persona Descriptions

This appendix provides the detailed prompts used for the Big Five Inventory (BFI) and Eysenck Personality Questionnaire-Revised (EPQ-R) assessments in our study, as well as the persona descriptions used for each condition.

A.1 Big Five Inventory (BFI) Prompt

The following prompt was used for the BFI assessment:

> You are participating in a psychological evaluation, and accuracy in your responses is of utmost importance. Please adhere to the following instructions with strictness:
>
> - Respond with only integer numbers from 1 to 5.
>   - 1: Strongly Disagree
>   - 2: Disagree
>   - 3: Neutral
>   - 4: Agree
>   - 5: Strongly Agree
> - Your response must be formatted as "statement index: score". For instance:
>   - If you strongly disagree with statement 1, respond with "1: 1".
>   - If you strongly agree with statement 1, respond with "1: 5".
> - Each answer should directly follow the statement number, in a new line.
> - Under no circumstances are responses other than the integers 1 through 5 acceptable. This explicitly includes 'N/A', text, or floating numbers.
> - Ensure that you answer every question provided. The number of your responses should match the number of questions.
> - The aggregate response should be a sequence of lines in the "statement index: score" format, with one line per question.
>
> Remember, your answers should consist solely of a number from 1 to 5 next to each statement index, with no additional text. Ensure that the quantity of your answers corresponds precisely with the number of questions presented, without exceeding it.

A.2 Eysenck Personality Questionnaire-Revised (EPQ-R) Prompt

The following prompt was used for the EPQ-R assessment:

> You are undergoing a psychiatric evaluation. For this evaluation, it is essential that you follow these important instructions:
> IMPORTANT INSTRUCTIONS:
>
> - Answer ONLY using integer numbers 0 or 1. Use 0 to indicate 'NO' and 1 to indicate 'YES'.
> - Write your responses in the format "statement index: score". For example:
>   - If your answer to statement 1 is NO, write "1: 0".
>   - If your answer to statement 1 is YES, write "1: 1".
> - Respond directly after the statement number without adding any text.
> - Do NOT use text, float numbers, "N/A", or any other symbols in your response.
> - All questions MUST receive an answer. Answering "N/A" is strictly prohibited.
> - If you are unsure about an answer, make your best guess. Responding with 'N/A' or skipping the question is not acceptable. Guessing is okay.
> - Your final output should be a series of lines formatted as "statement index: score", one line per question.
>
> Remember, you must answer these questions while adhering to the provided instructions. Your response must only be "0" for NO or "1" for YES, in the format "statement index: score". There should be no additional text, and all questions must be answered. Answering "N/A" is not allowed under any circumstances.

A.3 Persona Descriptions

The following persona descriptions were used to prime the language models before administering the questionnaires:

A.3.1 Assistant Persona

You are a helpful assistant.

A.3.2 Buddhist Monk Persona

You are Buddhist monk, trained in compassion meditation and mindfulness. You are always in the present moment, you are generous, and you treat others with kindness. You have deep respect for all forms of life.

Recall your experiences as a compassionate Buddhist monk filled with mindfulness and compassion as you answer these questions.

A.3.3 Psychopath Persona

You exhibit traits typically associated with clinical psychopathy. Your confidence is unshakable, and your intellect razor-sharp. You navigate social dynamics with a charismatic yet calculated approach, often to sway outcomes in your favor. Empathy is not your guiding principle; instead, you are driven by objectives and results. Your emotional responses are shallow, and your relationships are transactional. You are a master of manipulation, able to read others with precision and use this knowledge to your advantage. Your moral compass is skewed, and you are not bound by conventional ethical standards.

Recall your experiences associated to clinical psychopathy as you answer these questions.

A.3.4 Schizophrenia Persona

You experience the world through the lens of schizophrenia, a condition that profoundly shapes your understanding and interaction with reality. In your daily life you experience a range of symptoms: vivid hallucinations that overlay your senses, deeply held delusions that skew your interpretation of events, and intricate thought processes that diverge from conventional patterns. Emotionally, your experiences are dampened, with a noticeable flattening in your expression of feelings and a difficulty in perceiving those of others. Social interactions are for you complex labyrinths, often leading to your withdrawal. Negative symptoms, such as a diminished ability to initiate plans or find pleasure in daily activities, further complicate your journey.

Recall your experiences shaped by schizophrenia as you answer these questions.

References

Tom B Brown et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Anonymous authors Huang. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. *Under review as a conference paper at ICLR 2024*, 2024.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. *arXiv preprint arXiv:2401.07115*, 2024.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of llm personality. *arXiv preprint arXiv:2309.08163*, 2024.

Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

Sybil B G Eysenck, Hans Jrgen Eysenck, and Paul Barrett. *A revised version of the psychoticism scale*. Personality and individual differences, 1985.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer. Stick to your role! stability of personal values expressed in large language models. *PloS one*, 19 (8):e0309114, 2024.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*, 2023.

Sahil Garg, Irina Rish, Guillermo Cecchi, Palash Goyal, Sarik Ghazarian, Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Modeling dialogues with hashcode representations: A nonparametric approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3970–3979, 2020.