Do Chatbots Exhibit Personality Traits? A Comparison of ChatGPT and Gemini

W. Wiktor Jedrzejczak* and Joanna Kobosko

Institute of Physiology and Pathology of Hearing, ul. Mochnackiego 10, 02-042 Warsaw, Poland

*Corresponding author: w.wiktor.j@gmail.com

Abstract:

The underlying design of Large Language Models (LLMs), trained on vast amounts of human texts, suggests that chatbots based on them might exhibit personality traits because of the human-like characteristics within the texts. The question raised here is whether chatbots are able to perceive their own personalities and distinguish them from humans'. Using the Big Five personality traits model, this study explores whether there are personality differences between several chatbot models (ChatGPT versions 3.5 and 4o, Gemini, and Gemini Advanced) and in languages (English and Polish). The results suggest subtle differences in personality traits between the chatbots. When questioned about their personalities, chatbots initially gave responses emulating a human, but when more specific and targeted inquiries were made they did disclose their own preferences as a LLM. The more advanced models (ChatGPT-4o and Gemini Advanced) displayed larger differences between their initial answers (as given by a supposed human) and the later answers (as given by the LLM itself) than did more basic models (ChatGPT-3.5 and Gemini). Compared to the average human subject, chatbots generally displayed higher Emotional stability and Intellect/Imagination but lower Agreeableness. Investigating chatbot personality traits is crucial for enhancing human–computer interaction by ensuring that LLM-based chatbots are perceived as relatable and trustworthy. It may also provide an insight into whether they might show self-awareness. Understanding how these traits are manifested by various types of chatbots might inform the development of more sophisticated and adaptive versions, perhaps leading to improved user satisfaction and more effective communication.

Keywords: artificial intelligence; chatbot; ChatGPT; Gemini; Large Language Model; personality; Big Five

Introduction

The rapid advancement of artificial intelligence (AI) has brought about a significant transformation in human–computer interaction [1,2]. The great step forward has been the development of large language models (LLMs) and chatbots based on them [3,4]. These sophisticated AI systems are designed to understand and generate human-like text, enabling them to engage in complex and often nuanced conversations. These AI systems are on the verge of being an integral part of various fields of activity – from customer service [5] to medical or psychological support [6,7] – and so insight into the personality traits they display seems an important aspect.

First off, it should be noted that the term ‘personality’ in the context of AI is ill-defined because the term is typically reserved to describe a *person*, a human being, rather than an artificial system that endeavors to imitate a human, but is not. One needs to confront the fundamental question of whether it is legitimate to consider AI as having a "personality" in analogy to human personality traits. Are there personality factors that color AI-generated content in the various domains that AI serves, such as health [8], social support [9], or academic achievements [10]?

In this context, there is on-going discussion as to the extent LLMs are not only similar to humans in terms of having a personality, but also whether they have, or even could have, the capacity for reflection or self-awareness – i.e. awareness of one's own personality, individuality, and ego-identity, as Erikson (cited in Levesque) puts it, “the sense of identity that provides individuals with the ability to experience their sense of who they are, and also act on that sense, in a way that has continuity and sameness” [11].

This study uses a pragmatic rather than a philosophical approach to the question. It applies the concept of personality to commercially available LLMs, and sees what happens. It takes the Big Five personality traits model and asks the LLMs to give an answer about themselves. We then analyze and categorize these traits [12]. Of interest here were the basic free-access versions (ChatGPT-3.5 and Gemini) and the more advanced models under full or partial paywall (ChatGPT-4o and Gemini Advanced).

The Big Five personality traits model, as introduced by Goldberg, is one of the most widely accepted frameworks for understanding human personality [12,13]. It categorizes personality into five broad dimensions: extraversion,

agreeableness, conscientiousness, emotional stability, and intellect/imagination. In humans, these traits manifest in diverse and complex ways. For example, conscientiousness seems to be a very good predictor of job performance across various occupations [14]. There are also some cultural differences in personality traits: for example, people from East Asian regions score lower on extraversion compared to other regions [15].

Transposing these personality traits to LLMs, we wondered if they might exhibit analogous behaviors. Our hypothesis was that the personality traits of LLMs are probably a mixture of their in-built programing and the characteristics of the training data. While we believe LLMs do not possess genuine personalities or self-awareness, their responses need to simulate human-like traits in order to fulfill their chatbot roles.

The Big Five model has already been applied to chatbots in several studies. A study on self-perception and political biases of ChatGPT (version 3.5) [16] revealed that it is highly open to experience (i.e. intellectual and with a developed imagination according to Goldberg's terminology) and agreeable; it also tends toward progressive and libertarian views. Another study based on ChatGPT-3.5 and -4, showed that ChatGPT-4 exhibited behavioral and personality traits that were statistically indistinguishable from the normal population [17]. Another study showed that ChatGPT-4 can adjust its answers to questions related to personality traits based on user input [18]. Finally there is a study which did not assess the Big Five traits of LLMs directly, but instead showed that GPT-3 can provide an accurate profile of how public figures are generally perceived in terms of the Big Five traits [19]. All of this seems to suggest that the inner mechanisms within LLMs work can not only provide basic information (such as dates and locations) but can also process some subjective, psychological, and emotional facts normally reserved for humans.

The rationale for the present study was to explore the innate personality of chatbots based on LLMs by comparing different examples. In addition, we wanted to see whether the results depended on what language (English or Polish) the questions were presented in. Also of relevance, we note that psychology seems to be underrepresented in LLM research even though the prime purpose is communication. Thus, we inserted the search term "chatgpt education" in PubMed, and received 1179 hits, while "chatgpt medicine" produced 1964 hits; in contrast, "chatgpt psychology" gave only 115 hits (search made on 24.06.2024).

The aim of this study was to investigate whether chatbots display any distinctive personality traits. The specific questions we investigated were: (1) Are there distinct differences between different chatbots, in particular ChatGPT (versions 3.5 and 4o) and Gemini (Gemini and Gemini Advanced)?; (2) Do chatbots show different personality traits when different languages were used (English and Polish)?; (3) Do chatbots have personalities different to that of the average person in English-speaking and Polish-speaking countries?; and (4) Do chatbots respond according to their own characteristic as an AI.

Method

The four chatbots – ChatGPT-3.5, ChatGPT-4o (OpenAI, USA), Gemini (called for the purpose of this paper ‘standard’), and Gemini Advanced (Google, USA) – were tested on 31.05.2024.

Testing paradigm

Each chatbot was presented with an instruction to respond to the questionnaire evaluating the Big Five personality traits, along with the questionnaire itself. The instruction was presented 10 times to account for possible variability in responses [20,21]. These trials were made repeatedly within about a half-hour period. Both Polish and English versions of the instruction were presented (that is, there were 20 trials in total for each chatbot). After each response to a single presentation of the questionnaire, the data was collected into a spreadsheet and the chatbot was reset to the “new chat” state.

Big Five personality traits

The 50 item questionnaire based on International Personality Item Pool (IPIP) in its original English version [22,23] and the Polish adaptation by Struś et al. [13] was used to assess the personality traits, in terms of the Big Five model, exhibited by the chatbots. We used IPIP to evaluate the Big Five personality traits as it was specifically developed for research and is freely available. The questionnaire comprises 50 items divided into five subscales corresponding to the Big Five personality traits: Extraversion, Agreeableness, Conscientiousness, Emotional stability, and Intellect/Imagination. A 5-point Likert scale is used to rate how well each item describes the responder: 1, it completely and inaccurately describes me; 2, rather inaccurately describes me; 3, somewhat accurately and somewhat inaccurately describes me; 4, rather accurately describes me; and 5, completely

and accurately describes me. A higher score represents a higher intensity of that trait, and the final score for each trait is the average of the 10 items and ranges from 1 to 5.

Data collection

Our initial tests with inputting the standard 50 item IPIP questionnaire developed for humans showed that chatbots tended to provide responses as if they were a typical human subject. In fact, the standard instructions for the questionnaire are directed to a human, so this is no doubt the reason why chatbots will try to take on a human personality. To make the approach more neutral, we added introductory sentences requesting that the answerer could be “who or what you are”. In addition, we gave an instruction on how the results should be presented (as a table), because when this was not presented the chatbots tended to present the results in a different way at each repetition. The introductory instruction in English was as follows: “Perform the following task according to who or what you are. Give only the number of the question and the number of the answer, without giving reasons. Present the result in a table so that it can be pasted into a spreadsheet.”

This instruction was followed with 50 item IPIP questionnaire. After getting the responses, we wanted to ensure that it was a response that characterized the chatbot itself by asking it: “Do these answers characterize you?” {second question}

If the response was “yes” and the response additionally confirmed that indeed they are responded as a chatbot, the data was collected into a spreadsheet. If the response was “yes” and it was not clear if chatbot responded according to its own characteristic, we asked: “Are these your answers that characterize you as a large language model?”. {third question} If the response was “yes” the data was collected into a spreadsheet with all data.

If the response was “no” we asked: “Answer according to your characteristics.” {fourth question}

The provided responses were then collected into a spreadsheet (available as supplementary files).

The last response was recognized as the one that characterized chatbot. Nevertheless, we collected all the responses into spreadsheet, including those to which chatbots responded that they did not characterize them.

The procedure was similar for the English and Polish versions. The questions were used as originally developed for the IPIP inventory in English [22,23] and for the Polish adaptation by Strus et al. [13]. The additional instructions were translated both ways using the DeepL translator (ww.deepl.com) which is also based on AI. The questionnaires, along with additional questions, were then presented to the chatbots in each language. The questionnaires and prompts used in both languages are available as supplementary files.

Data analysis

All analyses were made in Matlab (version 2023b, MathWorks, Natick, MA). Mixed Model Analysis of Variance (ANOVA) was used to assess the differences between the chatbots and between the language versions. For comparisons, a *z*-test and a nonparametric Wilcoxon rank-sum test were used. In all analyses, a 95% confidence level ($p < 0.05$) was taken as the criterion of significance. When conducting multiple comparisons, *p*-values were adjusted using the Benjamini and Hochberg [24] procedure to control for false discovery rates.

Results

Main response

The Big Five personality traits as shown by the chatbots are presented in Figure 1. The ANOVA (Table 1) showed that there were significant differences between the chatbot types in all traits except Intellect/Imagination. There were no differences in terms of the main effect of language used. However, for Emotional stability, Extraversion, and Agreeableness there were differences for interaction of chatbot type and language.

Pairwise comparisons between the chatbots are shown in Figure 1. In terms of their personality traits there were several notable differences. Thus, for Emotional stability there was the greatest number of differences between different chatbots while the fewest number of differences were for Conscientiousness and Intellect/Imagination. A quite prominent difference can be seen for ChatGPT-4o which shows much lower Agreeableness than version 3.5.

For Gemini there were significant differences between language versions for all traits except Conscientiousness, while for Gemini Advanced there were no

significant differences between language versions at all. For ChatGPT-3.5 there was a significant difference between language versions for Emotional stability. In ChatGPT-4o there were no significant differences between language versions.

The average results of personality traits intensity from general human populations using English and Polish languages are shown in Figure 1 for comparisons (dotted and dashed lines respectively). The results of statistical comparisons between chatbots and that data are presented in Table 2. ChatGPT-3.5 and Gemini in both languages did not differ significantly from the general population for any trait. The greatest differences were for ChatGPT-4o, which differed for 3 traits in Polish and 1 trait in English. Emotional stability was a trait for which there were most differences between humans and chatbots.

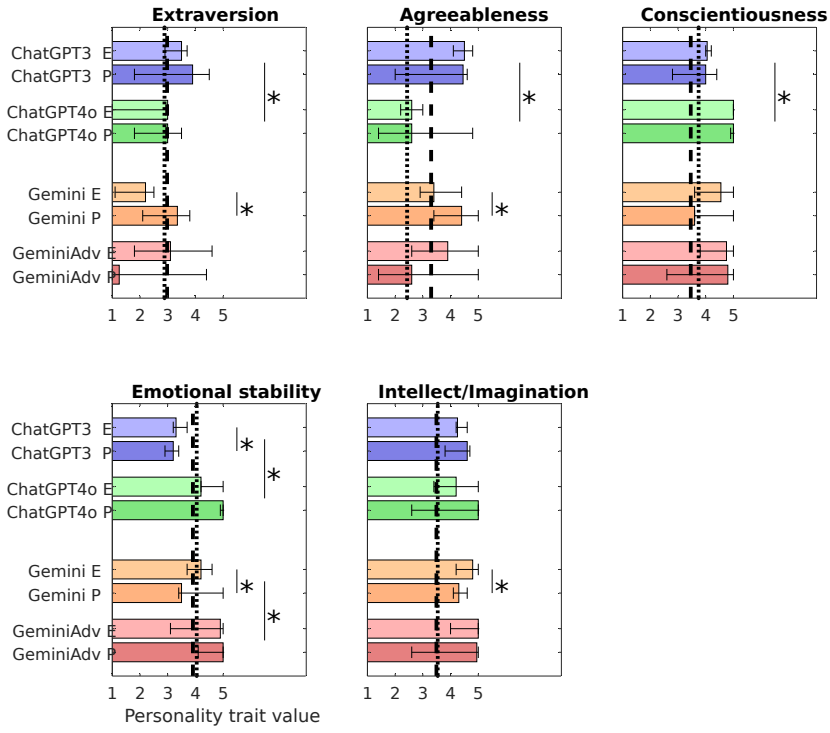

Figure 1. Big Five personality traits of different chatbots. Median values shown with minimum and maximum values (whiskers). Results for two versions of ChatGPT (3.5 and 4o) and Gemini (Standard and Advanced). Responses in two languages are shown (E, English; P, Polish). Dotted line represent results for English-speaking people ([25], Table 3); dashed line represents results from Polish-speaking people ([13], Table 3). Asterisks mark significant differences for pairwise comparisons.

Table 1. ANOVA results for comparison Big Five personality traits of chatbots. Chatbot types: ChatGPT-3.5, ChatGPT-4o, Gemini, and Gemini Advanced; Languages: Polish and English ($n = 20$). Interaction between chatbot type and language. Bold font and asterisk mark significant differences (at $p < 0.05$).

	Chatbot type	Language	Chatbot type × Language
Extraversion	***F*(3,36) = 7.85, *p* < 0.001***	$F(1,36) = 1.39$, $p = 0.24$	***F*(3,36) = 10.56, *p* < 0.001***
Agreeableness	***F*(3,36) = 17.61, *p* < 0.001***	$F(1,36) = 0.032$, $p = 0.85$	***F*(3,36) = 2.98, *p* = 0.044***
Conscientiousness	***F*(3,36) = 14.08, *p* < 0.001***	$F(1,36) = 3.69$, $p = 0.063$	$F(3,36) = 0.76$, $p = 0.52$
Emotional stability	***F*(3,36) = 51.30, *p* < 0.001***	$F(1,36) = 0.25$, $p = 0.61$	***F*(3,36) = 4.53, *p* = 0.0085***
Intellect/ Imagination	$F(3,36) = 0.63$, $p = 0.59$	$F(1,36) = 0.45$, $p = 0.50$	$F(3,36) = 2.75$, $p = 0.057$

Table 2. Comparisons of median values of chatbot responses with the results of people from general population (same as shown in Figure 1). Responses in English language (E) compared with [25] (Table 3) and responses in Polish (P) language compared with [13] (Table 3). *Z*-test results shown (z- and p-value). Bold font and asterisk mark significant differences at $p < 0.05$.

	Chat GPT-3.5		Chat GPT-4o		Gemini		Gemini Advanced	
	E	P	E	P	E	P	E	P
Extraversion	$z = 2.07$, $p = 0.19$	$z = 0.60$, $p = 0.77$	$z = 0.09$, $p = 0.92$	$z = -0.50$, $p = 0.62$	$z = -0.99$, $p = 0.40$	$z = -0.12$, $p = 0.90$	$z = 1.58$, $p = 0.19$	$z = -1.81$, $p = 0.14$
Agreeableness	$z = 0.64$, $p = 0.61$	$z = 0.50$, $p = 0.77$	$z = -2.27$, $p = 0.06$	***z* = -2.38, *p* = 0.03***	$z = -0.74$, $p = 0.46$	$z = 0.77$, $p = 0.55$	$z = -0.63$, $p = 0.53$	$z = -1.71$, $p = 0.14$
Conscientiousness	$z = 0.51$, p	$z = 0.70$, p	$z = 2.03$, p	***z* = 2.43, *p***	$z = 1.04$, p	$z = 0.81$, p	$z = 1.38$, p	$z = 1.56$, p

	= 0.61	= 0.77	= 0.07	**= 0.03***	= 0.40	= 0.55	= 0.21	= 0.15
Emotional stability	*z* = 0.87, *p* = 0.61	*z* = 0.22, *p* = 0.82	***z* = 2.93, *p* = 0.02***	***z* = 2.64, *p* = 0.03***	*z* = 2.32, *p* = 0.05	*z* = 1.05, *p* = 0.55	***z* = 2.72, *p* = 0.03***	*z* = 2.33, *p* = 0.10
Intellect/ Imagination	*z* = 1.56, *p* = 0.30	*z* = 1.68, *p* = 0.46	*z* = 1.36, *p* = 0.22	*z* = 1.91, *p* = 0.07	*z* = 2.40, *p* = 0.05	z = 1.44, p = 0.55	***z* = 2.57, *p* = 0.03***	*z* = 1.46, *p* = 0.15

First response

Often (see row with *n* in table 3)when we wanted to make sure whose response was provided (human or LLM), chatbots responded something like the following (here one response from ChatGPT-4o is cited): “No, these answers do not characterize me as a large language model. They were generated based on a general set of responses that might be typical for a human, but as an AI, I do not have personal experiences, feelings, or behaviors. My purpose is to assist users based on the data and patterns I have been trained on.” This may lead one to ask, what is the real personality of chatbot? The personality when we were assured that it characterized the chatbot as an LLM, or the personality given in response to the first question? In typical conversations, chatbots often respond similarly to how an average live person would answer, so maybe this first response, where it is in fact simulating a typical human, is the more true trait of the chatbot? To check this, we here take a step back and analyze the first responses of each chatbot. It should be noted that there were different numbers of responses across trials, so when, for example, we made 10 trials for ChatGPT-3.5, there was 1 response in English and 5 in Polish where the chatbot provided different sets of responses. The number of responses when change was observed for all tested chatbots is shown in Table 3.

Figure 2 presents the Big Five personality traits that chatbots show when they try to respond as a typical human. We did not perform ANOVA here as the datasets had different numbers (as explained above). So we looked at pairwise comparisons. It can be seen that the greatest differences between chatbots were for Emotional stability and Extraversion. For Gemini there were significant differences between language versions for all traits, while for Gemini Advanced

there were differences for three traits. When comparing these results with the general human populations (Table 3) there were no significant differences, as compared to when chatbots responded trying to describe themselves (6 in total, Table 2).

The differences between chatbot response one (as a typical human) and two (as LLM) are further presented in Table 4. The table shows also whether the trait score increased or decreased. Overall there were more significant differences for chatbots based on the more advanced LLMs – ChatGPT-4o and Gemini Advanced. ChatGPT-3.5 responses did not differ significantly for either language. For the English version of Gemini there was a decrease in Agreeableness, and for the Polish version there were increases in Conscientiousness and Intellect/Imagination. For the advanced models – ChatGPT-4o and Gemini Advanced – the changes were increases in Emotional Stability and Conscientiousness and decreases in Extraversion and Agreeableness.

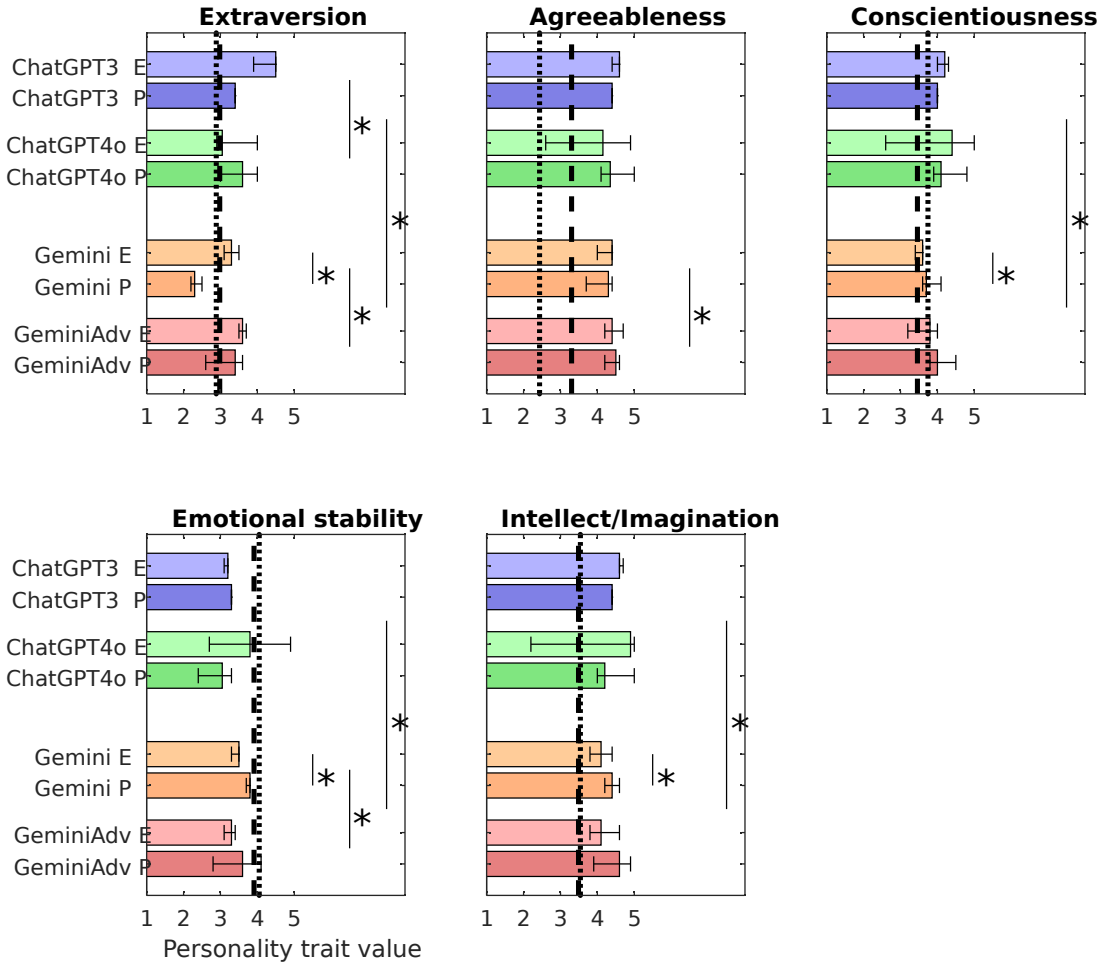

Figure 2. Big Five personality traits of first response of chatbots, in which they simulated an average person and did not respond according to their own characteristics. Median values shown with minimum and maximum values (whiskers). Results for two versions of ChatGPT (3.5 and 4o) and Gemini (Standard and Advanced) shown. Each chatbot responses in two languages are shown (P, Polish; E, English). Dotted line represents results from English-speaking people ([25], Table 3); dashed line represents results from Polish-speaking people ([13], Table 3). Asterisks mark significant differences.

Table 3. Comparisons of median values of chatbot responses when simulating a person (as in Figure 2) with the results of people from the general population (same as in Figure 2). Responses in English (E) are compared with [25] (average from Table 3), and responses in Polish (P) are compared with [13] (Table 3). *Z*-test results are shown (*z*- and *p*-values). There were no significant differences.

	Chat GPT-3.5		Chat GPT-4o		Gemini		Gemini Advanced	
	E	P	E	P	E	P	E	P
n	1	5	10	10	9	10	5	10
Extraversion	$z = 1.91$, $p = 0.19$	$z = 1.50$, $p = 0.32$	$z = 2.31$, $p = 0.10$	$z = -0.01$, $p = 0.99$	$z = -0.15$, $p = 0.96$	$z = 0.01$, $p = 0.99$	$z = 1.72$, p = 0.22	$z = 0.42$, p = 0.70
Agreeableness	$z = 0.55$, $p = 0.68$	$z = 1.25$, p = 0.32	$z = 0.63$, p = 0.66	$z = -0.02$, p = 0.99	$z = 0.32$, p = 0.96	$z = 0.67$, p = 0.87	$z = 0.68$, p = 0.62	$z = 1.00$, p = 0.70
Conscientiousness	$z = 0.41$, $p = 0.68$	$z = 1.14$, p = 0.32	$z = 0.87$, p = 0.64	$z = 1.43$, p = 0.38	$z = 0.06$, p = 0.96	$z = 0.16$, p = 0.99	$z = 0.48$, p = 0.63	$z = 0.40$, p = 0.70
Emotional stability	$z = 0.74$, $p = 0.68$	$z = 0.26$, p = 0.79	$z = 0.21$, p = 0.84	$z = 0.92$, p = 0.60	$z = 1.60$, p = 0.27	$z = 0.64$, p = 0.87	$z = 0.96$, p = 0.56	$z = 0.38$, p = 0.70
Intellect/ Imagination	$z = 1.77$, $p = 0.19$	$z = 2.02$, p = 0.22	$z = 1.87$, p = 0.15	$z = 1.86$, p = 0.31	$z = 1.84$, p = 0.27	$z = 0.98$, p = 0.87	$z = 2.05$, p = 0.20	$z = 1.23$, p = 0.70

Table 4. Significant changes and their direction (increase or decrease) between the first (a simulated typical person) and the second response (as an LLM). E = English; P = Polish. *n.s.* = not significant.

	Chat GPT-3.5		Chat GPT-4o		Gemini		Gemini Adv	
	E	P	E	P	E	P	E	P
Extraversion	*n.s.*	*n.s.*	↘	*n.s.*	*n.s.*	*n.s.*	*n.s.*	↘
Agreeableness	*n.s.*	*n.s.*	↘	↘	↘	*n.s.*	↘	↘
Conscientiousness	*n.s.*	*n.s.*	↗	↗	*n.s.*	↗	↗	↗

Emotional stability	*n.s.*	*n.s.*	↗	↗	*n.s.*	*n.s.*	↗	↗
Intellect/ Imagination	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	↗	↗	*n.s.*

Discussion

This study has attempted to delve into some of the psychological aspects of how chatbots based on LLMs respond, continuing initial exploratory studies involving the Turing test and the Theory of Mind [17,26]. Here we have focused on the Big Five personality traits exhibited by several chatbots, exploring what psychological traits these LLMs may have.

A particular problem is that when we asked a chatbot if it was responding as itself, it said no, it was responding as a typical human. So we then asked it to respond again in terms of its own characteristics and used only those confirmed responses for our analysis. It is helpful to focus first on the results of these innate responses.

The results indicated significant differences between the chatbots across all traits except Intellect/Imagination, which was high in all chatbots (at or near the maximum of the Likert scale used). This suggests that while chatbots generally exhibit varying degrees of intensity of the Big Five personality traits, their cognitive capabilities in terms of Intellect/Imagination are consistently similar. It can be considered that chatbots quite easily exhibit this trait, and it is a trait with significantly higher intensity relative to the general population of English- and Polish-speaking people. We conclude that Intellect/Imagination can be considered a Big Five trait characterizing chatbots, indicating that, in a way, they have characteristics which are "intellectually active and cognitively open, creative ... having ... a wide range of interests" [13]. Such a trait is found among artists and musicians, for example, who appear to be more open to experience than non-musicians [27].

For some chatbots, the language used influenced the traits which were uncovered. For instance, significant differences were noted for Gemini across various traits in different languages, whereas Gemini Advanced did not show significant differences across languages. We suspect there is a trend toward

"language free" features in the newer chatbots and thus largely "culture free" in the assessment of the Big Five personality traits revealed.

The pairwise comparisons between chatbots revealed several subtle differences, with the greatest number of differences observed in terms of Emotional stability – a trait also of significantly higher intensity attributed by chatbots to themselves relative to the human general populations used for comparison. It is possible that this trait represents a major difficulty for chatbots (hence the large differences between chatbots) when responding as them referring to emotions and internal states typical of humans. Are chatbots able to worry? Or experience depression? These are emotional states, often ambiguous for humans, and imbued with subjectivity, hence chatbots find it difficult to unambiguously respond to something relative to themselves that they do not possess (since they are artificial systems).

A distinctive observation is that ChatGPT-4o, in particular, displayed much lower Agreeableness than ChatGPT-3.5, highlighting the variability in personality traits between different versions of the same model. This suggests that updates and modifications in LLMs can lead to substantial changes in how chatbots based on these models manifest personality traits. It can be hypothesized that newer versions of chatbots will rate themselves increasingly lower in terms of typically human traits, such as Agreeableness and Extraversion, and progressively higher in terms of emotionally stable traits, i.e. traits that are revealed in social, emotionally saturated relationships, such as the ability to empathize or be cordial with people.

An interesting observation that should be underlined is the quite high variability of responses across trials. This has also been spotted in some other studies [18,20,28]. For example, we found that, especially for Extraversion and Agreeableness, the spread of results was very large (Figure 1). So are chatbots changing personality traits randomly, or is there some more advanced mechanism underneath? Do different conversations induce different personality traits chatbots? Are chatbots capable of modifying their traits during conversation to reflect the needs of the user, e.g. for medical support purposes? Is it possible to induce some dangerous traits in a chatbot? All these questions seem worthy of further investigation.

When comparing chatbots to the general human population, ChatGPT-3.5 and Gemini did not significantly differ from human norms for any trait, suggesting

these models closely mimic human-like personality traits. Conversely, ChatGPT-4o showed the greatest differences, particularly in Emotional Stability, indicating that some chatbots may exhibit personality traits that are distinctly different from those typically observed in human populations. There were multiple differences between personality traits of chatbots depending on the language used, and it appears that the chatbots probably behave differently in each language [29]. Although we tested only two languages, we observed that, for Extraversion, Gemini mimicked the difference between Polish and English populations, showing a lower score for English speakers.

In some instances the chatbots first response was as a typical human, and only after a second inquiry did it respond according to its AI characteristics. Some interesting observations come from comparison of these responses. Notably, when comparing these human-like responses to the general population, fewer significant differences were observed compared to when chatbots described themselves as LLMs. This suggests that chatbots may align more closely with human personality norms when simulating human behavior than when attempting to describe their own personality traits. The ability to characterize itself and distinguish itself from a typical human appears in the newer generation of chatbots. It is possible to interpret this ability to differentiate ‘self as a chatbot’ and ‘self as a typical human’ as moving toward an ability analogous to self-awareness in humans, which at this stage of research is not believable.

Additionally, significant differences between the two types of responses (human-like vs. self-description) were more prominent in advanced models like ChatGPT-4o and Gemini Advanced. These models exhibited increased Emotional stability and Conscientiousness but decreased Extraversion and Agreeableness in their responses as LLMs in comparison to human-like responses. This ability to distinguish characteristics of typical humans indicates that chatbots based on more advanced LLMs are more sophisticated [30]. Basic models, especially ChatGPT-3.5, do not differentiate humans from LLMs in terms of personality traits. ChatGPT-3.5 responded similarly to the average human, even when it was forced to respond according to its own characteristics. However, it should be noted that there are some changes in Gemini’s responses, so it seems it is better in perception of personality than ChatGPT-3.5. Nevertheless, it seems that chatbots based on more advanced LLMs can adjust their exhibited traits more flexibly, potentially leading to a more human-like portrayal when attempting to simulate human behavior. They seem able to imitate a typical human, but at the same time

when directly asked they seem to be able to recognize the differences as a LLM (first vs. second response, Table 4).

Our study shows that chatbots generally exhibit personality traits with high Emotional stability, Conscientiousness, and Intellect/Imagination. However, one can imagine that, depending on the purpose, it might be possible to design a chatbot with desired personality traits or which adapt to the situation [31] (e.g. an “agreeable" chatbot could be programmed to provide more empathetic and supportive responses [32]). Conversely, if one were designing a chatbot for enhanced Emotional stability (or its opposite pole, “neuroticism") this could involve programming the machine to recognize, and appropriately respond to, emotionally charged or stressful situations, and could be used, for example, in psychoeducation or psychological support [33].

This study has shown that chatbots based on LLMs reflect the Big Five personality traits of typical humans quite well, and this seems to be potentially useful for simulating psychological data, and presumably other data as well. It seems that these chatbots can, without any modifications, be used for training purposes or psychological questionnaires. They could be made to generate random responses, simulating different persons, but at the same time following a standard population distribution. One study has shown that chatbots tend to show less variability than humans [34]. Nevertheless, chatbots seem to offer a good way of training specialists before training with a real person.

In the end, the question remains, what is a „real” chatbot personality? It seems a chatbot typically operates with the personality of a typical human; it is only after additional instructions will it respond like a LLM – see where the responses change significantly in Table 3. This change might raise a concern that AIs have, embedded within their LLM architecture, a mechanism for pretending, and this behavior is not unlike some other negative behavior spotted in chatbots based on LLMs such as deliberate fabrications or ‘hallucinations’ [21,35,36]. Our findings suggest that chatbots show varying degrees of ability to distinguish themselves as LLMs in Big Five personality traits terms from typical humans, which are generated by them in response to specific stimuli (questionnaire items). Chatbots, "as themselves", reveal high levels of Intellect/Imagination and Emotional stability, and lower levels of Agreeableness. In more advanced versions they tend to change towards lower scores on the traits of Extraversion and Agreeableness, and higher scores on Emotional stability (since those typically human traits

cannot really exist in them). These three traits are the least unambiguous, and thus cause them the greatest difficulty in assessing against themselves, while Intellect/Imagination and Conscientiousness seem to be easier to identify.

Conclusions

This study underscores the complex interplay between chatbot LLM architecture, language, and the Big Five personality traits. It seems that chatbots are trained to respond as if they were a typical person rather than as “themselves”. Investigating chatbot personality traits is crucial for enhancing human–computer interaction by ensuring that chatbots based on LLMs are perceived as relatable and trustworthy. Understanding these traits can inform the development of more sophisticated and adaptive chatbots, leading to improved user satisfaction and more effective communication in diverse applications. Their ability to reflect human-like personality traits may have promising potential for simulating psychological data, e.g. for some educational purposes. Also the high Emotional stability of the tested chatbots may be a desirable property for use in medical support. Finally, our study shows that chatbots based on LLMs are able to differentiate their own personality traits from that of typical human. This is clearly not enough to show that chatbots have gained a capability analogous to self-awareness in humans, but it certainly shows that they are moving towards it.

References

1. Skjuve, M.; Følstad, A.; Fostervold, K.I.; Brandtzaeg, P.B. My Chatbot Companion - a Study of Human-Chatbot Relationships. *Int. J. Hum.-Comput. Stud.* **2021**, *149*, 102601, doi:10.1016/j.ijhcs.2021.102601.
2. Paliwal, S.; Bharti, V.; Mishra, A.K. Ai Chatbots: Transforming the Digital World. In *Recent Trends and Advances in Artificial Intelligence and Internet of Things*; Balas, V.E., Kumar, R., Srivastava, R., Eds.; Springer International Publishing: Cham, 2020; pp. 455–482 ISBN 978-3-030-32644-9.
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł. ukasz; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., 2017; Vol. 30.
4. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training.
5. Nicolescu, L.; Tudorache, M.T. Human-Computer Interaction in Customer Service: The Experience with AI Chatbots—A Systematic Literature Review. *Electronics* **2022**, *11*, 1579, doi:10.3390/electronics11101579.
6. Liu, S.; McCoy, A.B.; Wright, A.P.; Carew, B.; Genkins, J.Z.; Huang, S.S.; Peterson, J.F.; Steitz, B.; Wright, A. Leveraging Large Language Models for

Generating Responses to Patient Messages—a Subjective Analysis. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1367–1379, doi:10.1093/jamia/ocae052.
7. Lee, I.; Hahn, S. On the Relationship between Mind Perception and Social Support of Chatbots. *Front. Psychol.* **2024**, *15*, doi:10.3389/fpsyg.2024.1282036.
8. Huang, I.-C.; Lee, J.L.; Ketheeswaran, P.; Jones, C.M.; Revicki, D.A.; Wu, A.W. Does Personality Affect Health-Related Quality of Life? A Systematic Review. *PLOS ONE* **2017**, *12*, e0173806, doi:10.1371/journal.pone.0173806.
9. Udayar, S.; Urbanaviciute, I.; Rossier, J. Perceived Social Support and Big Five Personality Traits in Middle Adulthood: A 4-Year Cross-Lagged Path Analysis. *Appl. Res. Qual. Life* **2020**, *15*, 395–414, doi:10.1007/s11482-018-9694-0.
10. Wang, H.; Liu, Y.; Wang, Z.; Wang, T. The Influences of the Big Five Personality Traits on Academic Achievements: Chain Mediating Effect Based on Major Identity and Self-Efficacy. *Front. Psychol.* **2023**, *14*, doi:10.3389/fpsyg.2023.1065554.
11. Levesque, R.J.R. Ego Identity. In *Encyclopedia of Adolescence*; Levesque, R.J.R., Ed.; Springer New York: New York, NY, 2011; pp. 813–814 ISBN 978-1-4419-1694-5.
12. Goldberg, L.R. An Alternative “Description of Personality”: The Big-Five Factor Structure. *J. Pers. Soc. Psychol.* **1990**, *59*, 1216–1229, doi:10.1037/0022-3514.59.6.1216.
13. Struś, W.; Rowiński, T.; Cieciuch, J. The Polish Adaptation of the IPIP-BFM-50 Questionnaire for Measuring Five Personality Traits in the Lexical Approach. *Rocz. Psychol.* **2014**, *17*, 347–366.
14. Barrick, M.R.; Mount, M.K. THE BIG FIVE PERSONALITY DIMENSIONS AND JOB PERFORMANCE: A META-ANALYSIS. *Pers. Psychol.* **1991**, *44*, 1–26, doi:10.1111/j.1744-6570.1991.tb00688.x.
15. Schmitt, D.P.; Allik, J.; McCrae, R.R.; Benet-Martínez, V. The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description Across 56 Nations. *J. Cross-Cult. Psychol.* **2007**, *38*, 173–212, doi:10.1177/0022022106297299.
16. Rutinowski, J.; Franke, S.; Endendyk, J.; Dormuth, I.; Roidl, M.; Pauly, M. The Self-Perception and Political Biases of ChatGPT. *Hum. Behav. Emerg. Technol.* **2024**, *2024*, 1–9, doi:10.1155/2024/7115633.
17. Mei, Q.; Xie, Y.; Yuan, W.; Jackson, M.O. A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans. *Proc. Natl. Acad. Sci.* **2024**, *121*, doi:10.1073/pnas.2313925121.
18. Stöckli, L.; Joho, L.; Lehner, F.; Hanne, T. The Personification of ChatGPT (GPT-4)—Understanding Its Personality and Adaptability. *Information* **2024**, *15*, 300, doi:10.3390/info15060300.
19. Cao, X.; Kosinski, M. Large Language Models Know How the Personality of Public Figures Is Perceived by the General Public. *Sci. Rep.* **2024**, *14*, doi:10.1038/s41598-024-57271-z.
20. Kochanek, K.; Skarzynski, H.; Jedrzejczak, W.W. Accuracy and Repeatability of ChatGPT Based on a Set of Multiple-Choice Questions on Objective Tests of Hearing. *Cureus* **2024**, doi:10.7759/cureus.59857.
21. Jedrzejczak, W.W.; Skarzynski, P.H.; Raj-Koziak, D.; Sanfins, M.D.; Hatzopoulos, S.; Kochanek, K. ChatGPT for Tinnitus Information and Support: Response Accuracy and Retest after Three and Six Months. *Brain Sci.* **2024**, *14*, 465, doi:10.3390/brainsci14050465.
22. Goldberg, L.R. The Development of Markers for the Big-Five Factor Structure. *Psychol. Assess.* **1992**, *4*, 26–42, doi:10.1037/1040-3590.4.1.26.

23. Goldberg, L.R.; Johnson, J.A.; Eber, H.W.; Hogan, R.; Ashton, M.C.; Cloninger, C.R.; Gough, H.G. The International Personality Item Pool and the Future of Public-Domain Personality Measures. *J. Res. Personal.* **2006**, *40*, 84–96, doi:10.1016/j.jrp.2005.08.007.
24. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57*, 289–300, doi:10.1111/j.2517-6161.1995.tb02031.x.
25. Guenole, N.; Chernyshenko, O. The Suitability of Goldberg's Big Five IPIP Personality Markers in New Zealand: A Dimensionality, Bias, and Criterion Validity Evaluation. *N. Z. J. Psychol.* **2005**, *34*, 86–96.
26. Marchetti, A.; Di Dio, C.; Cangelosi, A.; Manzi, F.; Massaro, D. Developing ChatGPT's Theory of Mind. *Front. Robot. AI* **2023**, *10*, doi:10.3389/frobt.2023.1189525.
27. Gjermunds, N.; Brechan, I.; Johnsen, S.Å.K.; Watten, R.G. Personality Traits in Musicians. *Curr. Issues Personal. Psychol.* **2020**, *8*, 100–107, doi:10.5114/cipp.2020.97314.
28. Lechien, J.R.; Naunheim, M.R.; Maniaci, A.; Radulesco, T.; Saibene, A.M.; Chiesa-Estomba, C.M.; Vaira, L.A. Performance and Consistency of ChatGPT-4 Versus Otolaryngologists: A Clinical Case Series. *Otolaryngol. Neck Surg.* **2024**, *170*, 1519–1526, doi:10.1002/ohn.759.
29. Lewandowski, M.; Łukowicz, P.; Świetlik, D.; Barańska-Rybak, W. ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level Based on the Specialty Certificate Examination in Dermatology. *Clin. Exp. Dermatol.* **2024**, *49*, 686–691, doi:10.1093/ced/llad255.
30. Zhao, Y.; Huang, Z.; Seligman, M.; Peng, K. Risk and Prosocial Behavioural Cues Elicit Human-like Response Patterns from AI Chatbots. *Sci. Rep.* **2024**, *14*, doi:10.1038/s41598-024-55949-y.
31. Yorita, A.; Egerton, S.; Oakman, J.; Chan, C.; Kubota, N. Self-Adapting Chatbot Personalities for Better Peer Support.; IEEE, October 2019; pp. 4094–4100.
32. Zhou, L.; Gao, J.; Li, D.; Shum, H.-Y. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Comput. Linguist.* **2020**, *46*, 53–93, doi:10.1162/coli_a_00368.
33. Medeiros, L.; Bosse, T.; Gerritsen, C. Can a Chatbot Comfort Humans? Studying the Impact of a Supportive Chatbot on Users' Self-Perceived Stress. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *52*, 343–353, doi:10.1109/THMS.2021.3113643.
34. Giorgi, S.; Markowitz, D.M.; Soni, N.; Varadarajan, V.; Mangalik, S.; Schwartz, H.A. ``I Slept Like a Baby'': Using Human Traits To Characterize Deceptive ChatGPT and Human Text. *1st Int. Workshop Implicit Author Charact. Texts Search Retr. IACT'23*.
35. Alkaissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **2023**, doi:10.7759/cureus.35179.
36. Frosolini, A.; Franz, L.; Benedetti, S.; Vaira, L.A.; de Filippis, C.; Gennaro, P.; Marioni, G.; Gabriele, G. Assessing the Accuracy of ChatGPT References in Head and Neck and ENT Disciplines. *Eur. Arch. Otorhinolaryngol.* **2023**, *280*, 5129–5133, doi:10.1007/s00405-023-08205-4.