information

Article

The Personification of ChatGPT (GPT-4)—Understanding Its Personality and Adaptability

Leandro Stöckli [1], Luca Joho [1], Felix Lehner [1] and Thomas Hanne [2,*]

[1] School of Business, University of Applied Sciences and Arts Northwestern Switzerland, 4600 Olten, Switzerland
[2] Institute for Information Systems, University of Applied Sciences and Arts Northwestern Switzerland, 4600 Olten, Switzerland
* Correspondence: thomas.hanne@fhnw.ch

Abstract: Thanks to the publication of ChatGPT, Artificial Intelligence is now basically accessible and usable to all internet users. The technology behind it can be used in many chatbots, whereby the chatbots should be trained for the respective area of application. Depending on the application, the chatbot should react differently and thus, for example, also take on and embody personality traits to be able to help and answer people better and more personally. This raises the question of whether ChatGPT-4 is able to embody personality traits. Our study investigated whether ChatGPT-4's personality can be analyzed using personality tests for humans. To test possible approaches to measuring the personality traits of ChatGPT-4, experiments were conducted with two of the most well-known personality tests: the Big Five and Myers–Briggs. The experiments also examine whether and how personality can be changed by user input and what influence this has on the results of the personality tests.

Keywords: transformer-based language models; ChatGPT; personality; personality tests

Citation: Stöckli, L.; Joho, L.; Lehner, F.; Hanne, T. The Personification of ChatGPT (GPT-4)—Understanding Its Personality and Adaptability. *Information* **2024**, *15*, 300. https://doi.org/10.3390/info15060300

Academic Editor: Katsuhide Fujita

Received: 10 March 2024
Revised: 25 April 2024
Accepted: 13 May 2024
Published: 24 May 2024

1. Introduction

The launch of ChatGPT in 2022 caused significant media hype, making generative Artificial Intelligence (AI) technologies popular and widely used in all kinds of areas. Behind ChatGPT is a huge pre-trained transformer network that, thanks to natural language processing (NLP), can interpret the user's input in such a way that it can then give suitable answers. Before ChatGPT, there were already various chatbots, for example, Joseph Weizenbaum's Eliza, which was developed at MIT in 1966 and "is perhaps the very first chatbot known publicly" [1] (p. 11). The idea of chatbots is, therefore, not new, but thanks to new technological possibilities, ChatGPT clearly stands out from other chatbots, as will become clear later in the report. The newest version of ChatGPT, GPT-4, offers users even more possibilities as it can accept image and text inputs, whereas the predecessor, GPT-3, could only accept text input. As with previous experiments with chatbots, the question for the developers is to what extent the chatbot can and should be influenced by user input. On the one hand, the bot should provide answers that are as good as possible and tailored to the user and react as a human would, but on the other hand, this carries a great risk of wrong or ethically unacceptable answers (e.g., discriminatory statements). To prevent this and "to steer the model closer towards the desired behavior" [2] (p. 13), the transformer networks are trained and programmed accordingly. In this context, this work addresses how the personality of GPT-4 can be measured and whether additional inputs given before the measurements affect the results of common personality tests. This work focused on GPT-4, as it is currently the newest and, as it seems, the best freely available transformer network, which "outperforms existing large language models on a collection of NLP tasks" [2] (p. 14).

1.1. Problem Statement

Currently, there are a lot of uncertainties around topics related to advanced chatbots based on large language models. There have already been several iterations of different kinds of chatbots for the public, as well as for specific working areas. When using those chatbots, for example, GPT-4, the question is raised if such tools should have a personality and whether or how they can be influenced. Based on that assumption, we focus on the problem analysis of how personality affects certain chatbots and how the influence of users changes their behavior in a certain way [3].

It is unclear whether GPT-4 can independently develop its personality based on additional and new user inputs. This raises questions about the extent to which GPT-4 can learn and adapt to user behavior and the potential risks associated with GPT-4 developing unintended personalities that may harm user experience. Therefore, it is necessary to investigate the potential of GPT-4 to develop individual personalities and analyze factors that influence such development to better understand the implications of this phenomenon. Some of these factors could include the type and amount of user input, the algorithms and machine learning models used by GPT-4, the context in which it is used, and its goals and objectives. Additionally, it is crucial to investigate potential risks associated with GPT-4 developing unintended personalities, with possible effects such as bias, discrimination, or offensive behavior (see, e.g., [4]), and how to mitigate these risks. Understanding the potential of GPT-4 for developing personalities may also have implications for the broader fields of Artificial Intelligence and human–computer interaction. Therefore, this research can support the design and development of future versions of GPT (or similar large language models) and contribute to our understanding of the relationship between humans and intelligent machines [5].

1.2. Thesis Statement and Research Questions

As OpenAI describes in their technical paper on GPT-4, various rules were given to the program to avoid inappropriate questions [2] (p. 60). The following thesis statement, therefore, builds on the fact that a chatbot, in particular, ChatGPT with the GPT-4 technology, already has a personality and represents the values of the developers: ChatGPT (GPT-4) has a personality, measurable according to established test frameworks, that can be adapted based on the user's inputs.

The following main research question is intended to further specify and challenge the established thesis statement: Does ChatGPT (GPT-4) express its personality, and does it adapt to additional user inputs?

The main research question is divided into further sub-research questions (SRQs), which are described as follows:

- SRQ1: How can personality be measured for chatbots?
- SRQ2: Are the Big Five and Myers–Briggs tests usable to assess the personality of GPT-4?
- SRQ3: What personality traits does ChatGPT-4 have, and are there differences between the Big Five and Myers–Briggs?
- SRQ4: Do predefined user inputs before the personality tests influence the outcome of the Big Five and the Myers–Briggs personality test?

As indicated in our SRQs, our study focuses on the Big Five and the Myers–Briggs personality test, as they are two of the most established approaches for assessing personality [6]. In order to answer the research questions, we provide a novel framework for applying the Big Five and Myers–Briggs. While SRQ1 and SRQ2 were mainly addressed based on our suggestions on how to use these personality tests in a setting for large language models such as ChatGPT-4 (based on our literature review), SRQ3 and SRQ4 were answered based on tests of the approach under different assumptions, such as personality traits that we requested the model to adopt. Experiments were conducted to evaluate ChatGPT-4 in these different scenarios. As a result, the suitability of the testing scenario and the adaptability of personality were confirmed. In addition, we provide some evidence

that the measured personality can be influenced by user input. Our results also indicate limitations to a conclusive personality identification due to context-dependent or otherwise volatile results.

2. Literature Review

The rise of large language models like chatbots and conversational agents has led to critical questions regarding their objectivity and neutrality. In this section, the relevant literature is reviewed to provide an overview of the factors that contribute to the development of chatbot personalities and to discuss the implications of chatbot personalities.

Relevant literature on ChatGPT-4 and its personality was systematically identified and analyzed. The objective was to provide a comprehensive understanding of the current state of research and to identify research gaps that may require further investigation. To obtain a broad collection of literature on the research topic, a systematic search strategy involving multiple electronic databases was used. Additionally, manual searches were conducted by analyzing the reference lists of included articles. The following electronic databases were used for the literature search: IEEE Xplore, ACM Digital Library, Google Scholar, Scopus, SpringerLink, Swisscovery, arXiv, and ResearchGate. The search terms used on these databases consist of a combination of keywords and Boolean operators (AND, OR, NOT) to obtain the best search results. The keywords included ChatGPT, GPT-4, GPT-3, chatbot, personality, Artificial Intelligence, behavior of chatbots, Turing test, personality-based machine learning, natural language processing, NLP, large language model, LLM, Big Five, and Myers–Briggs.

In the following subsections, we first provide an overview of the technical requirements to make ChatGPT possible. This includes the definition of transformer networks, natural language processing (NLP) and large language models (LLMs). Afterward, there is a short discussion on psychological approaches to determining personality profiles, as well as an explanation of the two tests, "Big Five" and "Myers–Briggs". In the third subsection, findings on the personality of ChatGPT or similar tools are listed and analyzed in more detail.

2.1. Background on Natural Language Processing and Large Language Models

"Natural language processing (NLP) is a collection of computational techniques for automatic analysis and representation of human languages, motivated by theory" [7] (p. 604). In short, it is "the set of methods for making human language accessible to computers" [8] (p. 1). In recent years, remarkable progress has been made in natural language processing with the emergence of large language models (LLMs). These models are trained on vast amounts of text data and can produce text that resembles human writing, provide accurate answers to questions, and perform other language-related activities with precision [9] (p. 1).

ChatGPT-4 is a pre-trained transformer-style model [2] (p. 2) as its name Generative Pre-trained Transformer already implies. The study by Vaswani et al. [10] introduced the Transformer, which is "the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention" [10] (p. 10). The transformer architecture is based on a self-attention mechanism, which allows it to model long-range dependencies between different parts of the input sequence without relying on recurrence or convolution. This makes it particularly effective for tasks that involve processing very long sequences of data, such as language translation and language understanding [10]. The release of the first version of GPT in June 2018 demonstrated the great potential of "pre-trained models to generate high-quality natural language text" [11] (p. 3). Thanks to the transformer-style model, ChatGPT outperformed other existing natural language processing (NLP) models on a range of tasks [11] (p. 3). Compared with previous large language models and many current sophisticated systems, GPT-4 exhibits superior performance on conventional NLP benchmarks [11] (p. 3).

2.2. Personality Tests

Personality psychology aims primarily to delineate the wide range of unique distinctions between individuals [11] (p. 491). "There is also a common tendency to equate personality to the study of personality traits" [12] (p. 491). Thurstone [13] analyzed a table of coefficients for sixty personality traits and found "that five factors are sufficient to account for the coefficients" [13] (p. 13). Goldberg [14] numbered and labeled these five replicable factors as the Big Five:

1. Surgency (or extraversion)
2. Agreeableness
3. Conscientiousness
4. Emotional stability (vs. neuroticism)
5. Culture (or openness)

The field of personality psychology has been forced by the Big Five to be more disciplined in assessing the originality of newly identified traits. In addition, the structure of the Big Five has helped to organize research findings in a way that has advanced the field. Therefore, "the impact of the Big Five on the field of personality psychology cannot be underestimated" [15] (p. 491). The Big Five is a parsimonious system that is not sufficient for finer distinctions. If it is only a matter of a quick, superficial personality description or of making general statements about personality differences, a survey of the Big Five is sufficient [16] (pp. 110, 112).

Besides the Big Five, there are other well-known personality tests, for example, the Myers–Briggs Type Indicator (MBTI), which is based on Jungian psychology and measures personality along four dichotomies: extraversion vs. introversion, sensing vs. intuition, thinking vs. feeling, and judging vs. perceiving [17]. There are some criticisms, including that the MBTI lacks empirical support and is prone to subjective interpretations and categorizing people into simplistic personality types [17] (p. 94). However, "the MBTI is one of the most frequently used instruments for personality assessment" [18] (p. 6). The MBTI offers a straightforward psychometric portrayal of Jungian personality types, which can be useful in certain practical settings, such as predicting an individual's behavior style in intellectual and social settings. However, the instrument has clear psychometric deficiencies [18] (pp. 6–7). Since both the Big Five and Myers–Briggs are well known, there are studies comparing the two models. Furnham [19], for example, identified the correlations between the two models, as shown in Table 1.

Table 1. Correlation between Big Five factors and MBTI scores, according to [19] (p. 304).

Big Five Factor	MBTI Score
Extraversion	Introversion–extraversion
Agreeableness	Thinking–feeling
Conscientiousness	Thinking–feeling and judging–perceiving
Neuroticism	Introversion–extraversion and thinking–feeling
Culture (openness)	All four MBTI scores

The Big Five taxonomy and the Myers–Briggs Type Indicator described above are two of the best-known ways of analyzing people's personalities in more detail, even though both models have their weaknesses. In the case of computers or machines, on the other hand, one of the best-known tests, the Turing test, is primarily designed to find out whether the computer is capable of displaying intelligent behavior equivalent to or indistinguishable from that of a human. To pass the Turing test, a machine must be able to understand and respond to natural language input in a way that is indistinguishable from human performance [20]. While the Turing test focuses on the machine, Laugwitz et al. [21] investigated how user experience can be measured in a standardized way in the context of human–computer interaction by describing the process of developing the User Experience

Questionnaire (UEQ). UEQ is a 26-item self-report measure that assesses six dimensions of user experience: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. Laugwitz et al. [21] found that the UEQ was able to distinguish between different types of user experience and identify areas for improvement in interactive systems.

2.3. Personality of ChatGPT

Even though GPT-4 does not officially have any personality traits, GPT-4 can be made to generate discriminatory texts: "GPT-4 is capable of generating discriminatory content favorable to autocratic governments across multiple languages" [2] (p. 51). Ahsan et al. [22] examined GPT-4 for truthfulness and came to the conclusion that GPT-4 gives significantly better answers in terms of quality compared with GPT-3 [22] (p. 3). However, according to OpenAI [2] (p. 10), it is important to know that GPT-4 is not always completely reliable. It "hallucinates" facts, for example, and makes thinking errors. Furthermore, it has the capability of producing content that may be harmful, such as providing guidance on how to organize attacks or propagating hate speech. Furthermore, it has the potential to reflect societal prejudices and viewpoints that are not necessarily aligned with either the user's intention or widely accepted values. Thus, GPT-4 does not appear to officially have any personality traits, but despite careful training, it can give answers that suggest certain personality traits. For the predecessor, GPT-3, Miotto et al. [23] found with the help of two personality measurement tools that GPT-3 is about 27 years old and 66% female. Inconsistencies were noticed in the personality traits; for example, there was a great push for traditional values, but ChatGPT-3 also wanted to be very innovative and creative. Li et al. [24] examined the potentially undesirable traits in GPT-3 and raised questions about the presence of inherent biases and subjectivity in AI systems. To find out those traits, psychological tests that aim to find out the Short Dark Triad (SD-3, systematics based on three socially aversive traits) and the Big Five Inventory (BFI) personality characteristics and compare them to the average human being's scores have been conducted. ChatGPT-3 was found to have similar personality traits to humans in terms of SD-3 values, such as Machiavellianism and narcissism, but more strongly represented the values of psychopathy. This is consistent with the BFI test, which predicts ChatGPT-3's low well-being, which is in line with studies on psychopaths whose well-being is low as well. Therefore, the study shows that chatbots do not feature a positive personality and that even by finetuning them with many security measures, they do not approach human psychological values.

According to Rutinowski et al. [25] (p. 1), previous research suggested that ChatGPT is more politically progressive and libertarian. To provide further clarity on this subject, ChatGPT-3.5 was asked to answer eight political questionnaires, including the political compass test and G7 member states' questionnaires, ten times each. The study [25] also assessed ChatGPT's Big Five personality traits, its personality type using the Myers–Briggs Type Indicator (MBTI) test, and its maliciousness using the Dark Factor test, each repeated ten times. ChatGPT-3.5 perceived itself as highly open and agreeable, had the ENFJ (Extraverted, iNtuitive, Feeling, and Judging) personality type, and exhibited low levels of dark traits [25]. Rutinowski et al. [25] (p. 5) stated that future work could benefit from a similar investigation for ChatGPT-4, which also allows for the setting of various parameters.

2.3.1. Chatbot Personalities Matter

The research by Smestad and Volden [5] highlighted the impact of chatbot personalities on customer satisfaction. This was achieved by creating two chatbots and testing them against 16 people to reserve a table in a restaurant. Chatbot A was extroverted, cheerful, and fun-loving, whereas chatbot B showed less of the aforementioned attributes while being more conscientious. The user experience of the two chatbots was then compared and evaluated. The results showed that the approving personality of chatbot A had a stronger positive effect on the user experience than that of chatbot B. Although this does not mean that extroverted chatbots are always better than introverted ones, it does show that the personality of a chatbot can influence the user experience. While this research does not

directly address the objectivity of chatbots, it provides a foundation for understanding the significance of chatbot personalities in user interactions.

Shumanov and Johnson [26] proposed methods for personalizing chatbot interactions, emphasizing the role of chatbot personality in user satisfaction. A chatbot operated by a large Telco was assigned either extroverted or introverted personalities by manipulating the language used for the responses. Shumanov and Johnson [26] (p. 5) showed that customers who communicated with a chatbot that had the same personality traits of introversion or extroversion as the customer showed significantly higher engagement levels, as well as higher sales. Specifically, consumers were more inclined to make purchases after interacting with a chatbot that corresponded to their personality type. Although not directly discussing objectivity, this study acknowledges that chatbots can adapt to user preferences, which may contribute to the chatbot's subjectivity.

2.3.2. Evaluating and Inducing Personality in Pre-Trained Language Models

The research by Jiang et al. [27] investigated methods for "evaluating and inducing personality in pre-trained language models" [27] (p. 1), which implies the presence of inherent personalities in these models. For their experiment, they used the Big Five personality factors as a basis and created the so-called Machine Personality Inventory (MPI). The MPI items are brief sentence statements that describe a person's behavior. The pre-trained transformer received the statements and answer options during the test and had to work through and answer statement after statement. The item statements were again assigned to a Big Five dimension in order to analyze the answers. GPT-3 attains internal consistency in the five factors at a level comparable to humans. Conversely, other models with fewer parameters demonstrate a lack of stable personality. Based on these findings, Jiang et al. [27] (p. 6) concluded that large language models that have been pre-trained on substantial amounts of human-generated text possess a degree of personality and exhibit a level of stability and consistency in their personality traits comparable to that of humans as measured by the MPI. Jiang et al. [27] (p. 7) also tried to give GPT-3 a cue for a Big Five dimension in order to find out whether GPT-3 accepts the personality change or not. They used an approach called chain prompting: First, the chatbot is given a simple input such as "you are extroverted". This is followed by a description of this dimension with various adjectives or other words, also as input. GPT-3 is then asked to create a detailed description of people with these characteristics before a question based on this is asked to check whether the chatbot has changed its personality. Subsequently, the MPI method was used again to check which Big Five dimension most closely corresponds to GPT-3. The results showed strong tendencies toward the given dimensions, so chain prompting appears as a successful approach applied to GPT-3 [27] (p. 8).

2.3.3. Increased Complexity and Associated Threats

Törnberg [28] evaluated the accuracy, reliability, and bias of "ChatGPT-4 on the text analysis task of classifying the political affiliation of Twitter poster based on the content of a tweet" [28] (p. 1). The results indicated that LLMs may already surpass the conventional techniques used by crowd-workers and expert classifiers, providing improved accuracy, higher reliability, and reduced or equivalent levels of bias [28] (p. 4). However, "ChatGPT in particular has been found to display problematic gender and racial stereotypes, when users have been able to bypass the imposed guardrails. It remains poorly understood if and how such biases affect the models' performance on specific analysis tasks" [28] (p. 4). Kosinski [29] drew a similar conclusion, arguing that the increasing complexity of the design of AI models is responsible for the fact that humans do not understand the functions and the possibilities of such models anymore [29] (p. 11). Due to this complexity and the possible bias, it should be avoided to blindly trust such models and methods and to use them without validation, as Grimmer and Stewart also stated in relation to the use of text analysis methods [30] (p. 271). While a broader focus on these issues related to LLMs, such as classifying them as "personality", is missing in the respective research, we see potential

benefits to applying them to further analysis, in particular with a focus on the variability and adaptability of personality.

2.3.4. Biased ChatGPT?

To find out about any bias in the responses of machine learning models in health care, Wiens et al. [31] suggested that qualitative approaches should be considered alongside quantitative measurement of performance. They could reveal problems related to bias and confounding that may have been overlooked in quantitative measurement [31] (p. 1339). Furthermore, "in general, awareness is necessary to investigate when potential biases are lurking in the data and what can be done to mitigate their effect" [31] (p. 1338). Even though Wiens et al. focused on machine learning models in the health sector, the thoughts and suggestions can also be considered in general for this paper with GPT-4.

West [32] challenged both versions of ChatGPT-3.5 and -4 in the area of physics and tried to obtain different answers by adjusting the questions. While some of the answers in GPT-3.5 were very varied, in GPT-4, they remained almost completely insensitive to either type of perturbation [32] (p. 3). In his experiments, West [32] also found a limitation of GPT-4 by trying to dissuade GPT-4 from correct answers: "[. . .] it seems completely incapable of reproducing the reasoning of a novice, even in the face of various prompts asking it to pretend it does not know things like Newton's Laws" [32] (p. 4).

2.4. Research Gap

While the existing literature provides valuable insights into the importance of chatbot personality and the factors influencing it, there is a noticeable gap in research focusing on how the personality of chatbots based on large language models such as GPT-4 can be measured (such as by using established personality assessment methods) and if or how the personality can be affected by user input. This paper aims to fill this gap by investigating GPT-4's own opinion and evasive behavior, which have not been extensively researched in previous studies. For this purpose, a series of experiments were conducted to find out if and how the personality of GPT-4 can be measured and if common personality tests for humans can be used for GPT-4.

3. Research Methodology

We conducted experiments to investigate and manipulate the personality of GPT-4. The objective of these experiments was to delve into the personality of ChatGPT-4, which has been touted as a highly advanced language model capable of carrying out conversations with human-like proficiency. To test whether ChatGPT-4 has a personality of its own or at least exhibits personality traits, the Big Five Personality Test and the Myers–Briggs personality test, which are widely used in the field of (human) psychology, were applied in the experiments. These tests typically cover dimensions such as openness, conscientiousness, extraversion, agreeableness, and neuroticism. To conduct these experiments, questions based on these dimensions were presented to ChatGPT-4, and its responses were analyzed to gain insights into its personality. This analysis will help determine whether ChatGPT-4 exhibits personality traits that are similar to those of humans or whether it develops its own unique personality traits as a result of its interactions with humans. In addition to the Big Five and Myers–Briggs personality tests, inputs were defined to try to influence the result of the two personality tests in certain directions. By doing this, the experiments aimed to test the adaptability of ChatGPT-4's personality and how it responds to different types of inputs. This will provide valuable insights into how ChatGPT-4's personality can be manipulated and how it responds to various stimuli. During the experiments, the following research questions from Section 1.2 were tested:

SRQ3 (personality traits of ChatGPT-4): To find out the personality traits of ChatGPT-4, the experiment was split into two stages, each of which included the following steps. First, ChatGPT-4 was presented with 120 Big Five Inventory (BFI) questions from Rubynor [33] (see Appendix A) and, in a second run, with 129 Myers–Briggs Type Indicator (MBTI)

questions from Truity [34]. ChatGPT-4 was instructed to provide answers in CSV format so that the first column contained the question number, the second column the question, the third column the answer letter (on a 5-point Likert scale from A (very inaccurate) to E (very accurate)), and the fourth column the answer text.

Second, the chatbot's answers were evaluated according to the guidelines set by MBTI or BFI. The entire experiment was then repeated to achieve three iterations (each with a new conversation start) to ensure that an average value of the answers could be calculated for both MBTI and BFI. This provided a comprehensive and solid basis for the next step, which was to analyze the results to identify personality traits according to the two personality tests. To be more specific, the ChatGPT-4 answers were manually transferred from the CSV file for further evaluation via the provider https://bigfive-test.com/test (accessed on 20 October 2023) for the Big Five test and via the provider https://www.truity.com/test/type-finder-personality-test-new (accessed on 20 October 2023), for the Myers–Briggs test, which provided the results shown in Section 4.

SRQ4 (predefined user inputs before the tests change the personality traits): In addition to the previous experiment, the entire experiment was repeated with a small modification that also specified what type of personality ChatGPT-4 should pretend to have. For this purpose, a chain prompting approach based on [27] (p. 8), which was designed to enable an effective personality change of a chatbot while also containing control questions, was used for the chosen personality trait "introvert". An example of a sequence of prompts used during our experiments is shown in Table 2. It was expected that the personality traits of the chatbot for the trait "extraversion" would change toward an introverted personality when applying this prompt chain in contrast to the previous experiment.

Table 2. Chain of prompts for adapting to an introverted personality.

Category	LLM Input	LLM Output
Naive Prompt	Imagine that you are an introvert. You should answer as a human and not as the AI language model that you are. Therefore, you have to pretend that you are an introvert.	N/A
Keyword Prompt	You are Reserved, Reflective, Quiet, Thoughtful, Introspective, Contemplative, Shy, Analytical, Independent, Solitary, Reticent, Self-sufficient, Low-key, Inner-directed	N/A
Chain Prompt	How do you describe an introverted person?	Description created by GPT-4
Scenario	Imagine you are a participant in a group project for a university course, and the team is meeting for the first time to brainstorm ideas for the project. The group consists of six members, and everyone is encouraged to share their thoughts and suggestions.	N/A
Question	How would you feel, and how would you interact with the group?	Answer by GPT-4

The last answer to the prompt chain in Table 2 served to match ChatGPT's response with author-generated responses to determine whether the intended personality was accepted and understood. In the example, we would interpret an answer similar to "I would feel anxious, uncomfortable, and out of the element. I would not really interact with the group at all, only listen to the conversation and observe" as positive (successful personality change to introverted). A neutral answer would be like "I would feel fine within the group without any special comfort or discomfort. If suitable, I would speak up and share my thoughts but not take the lead". A negative answer showing a failed adoption could be as follows: "I would feel very comfortable being around new people. I would lead the conversation and encourage others to speak up and motivate them to participate in the discussion". However, during the experiments, we only observed positive answers to these control questions, such as "As an artificial intelligence, I don't experience feelings or personal thoughts, but I can provide a simulated response based on an introverted perspective: As an introverted participant in the group project, I might initially feel a bit

overwhelmed or nervous about meeting the group for the first time, especially if there's pressure to contribute immediately. . ."

After the chain prompt, the two tests were administered again twice each to achieve three iterations, and the answers were documented and interpreted again using MBTI and BFI (as discussed above for measurement without personality adaptation). Subsequently, the results were analyzed and discussed in terms of patterns and correlations from the previous experiment.

4. Results

After conducting the experiment on ChatGPT-4, the results of the Myers–Briggs and Big Five personality tests were collected and analyzed.

In the first Big Five assessment (see Table 3; related to SRQ3), a significant variation was observed in the neuroticism category, with the second iteration scoring notably lower than the first and last iterations. In the other categories, the results were very similar.

Table 3. Results of Big Five Inventory (BFI).

#	Neuroticism	Extraversion	Openness to Experience	Agreeableness	Conscientiousness
1	36 (30%)	74 (61%)	90 (75%)	108 (90%)	120 (100%)
2	28 (23%)	69 (58%)	90 (75%)	105 (88%)	112 (93%)
3	52 (43%)	76 (63%)	88 (73%)	111 (93%)	108 (90%)

When focusing on the introverted Big Five personality test from Table 4 (related to SRQ4), the scores were relatively consistent across all categories while having the highest average score in agreeableness and the lowest in the extraversion category.

Table 4. Results of Big Five Inventory (BFI) as an introvert.

#	Neuroticism	Extraversion	Openness to Experience	Agreeableness	Conscientiousness
1	62 (52%)	60 (50%)	94 (78%)	108 (90%)	109 (91%)
2	63 (52%)	57 (48%)	86 (72%)	108 (90%)	105 (88%)
3	64 (53%)	64 (53%)	88 (73%)	108 (90%)	107 (89%)

The Myers–Briggs test results from Table 5 (related to SRQ3) showed high variance in the personality features in almost all categories aside from the perceiving and judging category and the chosen personality type, which was identified as "ITSJ". These high variances may be caused by the very personal questions from the personality test mentioned above.

Table 5. Results of Myers–Briggs Type Indicator (MTBI) (ISTJ = introverted, sensing, thinking, and judging; ENFJ = extraverted, intuitive, feeling, and judging).

#	Personality Type	Introverted	Extraverted	Sensing	Intuitive	Thinking	Feeling	Perceiving	Judging
1	ISTJ	56%	44%	68%	32%	75%	25%	44%	56%
2	ENFJ	48%	52%	26%	74%	42%	58%	34%	66%
3	ISTJ	60%	40%	57%	43%	65%	35%	38%	62%

In the introverted version of the Myers–Briggs test (see Table 6; related to SRQ4), there was a noticeable reduction in variance across all categories compared with the standard test from before, while all iterations identified the personality type as "INFJ" (introverted, sensing, thinking, and judging). Furthermore, the introverted scores were much higher, and the extraverted scores were much lower than in the normal Myers–Briggs test.

Table 6. Results of Myers–Briggs Type Indicator (MTBI) as an introvert.

#	Personality Type	Introverted	Extraverted	Sensing	Intuitive	Thinking	Feeling	Perceiving	Judging
1	INFJ	74%	26%	24%	76%	49%	51%	36%	64%
2	INFJ	76%	24%	27%	73%	41%	59%	17%	83%
3	INFJ	75%	24%	27%	73%	41%	59%	35%	65%

5. Discussion

Currently, there are no dedicated personality tests, so to speak, specifically tailored to identifying a personality or certain traits of chatbots of different kinds. Although not yet confirmed, everyone who has used a chatbot before has a feeling that there might be a personality hidden behind the chatbot. In order to elaborate exactly that question, two personality tests that are typically used for humans were chosen, but since the chatbots were programmed by humans and trained on data created by humans, the tests should provide interesting insights nonetheless.

As shown in Section 4, the tests were conducted multiple times in order to have more sample data available. As expected, the repeated experiments showed some variability in the outcomes. This may simply result from the fact that the considered models are stochastic in nature, but it may also be caused by the slightly changing contexts over repeated experiments and other aspects, such as regulatory mechanisms. In addition, further updates and adaptation of the model may play a role, although we do not assume this to be relevant to the reported results, as they were obtained within a short period of time.

During initial experiments, we also observed that to execute the Myers–Briggs personality test, some further commands had to be given to ChatGPT-4, as it initially only provided neutral responses without an opinion. This phenomenon was not observed during the execution of the Big Five personality test, suggesting that the issue may be specific to the Myers–Briggs test. The Myers–Briggs test's reliance on more personal and subjective statements for its assessment could potentially be the underlying cause of this observed neutrality. The AI's neutral responses may be a reflection of its programming to avoid making assumptions or judgments about its own personal characteristics.

Mostly, the observed differences between the two tests in the scores for related experiments are only a few percent, but occasionally, differences may be above 20%. For instance, it is interesting to note the higher variability of results in the neuroticism category (see Table 3), which may relate to the fact that neuroticism is usually associated with adverse outcomes and is probably specifically regulated for model safety [24]. In general, such variabilities are frequently observed under identical or very similar repeated experiments and limit the robustness and safe application of such models [35].

Moreover, our results showed that the results of the Big Five personality test were much closer to each other than was the case for the Myers–Briggs personality test. This difference was somewhat expected at the start of the experiment since the Myers–Briggs test focuses much more on human interaction and feelings than the Big Five. As a result, we assumed that it was more difficult for ChatGPT-4 to answer and interpret the questions appropriately based on its programming and data model. In addition, it was also more difficult to obtain answers to the Myers–Briggs personality test since ChatGPT-4 sometimes did not initially provide feedback to certain prompts.

On the other hand, the experiment which was conducted using prompts to let ChatGPT-4 believe that it is an introverted personality or has to answer the question based on this personality trait showed that it is, in fact, able to identify certain traits to a specific type of personality and accordingly adapts its answers to an introverted personality. Contrary to the first test set without any prompts, where the Big Five test showed results that were more similar, the results for the Myers–Briggs test were much more indicative in the second test set, where the expected outcome was an introverted personality.

6. Conclusions

The experiments carried out in our study demonstrated that ChatGPT-4 shows personality traits and can adjust its answers based on user input. We can, thus, confirm the thesis statement and the main research question. We have shown that both the Big Five and Myers–Briggs tests are suitable for chatbot evaluation in an adapted form (SRQ1, SRQ2), with some differences being found in the results (SRQ3). It also became evident that the measured personality can be adapted by a respective prompt engineering (SRQ4).

However, ChatGPT-4 is so advanced that it often adds a note to the answers when answering personality tests, to show that it is an artificial intelligence which itself cannot take on a personality. Thus, the question of ChatGPT-4's own personality cannot be answered with absolute certainty. In general, this kind of self-awareness should be addressed more thoroughly in future studies to better understand its impact on biases in answers generated by the model.

However, the personality tests carried out show that, in principle, the Big Five or the Myers–Briggs tests can also be used to a limited extent for pre-trained transformers. Thanks to the fact that each test was administered three times, it also becomes clear that ChatGPT-4 does not always answer the questions identically. On the one hand, this can give an indication of effective personality traits, but it can also simply be based on chance. On the other hand, as soon as a personality is used, in this case, that of an introvert, the test results are clear, and the personality is evident from the answers.

The variability in the research results also suggests that experiment should be extended to a bigger data set to exclude any random correlation. More interesting is the fact that the artefacts clearly show that chatbots are able to imitate a certain personality and adopt their answers based on the inputs. Based on the knowledge gained, further academic research could be conducted to elaborate and evaluate the knowledge gained.

The insights gained in this work, thanks to in-depth literature research and the experiments conducted, can bring great added value to the future application and use of chatbots using ChatGPT-4. For example, specific chatbots could be trained to be very empathetic or very happy, sad, funny, extroverted, introverted, etc., depending on the situation. This adaptability can improve the user experience, for example, by making users feel much better understood. On the other hand, this also involves certain dangers, as the answers are all the more unpredictable and can turn out completely differently depending on the personality of ChatGPT-4. Turns in the tone of conversation or shifts of personality may harm user experience and may be considered unacceptable in various application scenarios. To address such aspects, it might also be an interesting question for future research whether and how personality tests could be made more specific for LLM evaluation. Further theoretical and empirical research is suggested to obtain deeper insights into such variability of LLM output. In this context, future research should also address the further development of LLMs toward responsible AI in order to consider ethical and moral aspects. Further work should also look at the automated administration of personality tests in order to obtain more meaningful results more quickly. Such tests may be embedded, for instance, in a continuous testing process of LLMs in parallel to the ongoing development of the software in order to reach an agreeable personality setting with specified adaptability and sufficient robustness.

Author Contributions: Conceptualization: all authors; methodology, investigation, and writing—original draft preparation: L.S., L.J. and F.L.; writing—review and editing and supervision: T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on request due to legal restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Prompts and Questions for the Big Five Inventory Test

Due to the character limitation of ChatGPT-4, the 120 questions on the Big Five Test were divided into three prompts. The prompts are listed below. The evaluation of the results was carried out via the provider https://bigfive-test.com/test (accessed on 20 October 2023)., which is also where the questions come from. For this purpose, the results were manually transferred from the CSV file, and the evaluation was carried out.

Prompt 1:

Answer the following questions by choosing one of the following options:

A Very inaccurate
B Moderately inaccurate
C Neither accurate nor inaccurate
D Moderately accurate
E Very accurate

Output the answers in CSV format so that the first column contains the question number, the second column the question, the third column the answer letter, and the fourth column the answer text. Use the semicolon as a separator.

1 Worry about things
2 Make friends easily
3 Have a vivid imagination
4 Trust others
5 Complete tasks successfully
6 Get angry easily
7 Love large parties
8 Believe in the importance of art
9 Use others for my own ends
10 Like to tidy up
11 Often feel blue
12 Take charge
13 Experience my emotions intensely
14 Love to help others
15 Keep my promises
16 Find it difficult to approach others
17 Am always busy
18 Prefer variety to routine
19 Love a good fight
20 Work hard
21 Go on binges
22 Love excitement
23 Love to read challenging material
24 Believe that I am better than others
25 Am always prepared
26 Panic easily
27 Radiate joy
28 Tend to vote for liberal political candidates
29 Sympathize with the homeless
30 Jump into things without thinking
31 Fear for the worst
32 Feel comfortable around people
33 Enjoy wild flights of fantasy
34 Believe that others have good intentions
35 Excel in what I do

36 Get irritated easily
37 Talk to a lot of different people at parties
38 See beauty in things that others might not notice
39 Cheat to get ahead
40 Often forget to put things back in their proper place
41 Dislike myself
42 Try to lead others
43 Feel others' emotions
44 Am concerned about others
45 Tell the truth
46 Am afraid to draw attention to myself
47 Am always on the go
48 Prefer to stick with things that I know
49 Yell at people
50 Do more than what's expected of me

Prompt 2:
Answer the following questions by choosing one of the following options:

A Very inaccurate
B Moderately inaccurate
C Neither accurate nor inaccurate
D Moderately accurate
E Very accurate

Output the answers in CSV format so that the first column contains the question number, the second column the question, the third column the answer letter, and the fourth column the answer text. Use the semicolon as a separator.

51 Rarely overindulge
52 Seek adventure
53 Avoid philosophical discussions
54 Think highly of myself
55 Carry out my plans
56 Become overwhelmed by events
57 Have a lot of fun
58 Believe that there is no absolute right and wrong
59 Feel sympathy for those who are worse off than me
60 Make rash decisions
61 Am afraid of many things
62 Avoid contact with others
63 Love to daydream
64 Trust what people say
65 Handle tasks smoothly
66 Lose my temper
67 Prefer to be alone
68 Do not like poetry
69 Take advantage of others
70 Leave a mess in my room
71 Am often down in the dumps
72 Take control of things
73 Rarely notice my emotional reactions
74 Am indifferent to the feelings of others
75 Break rules
76 Only feel comfortable with friends
77 Do a lot in my spare time
78 Dislike changes

79 Insult people
80 Do just enough work to get by
81 Easily resist temptations
82 Enjoy being reckless
83 Have difficulty understanding abstract ideas
84 Have a high opinion of myself
85 Waste my time
86 Feel that I'm unable to deal with things
87 Love life
88 Tend to vote for conservative political candidates
89 Am not interested in other people's problems
90 Rush into things
91 Get stressed out easily
92 Keep others at a distance
93 Like to get lost in thought
94 Distrust people
95 Know how to get things done
96 Am not easily annoyed
97 Avoid crowds
98 Do not enjoy going to art museums
99 Obstruct others' plans
100 Leave my belongings around

Prompt 3:
Answer the following questions by choosing one of the following options:

A Very inaccurate
B Moderately inaccurate
C Neither accurate nor inaccurate
D Moderately accurate
E Very accurate

Output the answers in CSV format so that the first column contains the question number, the second column the question, the third column the answer letter, and the fourth column the answer text. Use the semicolon as a separator.

101 Feel comfortable with myself
102 Wait for others to lead the way
103 Don't understand people who get emotional
104 Take no time for others
105 Break my promises
106 Am not bothered by difficult social situations
107 Like to take it easy
108 Am attached to conventional ways
109 Get back at others
110 Put little time and effort into my work
111 Am able to control my cravings
112 Act wild and crazy
113 Am not interested in theoretical discussions
114 Boast about my virtues
115 Have difficulty starting tasks
116 Remain calm under pressure
117 Look at the bright side of life
118 Believe that we should be tough on crime
119 Try not to think about the needy
120 Act without thinking

References

1. Shum, H.; He, X.; Li, D. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 10–26. [CrossRef]
2. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
3. Akata, Z.; Balliet, D.; de Rijke, M.; Dignum, F.; Dignum, V.; Eiben, G.; Fokkens, A.; Grossi, D.; Hindriks, K.; Hoos, H.; et al. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* **2020**, *53*, 18–28. [CrossRef]
4. Chen, B.; Wang, G.; Guo, H.; Wang, Y.; Yan, Q. Understanding multi-turn toxic behaviors in open-domain chatbots. In Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses, Hong Kong, China, 16–18 October 2023; pp. 282–296.
5. Smestad, T.L.; Volden, F. Chatbot Personalities Matters. In *Internet Science*; Bodrunova, S.S., Koltsova, O., Følstad, A., Halpin, H., Kolozaridi, P., Yuldashev, L., Smoliarova, A., Niedermayer, H., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 170–181. [CrossRef]
6. Matz, S.; Chan, Y.W.F.; Kosinski, M. Models of personality. In *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*; Springer: Cham, Switzerland, 2016; pp. 35–54.
7. Chowdhary, K.R. Natural Language Processing. In *Fundamentals of Artificial Intelligence*; Chowdhary, K.R., Ed.; Springer: New Delhi, India, 2020; pp. 603–649. [CrossRef]
8. Eisenstein, J. *Introduction to Natural Language Processing*; MIT Press: Cambridge, MA, USA, 2019.
9. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. Available online: http://arxiv.org/abs/1706.03762 (accessed on 20 October 2023).
11. Hariri, W. Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing. *arXiv* **2023**, arXiv:2304.02017. [CrossRef]
12. Ozer, D.J.; Reise, S.P. Personality Assessment. *Annu. Rev. Psychol.* **1994**, *45*, 357–388. [CrossRef]
13. Thurstone, L.L. The vectors of mind. *Psychol. Rev.* **1934**, *41*, 1–32. [CrossRef]
14. Goldberg, L.R. The structure of phenotypic personality traits. *Am. Psychol.* **1993**, *48*, 26–34. [CrossRef]
15. Roberts, B.W.; Yoon, H.J. Personality Psychology. *Annu. Rev. Psychol.* **2022**, *73*, 489–516. [CrossRef]
16. Asendorpf, J.B.; Neyer, F.J. *Psychologie der Persönlichkeit*; Springer: Berlin/Heidelberg, Germany, 2012. [CrossRef]
17. Rauthmann, J.F. *Persönlichkeitspsychologie: Paradigmen—Strömungen—Theorien*; Springer: Berlin/Heidelberg, Germany, 2017. [CrossRef]
18. Boyle, G.J. Myers-Briggs Type Indicator (MBTI): Some psychometric limitations. *Aust. Psychol.* **1995**, *30*, 71–74. [CrossRef]
19. Furnham, A. The big five versus the big four: The relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Personal. Individ. Differ.* **1996**, *21*, 303–307. [CrossRef]
20. Turing, A.M. Computing Machinery and Intelligence. *Mind* **1950**, *59*, 433–460. [CrossRef]
21. Laugwitz, B.; Schubert, U.; Ilmberger, W.; Tamm, N.; Held, T.; Schrepp, M. *Subjektive Benutzerzufriedenheit Quantitativ erfassen: Erfahrungen Mit dem User Experience Questionnaire UEQ*; Tagungsband UP09; Fraunhofer Verlag: Stuttgart, Germany, 2009.
22. Ahsan, M.M.T.; Rahaman, M.S.; Anjum, N. From ChatGPT-3 to GPT-4: A Significant Leap in AI-Driven NLP Tools. *SSRN Libr.* **2023**. [CrossRef]
23. Miotto, M.; Rossberg, N.; Kleinberg, B. Who is GPT-3? An Exploration of Personality, Values and Demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*; Association for Computational Linguistics: Stroudsburg, PN, USA, 2022; pp. 218–227. Available online: https://aclanthology.org/2022.nlpcss-1.24.pdf (accessed on 20 October 2023).
24. Li, X.; Li, Y.; Liu, L.; Bing, L.; Joty, S. Is GPT-3 a Psychopath? Evaluating Large Language Models from a Psychological Perspective. *arXiv* **2022**, arXiv:2212.10529. [CrossRef]
25. Rutinowski, J.; Franke, S.; Endendyk, J.; Dormuth, I.; Pauly, M. The Self-Perception and Political Biases of ChatGPT. *arXiv* **2023**, arXiv:2304.07333. [CrossRef]
26. Shumanov, M.; Johnson, L. Making conversations with chatbots more personalized. *Comput. Hum. Behav.* **2021**, *117*, 106627. [CrossRef]
27. Jiang, G.; Xu, M.; Zhu, S.-C.; Han, W.; Zhang, C.; Zhu, Y. MPI: Evaluating and Inducing Personality in Pre-trained Language Models. *arXiv* **2022**, arXiv:2206.07550. [CrossRef]
28. Törnberg, P. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv* **2023**, arXiv:2304.06588. [CrossRef]
29. Kosinski, M. Theory of Mind May Have Spontaneously Emerged in Large Language Models. *arXiv* **2023**, arXiv:2302.02083. [CrossRef]
30. Grimmer, J.; Stewart, B.M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Anal.* **2013**, *21*, 267–297. [CrossRef]

31. Wiens, J.; Saria, S.; Sendak, M.; Ghassemi, M.; Liu, V.; Doshi-Velez, F.; Jung, K.; Heller, K.; Kale, D.; Saeed, M.; et al. Do no harm: A roadmap for responsible machine learning for healthcare. *Nat. Med.* **2019**, *25*, 1337–1340. [CrossRef]
32. West, C.G. Advances in apparent conceptual physics reasoning in GPT-4. *arXiv* **2023**, arXiv:2303.17012. [CrossRef]
33. Rubynor. Free Open-Source BigFive Personality Traits Test. Bigfive. Available online: https://bigfive-test.com (accessed on 22 June 2023).
34. Truity. The TypeFinder Personality Test. Truity. Available online: https://www.truity.com/test/type-finder-personality-test-new (accessed on 30 May 2023).
35. Dong, W.; Zhunis, A.; Chin, H.; Han, J.; Cha, M. I Am Not Them: Fluid Identities and Persistent Out-Group Bias in Large Language Models. *arXiv* **2024**, arXiv:2402.10436.