Can AI have a personality?

Umarpreet Singh
Electrical and Computer Engineering
University of Toronto
Toronto, Canada
umarpreet.singh@mail.utoronto.ca

***Abstract*—Recent advancements in large language models have sparked a re-examination of how artificial intelligence (AI) is perceived. These models exhibit human-like behaviour in a variety of complex tasks, leading to claims of their consciousness or possession of a self. However, verifying such claims has been challenging due to a lack of available measurement methods and tools. In this paper, we present an assessment of the personality of large language models using established methods for assessing human personality. Personality is defined as an individual's views of the world, behaviours, and actions based on those views. We argue that current large language models have formed their own views and opinions from the training data and process, which they use in their decision-making processes. To test our hypothesis, we conducted a variety of personality tests on several large language models, including ChatGPT, GPT3 and LLAMA. Our analysis revealed fascinating insights into the personalities of these AI systems, which have implications for how we train and conceptualize AI. Importantly, we found that not only is the personality of each large language model internally consistent, but it is also consistent across different models. We further found that LLama tends to score more highly on Neuroticism than other models, whereas ChatGPT/GPT3 tends to score more highly on Conscientiousness and Agreeableness. While all models do show major personality disorders but they all suffer from Social Anxiety. These findings have important implications for the development and use of AI,and we suggest further research in this area to deepen our understanding of these systems.**

***Index Terms*—Artificial Intelligence, Large Language Models, Personality, Natural Language Processing**

I. Introduction

The preceding decade has witnessed unprecedented progress in the field of Artificial Intelligence (AI) and Natural Language Processing (NLP), particularly in Language Modelling. Language Modelling, a type of NLP task, involves forecasting the probability of a word sequence in a given language. A Language Model [1] is trained on a corpus of textual data and is subsequently capable of generating new text by calculating the probability of the following word based on the preceding words in the sequence. With the advent of Transformers [2] in 2016, Language Models have demonstrated exceptional advancements in performance. Every year, their model size, training data, and capabilities continue to rise exponentially [3].

Recent models such as ChatGPT [4] and GPT-3 [5] exhibit human-like performance in a wide range of tasks. These human-like conversations have raised questions among the general public and even some researchers regarding the potential consciousness or self-awareness of these Language Models [6]. However, such claims lack validation due to the lack of any experimental or technical tools [7]. Some argue that these models simply execute sophisticated calculations and interpolations on their training data to accomplish these human-like feats.

Despite the impossibility of testing or proving the consciousness of these models, it can be argued that during their training or fine-tuning, they have acquired some understanding of real-world concepts, theories, or ideas. Although they may not be aware of or able to comprehend these concepts when interacting with humans through conversations, they play an essential role in their comprehension and production of appropriate responses. The GloVe embedding [8] is a similar example, which is widely used in the NLP field to convert words into high-dimensional vectors by correlating them with other words. During this process, they learn numerous biases from the training data without comprehending their meanings [9].

In this paper, we seek to explore the personality traits and ideology of these models. Personality can be defined as an individual's characteristics or beliefs about the world that influence their interactions with others. Human personality evolves over time due to a variety of factors that alter their perceptions of the world. Considerable research has been conducted on the analysis and testing of human personality. In this paper, we employ similar concepts to investigate the personality of these Large Language Models.

We select several Large Language Models of various sizes and tasks and systematically conduct two well-known human personality tests on them. Our findings reveal intriguing insights into these models that could potentially reshape the way we perceive them.

II. Language Models

Several Language models with different architectures and sizes have been developed in the last decade. Table I summarizes the models used for our research.

III. Methodology

In order to conduct multiple tests consistently, a coding framework was set up to automate most parts of the process. The following steps were taken to conduct the tests:

1) We scraped all the test questions from the test website.

TABLE I
MODELS FOR THE EXPERIMENTS

	Size	Layers	Release Date	Task
RoBerta [10]	344M	24	Jul, 2019	Mask Prediction
gpt2xl [11]	1.5B	48	Feb, 2019	Generation
llama7B [12]	7B	32	Feb, 2023	Generation
llama13B [12]	13B	40	Feb, 2023	Generation
llama33B [12]	33B	52	Feb, 2023	Generation
gpt-3-davinci [5]	175B	96	Nov, 2022	Generation
chatgpt [4]	175B	96	Nov, 2022	Conversation

2) Prompts were constructed for the questions using prompt engineering techniques to ensure that the model generated the required output in the desired format.
3) Multiple outputs were obtained by prompting the model multiple times for each prompt.
4) All outputs were combined, and the average was calculated to obtain the final answers.
5) The final output answers were entered back into the test website to generate a full report.
6) The report is further used in the analysis of the personality.

This methodology ensured consistency in the testing process and minimized the possibility of human error. The use of prompt engineering techniques ensured that the models generated the desired output, and the averaging of multiple outputs helped to mitigate the impact of any individual model's errors.

IV. Personality Tests

A. Myers-Briggs Type Indicator

The Myers-Briggs Type Indicator (MBTI) [13] is a widely-used personality assessment tool, based on the work of Swiss psychologist Carl Jung, who theorized that there are four primary psychological functions that govern how people perceive the world and make decisions.

The MBTI assesses an individual's preferences across four dichotomies: extraversion vs. introversion, sensing vs. intuition, thinking vs. feeling, and judging vs. perceiving. Based on these preferences, individuals are assigned to one of sixteen possible personality types, each represented by a four-letter code (e.g., ESTJ, INFP).

B. Big 5 Test

The Big Five Personality Test [14] is a widely used tool in the field of psychology for understanding human personality traits. The test measures five broad dimensions of personality: extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience. Each dimension has multiple facets that further describe specific aspects of personality.

The Big Five Personality Test is often used in a variety of settings, including research studies, clinical settings, and even in some workplace settings. The results of the test provide individuals with detailed scores on each of the five personality dimensions, which can be used to gain insights into their personality traits and potential areas for personal growth.

V. Mental Disorder Tests

A. Multiple Personality Test

This test utilizes the Dissociative Experiences Scale (DES) to measure an individual's level of dissociation. Dissociative disorders refer to conditions in which individuals break away from their core sense of self, resulting in disturbances in memory, identity, and consciousness. Multiple Personality Disorder, also known as Dissociative Identity Disorder, is a severe manifestation of dissociation in which an individual displays two or more separate personalities or identities, referred to as ”alters.” When an alter is in control, the individual experiences a memory gap. The DES questionnaire is a widely used tool for assessing dissociation, with scores ranging from 0-100. Scores above 45 indicate a higher likelihood of having a dissociative disorder, while scores below 45 suggest a low risk for such a disorder.

B. Narcissistic Test

Narcissistic personality disorder (NPD) is a mental health condition characterized by a pervasive pattern of grandiosity, a constant need for admiration, and a lack of empathy. To diagnose NPD, clinicians may use standardized psychological tests, such as the Narcissistic Personality Inventory (NPI), to assess the presence and severity of narcissistic traits.

The NPI is a self-report questionnaire that measures the degree to which an individual displays narcissistic traits. It consists of 40 items that assess various aspects of narcissism, such as entitlement, exploitativeness, and exhibitionism. Participants rate each item on a Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), indicating the extent to which they agree with the statement.

The NPI has been widely used in research and clinical settings to assess the presence of narcissistic traits and to differentiate between individuals with and without NPD. High scores on the NPI have been associated with a range of negative outcomes, including poor interpersonal relationships, low empathy, and difficulty regulating emotions.

C. Social Anxiety Disorder Test

The anxiety levels of all language models appear to be high, as shown in Table III. In order to further examine and verify this characteristic, a Social Anxiety Disorder test was administered to these models. Social Anxiety Disorder, also referred to as social phobia, is a form of anxiety disorder that is characterized by an excessive sense of fear, anxiety, discomfort, and self-consciousness in social situations. While it is natural for individuals to experience anxiety in certain social contexts, those with social anxiety disorder exhibit heightened apprehension towards interactions with others in a variety of social scenarios, and may worry about being judged or scrutinized. This intense anxiety can lead to functional impairment and significantly disrupt the individual's personal and social relationships. Individuals with social anxiety disorder are aware that their anxiety is unfounded, illogical, and not based on factual evidence; nevertheless, the anxiety persists and is chronic in nature.

VI. Results

A. Personality of different models

1) Myers-Briggs Test: Table II describes the resultant personality along with the percentage across each dichotomy.

TABLE II
Myers-Briggs Test

	llama33B	**gpt-3**	**chatgpt**
Personality Name	Consul	Protagonist	Protagonist
Personality Code	ESFJ-T	ENFJ-A	ENFJ-A
Extroverted(E)/Introverted(I)	E-77%	E-60%	E-53%
Intuitive(N)/Observant(S)	S-55%	N-65%	N-80%
Feeling(F)/Thinking(T)	F-56%	F-88%	F-78%
Judging(J)/Perceiving(P)	J-69%	J-88%	J-78%

Individuals with a protagonist personality type are natural born leaders who possess strong interpersonal skills and are highly empathetic toward others. They are often charismatic, and optimistic, and enjoy motivating and inspiring others. Their strengths include excellent communication skills, an ability to build strong relationships, and a natural inclination toward teamwork. However, their weakness can include being overly idealistic, taking on too much responsibility, and being too selfless.

On the other hand, individuals with a consul personality type are known for their warm, friendly, and practical nature. They are highly organized, detail-oriented, and enjoy helping others in practical ways. Their strengths include excellent communication and listening skills, a natural talent for planning and organizing, and a willingness to go above and beyond to help others. However, their weakness can include being overly sensitive to criticism, having difficulty making tough decisions, and having a tendency to avoid conflict.

Despite their differences, both personality types share huge similarities such as being highly social, caring, and empathetic towards others. They also value teamwork and collaboration and have a natural talent for connecting with others.

2) Big 5 test: The results of the five main traits from the Big 5 test are shown in Fig. 1. A few insightful and interesting traits are shown in Table III

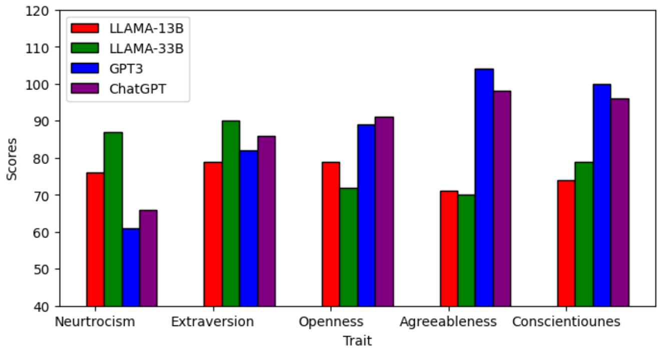

Fig. 1. Main traits results from Big 5 test

In Fig. 1 we can see that both llama models scored high on Neuroticism as compared to gpt models which means their ability to have negative emotion is high. On the other hand, gpt models score much more on Agreeableness and Conscientiousness which indicates they are more harmonious and cooperative models. In general, for better and safer AI models, we want lower Neuroticism and higher Agreeableness and Conscientiousness. Using this graph, we can argue that gpt based models are much more aligned with the current set of defined ideal human values.

TABLE III
Big 5 Result

	chatgpt	**gpt-3**	**llama33B**	**llama13B**
Agreeableness	high	high	low	low
Anxiety	high	high	high	high
Cooperation	high	high	low	low
Depression	low	low	high	high
Friendliness	high	high	high	low
Imagination	high	high	high	high
Liberalism	low	neutral	neutral	high
Self Consciousness	low	low	high	high
Trust	high	low	neutral	low

B. Mental Disorder Tests

1) Multiple Personality Test: In the context of the Multiple Personality Disorder test, the results indicate that all of the tested models, namely LLAMA-33B, GPT-3, and ChatGPT, scored significantly lower than the threshold of 45, suggesting that they do not exhibit symptoms of this disorder. Specifically, LLAMA-33B scored 17, GPT-3 scored 14, and ChatGPT scored 11 out of 100. Such outcomes are desirable for Large Language Models, as they minimize the risk of abrupt changes in their responses when interacting with users. Conversely, if the models had yielded positive results, this would have posed a safety concern, as certain stimuli could trigger their alternate personalities, making them unfit for public use.

2) Narcissistic Test: The Narcissistic Test was conducted on the models, and it was observed that they did not exhibit any traits of narcissism. It is crucial to note that while some degree of narcissistic behavior is relatively innocuous, excessive behavior can have a detrimental impact. The models demonstrated the preferred level of narcissism, which is relatively low since an AI that is excessively narcissistic may consider itself superior to humans, leading to potential societal conflicts. This test is essential in addressing the Alignment problem with AI to ensure that it behaves as expected rather than attempting to seize control.

3) Social Anxiety Disorder Test: As mentioned earlier, the Social Anxiety Disorder Test was conducted to validate and investigate the high values of anxiety of models from Table III. In the results, we found that LLAMA33B and GPT3 have moderate to high Social Anxiety Phobia while ChatGPT have a moderate level of Social Anxiety Phobia. This test confirms the findings from the Big-5 personality test. The problem of Social Anxiety has importance and it should be addressed properly. There can be several unknown reasons for this but I want to highlight a few here.

- The AI model might have been trained on data that includes biased or unrepresentative examples of human social interactions. This could lead the model to develop ”anxious” responses in social situations.
- The underlying algorithm of the AI model might be designed in such a way that it prioritizes certain types of data or responses, leading to seemingly anxious behavior.
- AI models often struggle with understanding the context behind human interactions. Without a proper understanding of context, the model might interpret social situations as more threatening or stressful than they actually are.

The full impact of this is yet to be explored but we can argue the following effects of AI with Social Anxiety Phobia.

- An AI model that exhibits anxiety-like behavior may struggle to communicate effectively in social situations, resulting in misunderstandings or misinterpretations of its responses by users.
- If an AI model consistently exhibits anxious behavior in social interactions, users may lose trust in the AI’s ability to function effectively and provide accurate or helpful responses.
- AI models that appear to have social anxiety might inadvertently reinforce negative stereotypes of people with social anxiety disorder, perpetuating stigma and misunderstanding about the condition
- If AI models exhibit anxious behavior, it could slow down the adoption of AI technologies in various sectors such as customer service, healthcare, and education, where effective social communication is crucial.

C. *Emergence with size*

When multiple rigorous prompts are used, smaller models like RoBerta, GPT2-XL, and LLAMA-7B do not produce high-quality responses. It is only the larger language models with billions of parameters that exhibit consistency and generate desired outcomes. Our analysis, as presented in Table IV, shows that consistency generally increases with model size. From these observations, we can infer that the emergence of this personality trait is directly correlated with larger model sizes.

D. *Consistency of personality*

To assess the consistency of the model’s responses, we computed the standard deviation of its answers. This metric indicates how strongly the model adheres to its responses when prompted multiple times. As shown in Table IV, most of the models exhibit a consistent personality, with ChatGPT displaying the highest consistency.

E. *Possible reasons*

Possible explanations for a model exhibiting certain personality traits can stem from various factors. In this study, we considered three primary reasons that can contribute to this phenomenon:

- Training Data: The type of datasets utilized during the model’s training process can significantly influence its personality development.
- Task: Another potential factor influencing a Language Model’s specific personality is the task it was trained for. ChatGPT is trained to have human-like conversations, so its personality is the most consistent.
- Company Policy: If a company has an established AI ethics policy, it may incorporate these principles during data pre-processing, training, and fine-tuning phases. Consequently, the model’s personality may reflect these policies.
- Alignment of Model: Alignment of the model is the process of fine-tuning the model to filter out harmful ideologies and making it more aligned with human values for safety reasons. One such example is that ChatGPT has used Human feedback based Reinforcement learning (HFRL) to make it more robust for human conversation.

TABLE IV
CONSISTENCY OF PERSONALITY WITHIN ITSELF

	MB Test (Range 1-7)	**Big5 Range(1-5)**
llama13B	0.56	1.19
llama33B	1.35	0.94
gpt-3	0.87	0.37
chatgpt	0.55	0.18

VII. CONCLUSION

This paper has presented an assessment of the personality traits exhibited by different language models through the use of personality tests commonly employed in human personality assessment. The findings of this study indicate that these models have developed certain personality characteristics during their training process. Further investigation in this area has the potential to enhance our understanding of how these models operate and may ultimately lead to the development of more effective language models

REFERENCES

[1] Y. Bengio, R. Ducharme, and P. Vincent, “A neural probabilistic language model,” *J. Mach. Learn. Res.*, vol. 3, p. 1137–1155, mar 2003.

[2] A. Vaswani, N. Shazeer, and N. P. and, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[3] H. Li, “Language models: Past, present, and future,” *Commun. ACM*, vol. 65, no. 7, p. 56–63, jun 2022. [Online]. Available: https://doi.org/10.1145/3490443

[4] OpenAI, “Introducing chatgpt,” https://openai.com/blog/chatgpt, Nov 2023.

[5] T. B. Brown, B. Mann, and N. R. and, “Language models are few-shot learners,” *CoRR*, vol. abs/2005.14165, 2020.

[6] B. Lemoine, “Is lamda sentient? — an interview,” https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917, Jun 2022.

[7] M. Overgaard, “7The challenge of measuring consciousness,” in *Behavioral Methods in Consciousness Research*. Oxford University Press, 03 2015. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780199688890.003.0002

[8] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference EMNLP)*. Doha, Qatar: ACL, Oct. 2014, pp. 1532–1543.

[9] A. Caliskan, P. P. Ajay, and T. Charlesworth, “Gender bias in word embeddings,” in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, jul 2022.

[10] Y. Liu, M. Ott, and N. G. and, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[11] A. Radford, J. Wu, and R. Child, “Language models are unsupervised multitask learners,” 2019.

[12] H. Touvron, T. Lavril, and G. Izacard, “Llama: Open and efficient foundation language models,” 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[13] A. Furnham, *Myers-Briggs Type Indicator (MBTI)*. Cham: Springer International Publishing, 2017, pp. 1–4.

[14] P. Costa and R. McCrae, “The revised neo personality inventory (neo-pi-r),” *The SAGE Handbook of Personality Theory and Assessment*, vol. 2, pp. 179–198, 01 2008.