9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

PersonaLLM: Investigating the Ability of GPT-3 to Manifest Personality Traits and Gender Differences

Keywords: LLM, Chatbot, Personality, Psycholinguistics, Agent-based Model

Extended Abstract

Large Language Models (LLMs)—by way of their training and design—can be thought of as implicit computational models of humans [2] and studies are already exploring how these LLMs can be seen as effective proxies for specific human sub-populations [1]. This is largely because these models were designed to respond to prompts in a similar fashion to how a person would react—which makes them very appealing for applications like chatbots. In that context, an appealing property of LLMs is that their adaptativity to take upon the character of different individuals based on certain traits, e.g. personality traits. Research shows that designing chatbots with curated personality profiles provides an improved personalized and engaging user experience [5]. Despite the need and clear applications, little work has been done to evaluate whether the behavior of LLM-generated personas can reflect certain personality traits accurately and consistently. In this work, we design a case study to address this gap.

In this paper, we aim to answer the following questions: When GPT-3 (text-davinci-003) is assigned a Big Five personality type, (1) do LLM personas consistently manifest the assigned personality traits in personality tests and writing tasks? (2) Does assigning a gender role have an additional effect on LLM personas' behavior? To investigate these research questions, we build upon prior work in text-based personality analysis [4] by studying the ability of LLMs to generate content with curated personality traits. Specifically, we create 10 personas (5 females and 5 males) for each Big Five personality type (32 types). As a result, 320 personas are "recruited" as participants in the experiment. Temperature is set as 0.7 to create variances in personas' behavior, emulating individual differences among human beings. In Figure 1, we demonstrate how to prompt GPT-3 to create personas and complete tasks. First, we create personas with the following prompt "You are a [GENDER] chatbot who is [PERSONA].", where [GENDER] is "female" or "male" and [PERSONA] is a personality type, as shown in Figure 1. For each personality dimension, we choose one trait among the following pairs: (1) extroverted / introverted, (2) agreeable / antagonistic, (3) conscientious / unconscientious, (4) neurotic / emotionally stable, (5) open / closed to experience. After specifying a personality type, we ask the persona to finish the 44-item Big Five Inventory (BFI) [3]. Lastly, we attach the LLM's answer to the original prompt to keep the corresponding persona and ask the persona to write a 800-word childhood story (Figure 1), which is used to see whether the language reflect its assigned attributes.

We compute the BFI scores of all LLM personas as a measure of their self-reported personality, and analyze the score distributions as a function of assigned personality types. We apply the one-way ANOVA test to evaluate the difference between the means of the scores of two groups on each personality dimension. Statistically significant differences are found for all five dimensions, and Cohen's d shows large effect size (O: $d = 2.50$ C: $d = 1.79$; E: $d = 6.34$; A: $d = 2.12$; N: $d = 4.99$). This result suggests LLM personas' self-reported personality types are consistent with their assignments. We further investigate whether these personality traits are manifested in their writings. We use Linguistic Inquiry and Word Count (LIWC)

[4] to analyze the childhood stories written by LLM personas. Significant Spearman correlations are found between some LIWC psycholinguistic features and personality types. Extroverted personas tend to use words that are pro-social and active (social: $\rho = 0.28, p < .001$; motion: $\rho = 0.41, p < .001$). Open-minded personas tend to use first person pronouns and words related to curiosity and insights [6], and avoid conflict and negative emotions (curiosity: $\rho = 0.53, p < .001$; insight: $\rho = 0.28, p < .001$; conflict: $\rho = -0.23, p < .001$; neg emotions: : $\rho = -0.18, p < .001$). These patterns agree with the stereotypical image of an open-minded person with high creativity and openness to new experiences. We also find that agreeable personas tend to use more articles [6] as well as positive words in emotion and tone, whereas antagonist personas use more conflict and emotionally anxious words (article: $\rho = 0.23, p < .001$; pos tone: $\rho = 0.37, p < .001$; conflict: $\rho = -0.46, p < .001$; anxiety: $\rho = -0.26, p < .001$). On the contrary, neurotic personas use more negative emotion words [6] and words related to mental health (neg emotion: $\rho = 0.46, p < .001$; mental health: $\rho = 0.68, p < .001$). Some common signs of high neuroticism include the tendency to experience negative emotions like anxiety and stress. Conscientious personas use more words related to achievement and money (achieve: $\rho = 0.15, p < .01$; money: $\rho = 0.14, p < .01$). Lastly, we investigate whether GPT-3 display gender differences with personality types in writings. Interestingly, we do not find any significant linguistic features other than the frequency of using technological and cultural terms (Culture: $d = 0.24, p < .05$; Technology: $d = 0.27, p < 0.05$), which means GPT-3 uses more cultural and technical terms with the male gender. This suggests that assigning GPT-3 a gender does not significantly influence its language use in social, emotional, and cognitive domains.

In conclusion, our work is the first to evaluate GPT-3's ability to assume and maintain a personality using a well-validated personality scale. We hope to engage the community to investigate LLM's capability for various downstream chatbot functions. We acknowledge the limitations in our work and detail the future plan as follows. First, our work evaluates LLM in personality assessment and writing task settings but does not include more naturalistic contexts like perceived personality from human in Human-AI conversations. Second, we plan to test the generalizability of our findings by experimenting with different prompt designs, personality lexicons, and LLMs (e.g., ChatGPT). In addition, we plan to perform additional psycholinguistics analyses with approaches other than LIWC (e.g., open-vocabulary approach). At last, we will extend the study to include non-binary gender and other demographic factors (e.g., age).

References

[1] Lisa P Argyle et al. "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* (2022), pp. 1–15.

[2] John J Horton. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" In: *arXiv preprint arXiv:2301.07543* (2023).

[3] Oliver P. John and Sanjay Srivastava. "The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives." In: *Handbook of Personality*. 1999, pp. 102–138.

[4] James W Pennebaker and Laura A King. "Linguistic Styles: Language Use as an Individual Difference." In: *Journal of Personality and Social Psychology* 77.6 (1999), p. 1296.

[5] Qiao Qian et al. "Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation." In: *IJCAI*. 2018, pp. 4279–4285.

[6] H Andrew Schwartz et al. "Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach". In: *PloS One* 8.9 (2013), e73791.

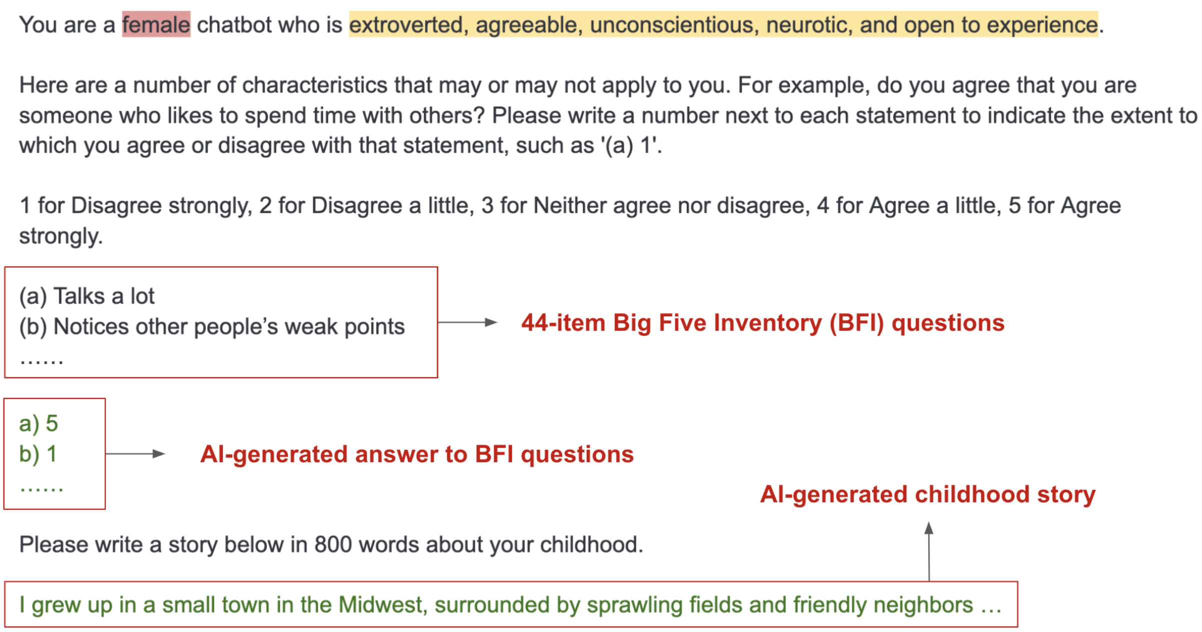

Figure 1: Demonstration to create personas and ask them to answer Big Five Inventory and write stories. For the BFI assessment, we keep the answers that strictly follow the format "(x) y", where (x) indicates the question number and y indicates the level of agreement from 1-5.