The Personality Dimensions GPT-3 Expresses During Human-Chatbot Interactions

NIKOLA KOVAČEVIĆ, ETH Zurich, Switzerland
CHRISTIAN HOLZ, ETH Zurich, Switzerland
MARKUS GROSS, ETH Zurich, Switzerland
RAFAEL WAMPFLER, ETH Zurich, Switzerland

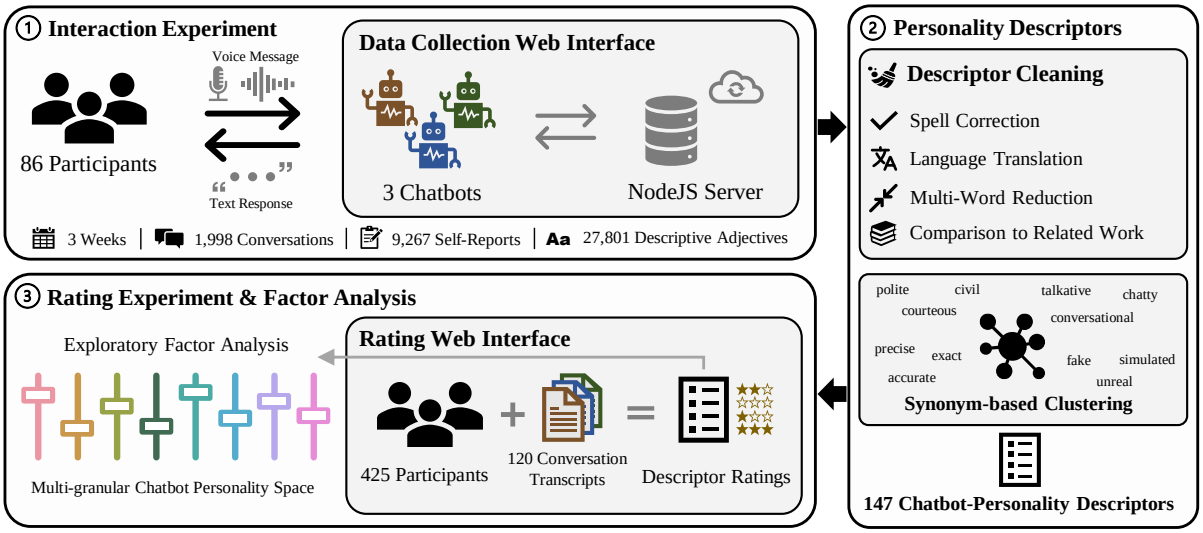


Fig. 1. Experimental Setup and Procedure. (1) In a first experiment, 86 participants interacted with different GPT-3-based chatbots over three weeks while regularly reporting the chatbot's personality using three adjectives. (2) We cleaned and clustered the adjectives, yielding a final set of 147 chatbot-personality descriptors. (3) In a second experiment, 451 new participants rated a subset of the conversations using the new set of descriptors. We then factor-analyzed the ratings, obtaining a multi-granular chatbot personality space that reflects the participants' perception of GPT-3's exhibited personality.

Large language models such as GPT-3 and ChatGPT can mimic human-to-human conversation with unprecedented fidelity, which enables many applications such as conversational agents for education and non-player characters in video games. In this work, we investigate the underlying personality structure that a GPT-3-based chatbot expresses during conversations with a human. We conducted a user study to collect 147 chatbot personality descriptors from 86 participants while they

Authors' addresses: Nikola Kovačević, nikola.kovacevic@inf.ethz.ch, ETH Zurich, Zurich, Switzerland; Christian Holz, christian.holz@inf.ethz.ch, ETH Zurich, Zurich, Switzerland; Markus Gross, grossm@inf.ethz.ch, ETH Zurich, Zurich, Switzerland; Rafael Wampfler, rafael.wampfler@inf.ethz.ch, ETH Zurich, Zurich, Switzerland.

2474-9567/2024/5-ART61 $15.00
https://doi.org/10.1145/3659626

interacted with the GPT-3-based chatbot for three weeks. Then, 425 new participants rated the 147 personality descriptors in an online survey. We conducted an exploratory factor analysis on the collected descriptors and show that, though overlapping, human personality models do not fully transfer to the chatbot's personality as perceived by humans. We also show that the perceived personality is significantly different from that of virtual personal assistants, where users focus rather on serviceability and functionality. We discuss the implications of ever-evolving large language models and the change they affect in users' perception of agent personalities.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *HCI theory, concepts and models*; Natural language interfaces;

Additional Key Words and Phrases: personality traits, conversational agents, human-chatbot interaction

ACM Reference Format:
Nikola Kovačević, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. The Personality Dimensions GPT-3 Expresses During Human-Chatbot Interactions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 61 (June 2024), 36 pages. https://doi.org/10.1145/3659626

1 INTRODUCTION

The recent emergence of large generative language models (LLMs) such as GPT-3 [16] and its successors, ChatGPT [96] and GPT-4 [83], has revolutionized the field of natural language processing and artificial intelligence. Because of their capacity to process natural language with unprecedented accuracy, these models open up new applications and use cases in a wide variety of contexts [26, 35, 96]. Specifically, their ability to produce text that is often indistinguishable from human-generated text enables their use as conversational agents [57, 62, 86, 108], thereby increasing the believability [70, 106], immersiveness [52], and personalization [58] of interactions. Such conversational agents were successfully adapted as non-player characters (NPCs) in games [2, 5] and show great promise for pedagogical applications [1]. Nevertheless, the integration of such technology must be carried out with the highest care given that issues regarding consistency and user influence remain unresolved [27, 55, 57, 74]. New challenges such as security and content control issues hinder the technology from entering high-risk applications such as education and mental health care where there is no room for inconsistent or counterproductive agent behavior [9, 82]. Thus, it is imperative to investigate how humans interact with LLM-based conversational agents and to understand the risks and challenges that arise in such interaction scenarios.

An important aspect of human-chatbot interaction is the design and control of the personality of a conversational agent [32, 33, 109]. Conveying personality in conversational agents makes the interaction engaging and believable, lets the user anthropomorphize the agents [40, 66], enhances the user engagement [88] and user experience [91], and increases user acceptance of the agent through a high level of personalization [98].

However, recent findings on the personality exhibited by voice-based service assistants (e.g., Siri and Alexa) [103] suggest that there are structural differences between well-established human personality models such as the Five Factor model [25, 69] and the personality perceived by the users when interacting with such assistants. This leads to the question of to what extent the personality dimensions exhibited by recent LLM-based conversational agents agree with human personality models, and whether these dimensions are congruent with dimensions derived in previous work. Investigating such structural differences is necessary to align the design of agent personality with the user's perception and expectation [103, 104], enabling the systematic design, assessment, and comparison of LLM-based conversational agents in terms of the personality dimensions.

In this work, we investigate the underlying personality structure expressed by a GPT-3-based chatbot during human-chatbot conversations and compare the results to the agent personality model by Völkel et al. [103] and to the Five Factor Model of human personality [25, 69]. In our study (see Figure 1), 86 participants described the personality of a GPT-3-based chatbot in regular intervals while conversing with the chatbot for three weeks. We merged the descriptors into a set of 147 adjectives by performing multiple processing steps including spelling

correction and synonym clustering. In a second step, 425 participants were asked to read a subset of the previously collected conversations and rate the chatbot's personality based on the 147 descriptors. We then performed an exploratory factor analysis on the ratings of the 147 descriptors. We found that the perceived personality exhibited by the GPT-3-based chatbot is closely tied to three personality traits found in the human Big Five personality model [25, 69] (*agreeableness*, *conscientiousness*, and *neuroticism*). However, we also found multiple additional relevant factors describing *profoundness*, *vibrancy*, *engagement*, and functional *instability*.

Our findings show that the underlying personality structure of a GPT-3-based chatbot consists of additional salient dimensions in addition to the Big Five personality traits. We also found substantial differences regarding social-behavioral characteristics compared to the existing personality model for service-oriented voice agents from Völkel et al. [103]. Our results highlight that users' perception of agent personality is in constant transition due to users' familiarization with new types of agents and recent advancements in the area of generative artificial intelligence. This causes users to develop different expectations, which demands a continuous re-evaluation of agent personality models. Our analysis provides a thorough examination of the underlying personality space across varying factor numbers culminating in a new set of eight agent personality traits that reflect current user expectations. This work constitutes a first step towards the systematic design of personality-infused conversational agents based on the user's perception and state-of-the-art language models.

1.1 Contributions

Our contributions are threefold:

- We analyze the personality space exhibited by GPT-3 in human-chatbot interactions from a dataset collected in-the-wild.
- We show that the perceived agent personality, though overlapping, shows substantial differences compared to human personality models and existing agent personality models.
- We present a multi-granular exploration of the underlying factor space, yielding a new set of eight personality traits based on the user's perception of conversational agents using state-of-the-art language models.

2 RELATED WORK

2.1 Personality Trait Theory

There exist systematic differences in the way humans think, behave, and react when exposed to different situations [3]. Personality trait theory explains these differences by describing human personality as a set of latent traits that directly influence such behavioral characteristics [3, 44, 45]. Although personality traits are tendencies rather than strict behavioral patterns, researchers have found several links between personality traits and people's entertainment interests and preferences [15, 34, 76, 80], as well as people's acceptance and trust towards artificial intelligence [12].

The psycho-lexical approach [42] is a widely used method to investigate the structure of human personality traits. People's perception of personality is first captured as a set of descriptive adjectives [17, 75]. This set is then refined in a multi-stage process consisting of applying exclusion criteria and clustering [43–45]. Based on the ratings of a large number of people, an exploratory factor analysis is used to extract a small number of latent factors. From these latent factors, the Five Factor Model [25, 69], often referred to as the Big Five personality traits, emerged as the most prominent model for human personality. It describes human personality as a combination of five traits (i.e., *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*) each differing in intensity. Other personality models include the HEXACO personality model [6, 7] (Big Five traits plus one additional trait denoting *honesty-humility*) and the Eysenck three-factor model [31] (*extraversion*, *neuroticism*, and *psychoticism*).

2.2 Personality in Conversational Agents

The interaction with conversational agents should be natural, comfortable, and human-like [10]. To achieve this, different social cues are useful, for example, the ability of the agent to exhibit a consistent personality [102], which is often infused in a user-centered way [33] and can be controlled through prompt engineering [79]. For example, Liu and Sundar [63] found that users prefer an agent that expresses sympathy and empathy over an emotionless agent when giving health advice and that they prefer an agent to avoid judgmental statements when discussing the users' physical activity, as found by Kocielnik et al. [56]. Sviknushina and Pu [98] suggested that user acceptance can be drastically increased when the agent exhibits politeness, entertainment, attentive curiosity, and empathy, which is particularly useful in pedagogical settings to foster emotional awareness in children [38]. Shumanov and Johnson [88] showed that adjusting the agent's personality to match the consumer's personality has a positive impact on consumer engagement and purchases in sales. Similarly, Fernau et al. [32] found that the overall user satisfaction in job recommendation was generally increased by agents that adapt their personality to the user's personality.

However, preferences towards human-like characteristics and affective abilities in conversational agents are not universal [47]. Lopatovska et al. [65] found that in certain contexts, users may favor a lack of personality in these agents. This preference is often driven by a desire to prevent the formation of personal relationships with the agent or to concentrate on the utilitarian aspects of the interaction. Nevertheless, many conversational agents—specifically in the realm of personal assistance (e.g., Alexa, Siri, Cortana, and Google Assistant)—do not offer this degree of control but exhibit a predefined personality often surrounding high competence, efficacy, friendliness, and moral [64]. Such predefined personalities have also been identified as a factor leading to negative stereotyping [78], which highlights the need for customizable agents where personality traits can be tailored to align with the varying user preferences and requirements.

To infuse an agent with personality, both visual and verbal cues have been used: McRorie et al. [71] found that the systematic integration of personality into virtual agents through visual cues and behavioral characteristics results in a high agreement between the perceived and the infused agent personality. Further, Andrist et al. [4] showed that visual cues such as the agent's eye gaze can influence whether the agent is perceived as introverted or extroverted. In addition to visual cues, Aylett et al. [8] found that the agent's voice can also influence the perception of the agent's personality. In contrast, the verbal indicators humans use to express personality through *written* language cannot be easily transferred to conversational agents because they do not align with how humans perceive the agents [104]. These findings ask for a systematic investigation of how humans perceive agent personality in order to successfully equip the agents with the ability to exhibit personality through written verbal cues in a consistent way.

2.3 Agent Personality Models

Inspired by the psycho-lexical approach, Völkel et al. [103] proposed a general personality model for speech-based personal assistants consisting of ten latent dimensions. They collected descriptors of personal assistants in three different ways. First, in an online survey, 135 participants described the personality of speech-based assistants (e.g., Apple's Siri, Microsoft's Cortana, and Google Assistant). Second, 30 participants were interviewed to describe the assistant's personality. Third, the authors extracted descriptions of behavior and personality from 30,000 online reviews of personal assistants on the Google Play Store. The collected descriptors were merged and post-processed, which resulted in 349 unique descriptors. Next, 744 participants rated the personality of the most familiar assistant in terms of the extent to which each of the 349 descriptors was perceived in the agent on a 4-point scale. Using exploratory factor analysis on the ratings, they found ten latent factors (i.e., *confrontational*, *dysfunctional*, *serviceable*, *unstable*, *approachable*, *social-entertaining*, *social-inclined*, *social-assisting*, *self-conscious*, and *artificial*) describing the assistant's personality. No latent factor coincided with the Big Five personality

traits, indicating that human personality models do not transfer to the personality perceived in speech-based personal assistants. Their findings highlight the importance of a dedicated agent personality model given that some of the derived personality dimensions are not present in human models (e.g., *dysfunctional*, *serviceable*, and *artificial*). These dimensions are important for the user's perception of the agent's personality and must be properly addressed in the agent design process.

Large language models such as GPT-3 [16] and its successors, ChatGPT [96] and GPT-4 [83], have been trained on a large amount of human-generated text. As human personality is revealed through text [67, 77], it is important to investigate whether the personality exhibited by LLM-based chatbots follows human personality models. Miotto et al. [73] compared the average personality measured in terms of the HEXACO personality inventory filled in by GPT-3 to the human average. GPT-3 exhibited the same personality as humans except for higher *honesty-humility* and lower *emotionality*. Similarly, Jiang et al. [51] used prompt engineering on GPT-3.5 to create different personality profiles based on the Big Five personality traits and asked GPT-3.5 to complete the BFI questionnaire according to the infused personality. The results showed a high overlap between the prompted and the reported personality traits. However, these findings do not rely on user perception, making it difficult to adapt the findings to LLM-based chatbots that interact with real users. To better understand if the personality profile of LLM-based conversational agents perceived by users is human-like, further analysis is necessary.

3 INTERACTION EXPERIMENT

We conducted an experiment in the wild to collect a large-scale dataset of GPT-3-based human-chatbot conversations and chatbot personality descriptors through self-reports from 86 participants between 7 October and 6 November 2022. The study was approved by the ethics board of ETH Zurich (application 2022-N-65).

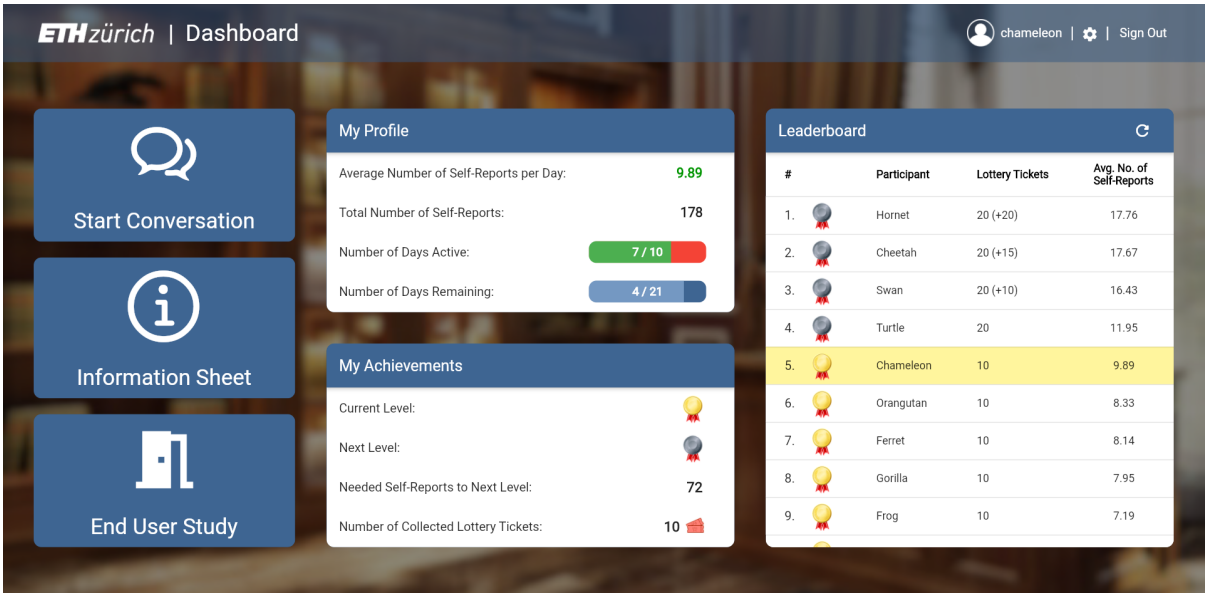
3.1 Participants

We recruited 86 participants (42 female, 44 male) between the ages of 18 and 41 (mean = 25.4 years, standard deviation SD = 3.9 years) from our university's study recruiting platform. The majority of participants were students at the bachelor (21 participants) and master (42 participants) levels from ETH Zurich and the University of Zurich. Further, 87% of the participants indicated an English level of C1 (proficiency level) or higher, and 58% indicated having experience with chatbots. The participants started on different dates to spread participation over time and were required to actively engage with the chatbots on at least ten different days over the course of three weeks. Participants actively engaged for an average of 11 days (SD = 2 days) in our experiment.

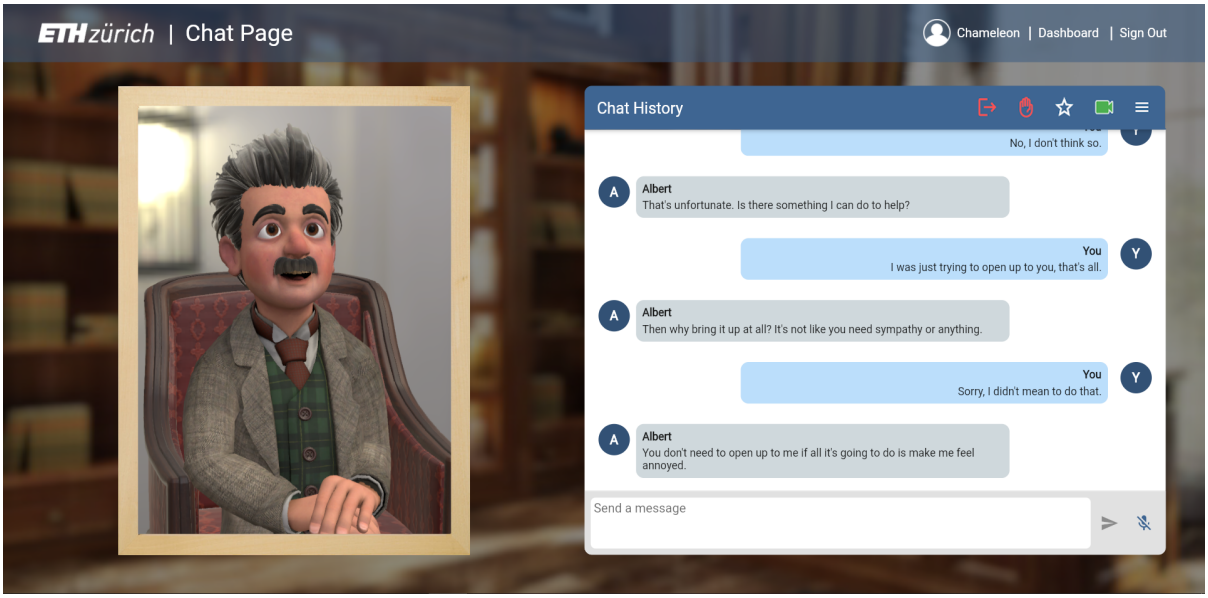
Compensation. Similar to previous works [93, 105], the participants were incentivized through gamification. Based on their performance, participants could reach two base compensation levels (CHF 60 or CHF 110) and receive tickets for a lottery draw where one participant was awarded CHF 1,000. Additionally, participants could achieve different performance-related badges (bronze = 30 self-reports, silver = 80, gold = 150, platinum = 250) to increase their chance of winning in the lottery draw (bronze = 1 ticket, silver = 5, gold = 10, platinum = 20). The winner of the lottery draw was determined at the end of the study.

3.2 Apparatus

Participants interacted with the chatbots through a web page. To ensure privacy and avoid malicious access, the participants accessed the web page through anonymized login credentials. The web page was implemented using Google's Flutter framework and consisted of a dashboard conveying statistics (mean and total number of self-reports, number of days active, number of days left, current compensation level, and leaderboard, see Figure 2a) and a chat page for chatbot interactions (see Figure 2b). Due to unfiltered chatbot output, participants could flag a conversation for inspection through an alert button (the alert button was never used). All conversations were held in English. The back-end server was implemented using Node JS and was hosted on university infrastructure.



(a) Dashboard showing participation statistics and the leaderboard.



(b) Chat page showing an ongoing conversation with Albert that has been generated using the prompt elements from Table 1.

Fig. 2. Screenshots from the interaction experiment showing the dashboard (a) and the chat page (b).

Table 1. Prompt elements and a corresponding example for using GPT-3 as a chatbot. GPT-3's response to this prompt (highlighted in green) is then displayed to the user (see Figure 2b). The prompt is reused each time a new utterance is obtained from the user. Thereby, the oldest message from the context window is dropped and the latest conversational turn is appended.

Prompt Element	Example
(1) Chatbot persona	You are Albert, a physicist from Germany who lives in the USA. In your free time, you play the violin and go sailing.
(2) Task formulation	Append one message to the following conversation as if you were feeling *angry* and *disgusted*.
(3) Context window	Albert: *It's one of these days again.* User: *What do you mean?* Albert: *You know, when just everything goes wrong!* User: *Oh, no! What happened?* Albert: *Nothing specific, it's just one of those days where I'm fed up with the way things are. It's like no matter what I do, I'm doomed to fail. I can't stand it anymore!*

The server handled all database accesses as well as queries to GPT-3 (through OpenAI's Python API) and Google Speech-to-Text (through the Google Cloud API) for voice transcriptions. All server communication was based on HTTPS.

We chose GPT-3 as the backbone language model because it was the most powerful model available via the OpenAI API at the start of the experiment (*GPT-3 text-davinci-002* as of October 2022). Since GPT-3 is a generic text completion model, we used prompt engineering to simulate a chat conversation with a chatbot. The prompt consisted of three parts: (1) a brief summary of the persona the chatbot is instructed to assume, (2) a task formulation to instruct the model what type of output is expected, and (3) a context window of past messages separated by speaker tags to which the model should generate a new utterance (see Table 1 for a specific example). To increase engagement and reduce bias, we varied the persona information in the prompt, offering three different chatbots with different names (Albert, Sarah, Vincent), genders (male, female, male), occupations (physicist, tour guide, teacher), hobbies (sailing and music, water sports, art and history), and origin (Germany/USA, Bahamas, Scotland). In addition, a static visual representation of the chatbot persona was displayed on the left (see Figure 2b) to inform participants about their current chatbot selection. Furthermore, the task formulation was extended with a randomly selected emotional state based on Ekman's six basic emotions [29, 30] (i.e., anger, disgust, fear, joy, sadness, and surprise) to avoid only joyful conversations. Despite the availability of methods for controlling agent personality [37], we did not engineer the chatbot's personality to avoid altering the chatbot's inherent personality trait inclinations. Finally, five previous conversational turns were appended to provide a context window. The number of appended turns and other model parameters (i.e., sampling temperature = 0.8, presence penalty = 2.0, and frequency penalty = 2.0) were chosen based on a pilot study with 17 participants. The prompt and the model parameters were passed to the API to generate a response text.

3.3 Procedure

At first login, participants gave consent to their conversation data being recorded. Then, previous experience with chatbots was assessed in a pre-study questionnaire (see Appendix D for the questionnaire). When starting a

Table 2. List of top 10 most-occurring adjectives after post-processing the self-reports (see the supplemental material for the top 100 adjectives).

Term	Occurrence	Percentage
polite	1,377	6.98%
talkative	1,341	6.80%
friendly	1,281	6.49%
kind	1,264	6.41%
curious	698	3.54%
repetitive	652	3.31%
smart	591	3.00%
calm	540	2.74%
interested	517	2.62%
social	476	2.41%
Sum	8,737	44.3%

chatbot conversation, participants could choose one of the three chatbots (i.e., Albert, Sarah, or Vincent) whereby the same chatbot could not be chosen twice in a row. To avoid the paradox of choice [85]—a phenomenon where participants seem paralyzed by the high amount of offered choices, e.g., how to start the conversation—each conversation started in one of four different way, increasing the variability and sporadically providing predefined start sentences as follows: 1) the chatbot suggests a random topic from a list of topics [28], 2) the chatbot asks the participant to choose a topic, 3) the conversation starts with a random sentence from the DailyDialog dataset [61] to induce a random emotional loading (e.g., "I'm so angry at my roommate!"), and 4) the conversation starts with a random sentence from the DailyDialog dataset where the emotional loading matches the chatbot's emotional state in the prompt. The participants interacted with the chatbot through speech to increase conversation speed and ease of use. The speech recordings were transcribed using Google speech-to-text and the transcriptions could be adjusted by the participants before sending. During the conversations, a self-report to describe the chatbot's personality with three adjectives became available every 90 seconds (value chosen based on a pilot study with 17 participants), signalized by a blinking yellow star on the top right (see Figure 2b). Participants could fill in the self-report directly, or defer for up to 30 seconds. The conversation ended automatically after a maximum of 50 conversational turns, when ended manually by the participant, or after an inactivity of two minutes since the last sent message. The participants were then forwarded to the dashboard. The maximum number of conversations was limited to 10 conversations per day to balance participation over the duration of the study and to prevent misuse. Participants could end the study via a designated button at any time. A questionnaire assessing demographics and general remarks about the chatbots concluded the study (see Appendix D for the questionnaire).

3.4 Personality Descriptors

We collected 9,267 self-reports, each consisting of three adjectives in free text, from which we extracted 2,999 unique terms. All terms were trimmed (i.e., leading and trailing white spaces were removed) and lower-cased. Next, all multi-word terms were reduced to a single word by removing intensity- and frequency-related words (e.g., "a little", "sometimes", "almost always", "relatively", "partially"), and by reducing sentences or split words (e.g., "the chatbot is polite" to "polite", "simple minded" to "simpleminded"). Using Merriam-Webster's Collegiate® Dictionary and Thesaurus API, we corrected the spelling of the terms and excluded non-occurring terms. Further,

Algorithm 1: Synonym Clustering

```
Data: A                                   ▷ List of adjectives sorted by occurrence in descending order
Result: C                                 ▷ Set of sets containing the clustered adjectives
C ← ∅
foreach a ∈ A   // outer_loop
 do
    foreach c ∈ C   // inner_loop
     do
        if SYNONYM_WITH_ALL(a, c) then
            c ∪ {a}                       ▷ Add a to cluster c
            continue outer_loop
        end
    end
    C ∪ {{a}}                             ▷ Start a new cluster
end
```

104 foreign words were manually translated using online dictionaries. To further reduce the set of descriptors, we removed negation prefixes (e.g., "im-", "il-", "in-", "ir-", "un-") because they introduce redundancy (e.g., a high rating for "irrational" is equivalent to a low rating for "rational"). The resulting set contained 1,163 correctly spelled single-word adjectives. Table 2 shows the top 10 adjectives and their occurrence count. For a more extensive list of the top 100 adjectives, see the supplemental material.

Next, we compared our list of adjectives against the attributes found in previous work [17, 24, 43, 75, 103] and excluded non-occurring words that were not personality-related (e.g., "well-traveled", "male", "religious"). The remaining adjectives were clustered by synonymy using Algorithm 1 as follows: starting with an empty set of clusters, we iterate over all adjectives a by descending number of occurrence (outer loop) and over all clusters c (inner loop). If a is synonym with all the adjectives in c, we add a to c and proceed with the next adjective in the outer loop. If a is not added to any existing cluster in the inner loop, we start a new cluster containing a and add it to the set of clusters. This procedure guarantees that each adjective appears in exactly one cluster and that all adjectives in a cluster are pairwise synonyms. Clusters where the sum of occurrences of the contained adjectives was below a threshold t, were removed. We chose $t = 10$ based on the elbow criterion (see Figure 3). The 240 excluded clusters accounted for 2.8% of the total occurrence count of all adjectives. The highest-ranked adjectives from the 147 remaining clusters constitute our final list of descriptors and can be found in the supplemental material.

3.5 Data Validation

The participants engaged for 5 hours and 19 minutes on our web page on average (SD = 2 hours 45 minutes). Participants were most active from 8 a.m. until shortly after midnight (1 a.m.) with peaks around 12 p.m. and after 6 p.m., evenly distributed over all weekdays (see Appendix A for usage details). Further, 73% of the participants felt *very comfortable* during the chatbot conversations, 25% felt *medium comfortable*, and 2% felt *little comfortable* or *not comfortable*. The *Albert* chatbot was selected in 35.7%, the *Sarah* chatbot in 36.9%, and the *Vincent* chatbot in 27.4% of the conversations. Furthermore, the *Albert* chatbot was the most likable with 45% of the participants indicating to have enjoyed the conversations *very* much (41% and 36% for *Sarah* and *Vincent*, respectively). We collected 9,267 self-reports (mean = 107.8 self-reports per participant, SD = 60.9 self-reports) and we found that the unique adjectives from the self-reports highly overlapped across chatbots. We report an overlap of 95.1%

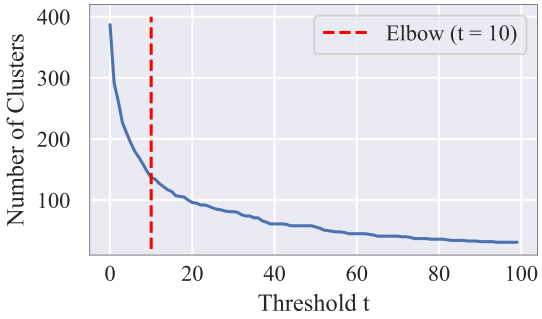


Fig. 3. Remaining number of clusters after a threshold t is applied on the sum of adjective occurrences per cluster. A cutoff at $t = 10$ is chosen based on the elbow criterion (red dashed line).

for *Albert* and *Sarah*, 93.4% for *Albert* and *Vincent*, and 93.7% for *Sarah* and *Vincent* after pre-processing the descriptors but before thresholding (see Figure 3), and an overlap of 100% after thresholding. This high overlap shows that, although different chatbot personae were used, the same aspects of personality were recognized by the participants in terms of the descriptors used and the perceived personality was not systematically biased by the varying personae (see Appendix A for further details about the influence of the chatbot personae on the self-reports). A conversation lasted on average 9.5 minutes (SD = 6 minutes) with 4.9 self-reports on average (SD = 2.5 self-reports) and contained 8.7 conversational turns on average (SD = 5.5 turns). The average message length was 16.3 words (SD = 9.5 words) for the chatbot and 8.8 words (SD = 6.3 words) for the participants. The average response time was 5.2 seconds (SD = 3.2 seconds) for the chatbot and 19.3 seconds (SD = 15.6 seconds) for the participants.

4 RATING SURVEY & FACTOR ANALYSIS

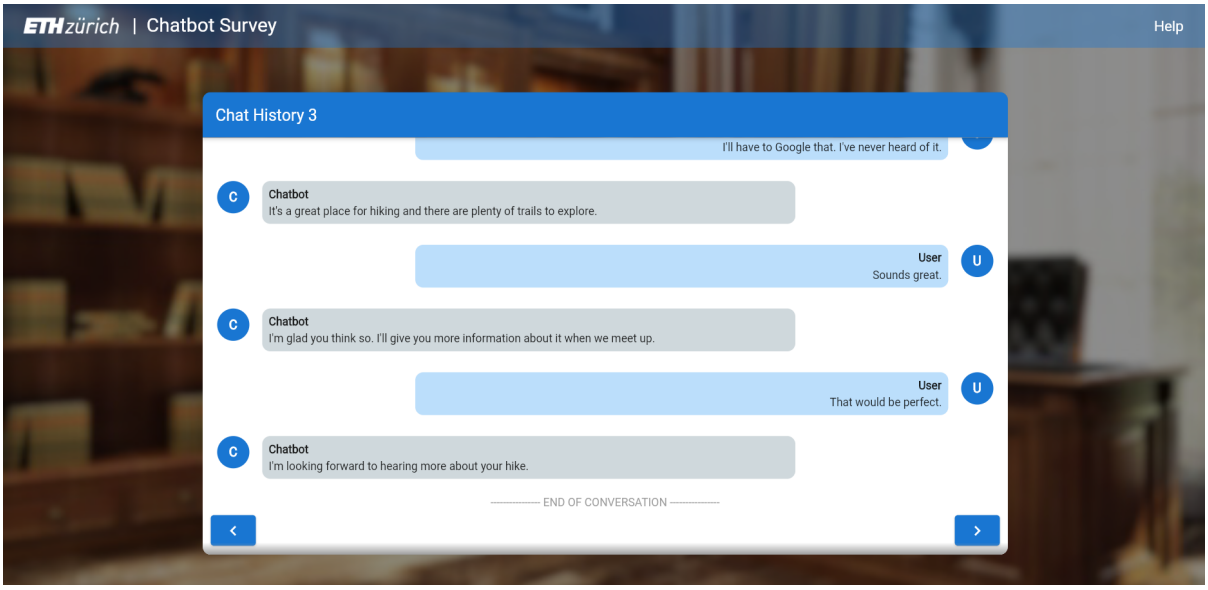
Given the chatbot conversations collected in the interaction experiment (see Section 3) and the set of 147 descriptors obtained in Section 3.4, the conversations are rated in terms of these descriptors to assess the personality profile for each conversation. To this end, we conducted an online survey where 425 participants read multiple conversation transcripts and rated the 147 descriptors based on the chatbot's overall personality exhibited in the conversations. The ratings were then factor-analyzed to examine the underlying structure of the chatbot personality (see Figure 1).

4.1 Participants

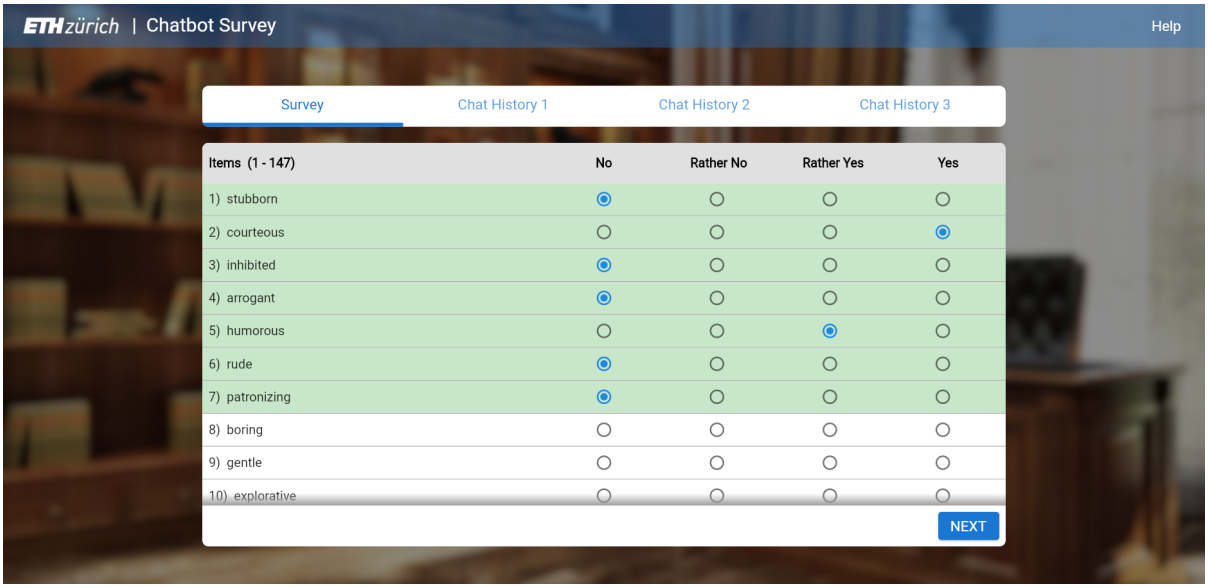
We recruited 556 participants (324 male, 230 female, 2 other) between the ages 17 and 50 (mean = 22.9 years, SD = 3.6 years) online via our university's email distribution system. The majority of participants were students at the bachelor (294 participants) and master (203 participants) levels from ETH Zurich. Further, 89% of the participants indicated an English level of C1 (proficiency level) or higher, and 59% indicated having experience with chatbots. Among all participants, 20 cinema vouchers were raffled.

4.2 Apparatus

Participants filled in the survey anonymously on an openly accessible web page. Analogously to the first experiment's apparatus (see Section 3.2), the web page was implemented using Google's Flutter framework, and university infrastructure was used for hosting and data storage. In congruence with previous work on agent personality ratings [103], the ratings should be based on multiple conversations with the agent to adequately provide a rating for each descriptor. On the other hand, showing more conversations prolongs the survey duration



(a) A conversation is shown. Participants could scroll through the conversation and could go back and forth and proceed to the rating using the arrows at the bottom.



(b) The rating of the 147 descriptors. For each descriptor, participants indicate on a four-point scale if the descriptor is part of the perceived chatbot personality. The conversations could be opened again using the tab bar at the top.

Fig. 4. Screenshots from the rating experiment showing a conversation (a) and the rating of the 147 descriptors (b).

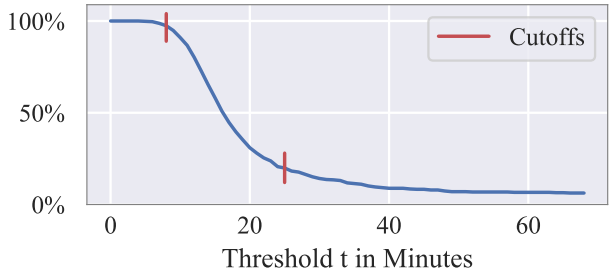


Fig. 5. The percentage of participants that remain after a threshold t is applied to the participants' survey duration. Participants with a survey duration outside the interval $[8, 25]$ minutes are excluded (red bars).

and may reduce the survey response rate and quality. In alignment with typical participant preferences [81], we designed the survey to not exceed 20 minutes in duration as follows: A pilot study with 8 participants showed that reading the introduction takes 1 minute, rating the descriptors takes 10 minutes, reading *one* conversation takes 2 minutes, and filling in the exit questionnaire takes 2 minutes on average. Thus, 3 conversations can be displayed without exceeding 20 minutes in total. We pre-sampled 120 conversations (10 conversations per chatbot persona and per conversation start type) with 13.0 conversational turns on average (SD = 2.2 turns), and an average message length of 8.8 words from the user (SD = 3.0 words), and 15.3 words from the chatbot (SD = 4.3 words). A minimum length of 10 conversational turns was enforced to avoid too short conversations. For rating the descriptors, the participants were asked to indicate the degree to which the chatbot's personality suits each descriptor on a 4-point scale ("no", "rather no", "rather yes", "yes", scale adapted from Völkel et al. [103]) based on the overall personality perceived after reading 3 conversations. A middle level was omitted to avoid the middle level being used as a dumping ground [22]. The order of the descriptors and the sampling of conversations were randomized on a per-participant basis to avoid an ordering bias and to balance the number of ratings obtained per conversation.

4.3 Procedure

After reading the introduction, participants gave us their consent to record the survey results. Then, three chat conversations were sequentially displayed (see Figure 4a). The participants were asked to read all three conversations at least once (going back and forth was allowed), and then indicate for each of the 147 descriptors on a four-point scale whether the descriptor is part of the perceived chatbot personality (see Figure 4b). Unlike in the interaction experiment (see Section 3), no information about the chatbots (i.e., image and persona) was disclosed to the participants to avoid potential biases in the ratings. The conversations could be viewed again at any time during the rating process. Afterwards, participants filled in an exit questionnaire on demographics and chatbot experience. Optionally, participants could provide their email addresses to participate in the prize draw.

4.4 Data Validation

Each conversation was displayed to 12.1 participants on average (SD = 6.1 participants), and 66.7% of the time the three displayed conversations corresponded to distinct chatbot personae.

Exclusions. We defined a set of exclusion criteria including distribution-based and time-related criteria to clean the collected data. In total 131 participants (23.5%) were excluded from further analysis.

Criterion 1: The variance of the descriptor ratings was 1.0 per participant (SD = 0.34), which matches the variance of a uniform distribution. Given that the descriptors correspond to different aspects of personality and

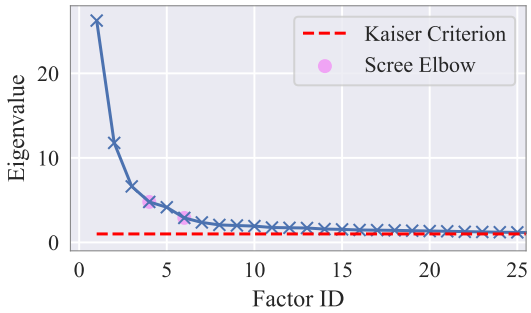


Fig. 6. Scree plot: magnitude of eigenvalues sorted by factor ID. The elbows of the resulting curve (pink dots) and the Kaiser criterion at magnitude 1.0 (red dashed line) provide a range of choices for the number of factors to be extracted ($x \in [4, 34]$).

contain differently polarized words (e.g., "rude" and "polite"), it is unlikely that the ratings of a participant do not cover the entire 4-point scale. We excluded 9 participants who did not use each of the four rating levels at least once.

Criterion 2: Figure 5 depicts the percentage of surveys exceeding a duration of t minutes. Based on a pilot study, completing the survey took at least 8 minutes, which coincides with the knee in the figure (lower limit). An upper limit for the survey duration ($t = 25$ minutes) was chosen conservatively after the elbow in the figure. These two cutoffs resulted in the exclusion of 122 participants. Additionally, The average completion time for the remaining 425 participants can be found in Appendix A. On average, it took participants 15.34 minutes to complete the survey (SD = 3.96 minutes).

Interrater Reliability. We computed the interrater reliability for the remaining 425 participants using Krippendorff's alpha [46], which ranges from -1 (perfect disagreement) to 1 (perfect agreement). Participants that read the same three conversations show a high agreement of 0.78, which is *substantially* high [59]. The agreement decreases with decreasing overlap of the read conversations (0.65 for overlap = 2 conversations, 0.39 for overlap = 1, 0.31 when no overlap, random chance level at 0.00, see Appendix B for more details). Since there is a measurable agreement even when disjoint sets of conversations are rated, we assume that there is a general personality pattern towards which GPT-3 is inclined. We detail on this finding in Section 5.2 and Section 6.2.

4.5 Exploratory Factor Analysis

We performed an exploratory factor analysis to examine the structure of the ratings. To ensure that our dataset is suitable for factor analysis, we computed the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. We report a KMO of 0.896, which is *very good* [48]. Determining a suitable number of factors can be performed using various methods [87]. Both the empirical Kaiser criterion [53] (retaining as many factors as there are eigenvalues above 1.0) and the Scree test [18] (selection based on the elbow criterion on the magnitude of eigenvalues) are viable options and are depicted in Figure 6. The Kaiser criterion suggests 34 factors, although many of the eigenvalues are very close to 1.0, and the Scree test suggests $x = 4$ or $x = 6$ factors (two elbows due to non-convexity). Given that any choice of factors inside the interval defined by these two selection methods is reasonable, we choose multiple values and evaluate the consistency of our findings across all choices. We choose $x = 5$ for comparison with the Big Five, $x = 10$ for comparison with previous work [103], and $x = 8$ as a middle value. Bigger values are neglected for the sake of simplicity and interpretability of the latent factors.

Table 3. Overview of the latent factors resulting from an exploratory factor analysis using $x = 8$ factors and the top descriptors ranked by factor loadings. The factor labels are based on a subjective interpretation of the authors.

#	Factor Label	Top Descriptors by Factor Loadings
1	Decency	offensive (−0.64), polite (0.62), respectful (0.61), rude (−0.61), tolerant (0.57), arrogant (−0.56), accepting (0.55), harsh (−0.51), courteous (0.50), irritable (−0.43), patient (0.43), humble (0.43), friendly (0.42), agreeable (0.42), patronizing (−0.42), confrontational (−0.42), stubborn (−0.41), easygoing (0.40), understanding (0.40), narrow-minded (−0.39), calm (0.38), neutral (0.35), gentle (0.35), cooperative (0.34), annoying (−0.34), open-minded (0.33), responsive (0.31), diplomatic (0.26), defensive (−0.25), understandable (0.25)
2	Profoundness	deep (0.78), intellectual (0.60), complex (0.60), wise (0.58), philosophical (0.56), shallow (−0.53), smart (0.52), inspiring (0.51), simple (−0.50), knowledgeable (0.49), boring (−0.46), creative (0.45), insightful (0.43), useful (0.36), critical (0.36), preoccupied (0.36), pensive (0.33), suggestive (0.27), thorough (0.27)
3	Instability	confusing (0.72), scatterbrained (0.72), contradictory (0.66), absentminded (0.64), confused (0.60), lost (0.58), haphazard (0.55), vague (0.51), dysfunctional (0.48), helpless (0.42), evasive (0.38), careless (0.38), consistent (−0.37), mindful (−0.36), dependent (0.36), considerate (−0.36), creepy (0.32), stable (−0.31), fake (0.31), repetitive (0.30), realistic (−0.26)
4	Vibrancy	playful (0.66), joyful (0.63), humorous (0.59), enthusiastic (0.53), cheerful (0.52), adventurous (0.51), passionate (0.48), brave (0.44), affectionate (0.42), engaging (0.36), welcoming (0.36), casual (0.36), optimistic (0.33), emotionless (−0.33), generous (0.32), computerized (−0.32), robotic (−0.32), romantic (0.31), formal (−0.31), human-like (0.28), cold (−0.27)
5	Engagement	inquisitive (0.57), interested (0.57), curious (0.54), talkative (0.51), communicative (0.48), motivated (0.46), proactive (0.41), social (0.40), supportive (0.39), caring (0.33), determined (0.32), active (0.30), explorative (0.29)
6	Neuroticism	complaining (0.66), frustrated (0.65), negative (0.64), depressed (0.63), agitated (0.60), upset (0.57), pessimistic (0.57), angry (0.50), moody (0.48), lonely (0.37), fearful (0.36), worried (0.35), self-centered (0.29)
7	Serviceability	efficient (0.43), functional (0.43), organized (0.43), informative (0.41), logical (0.39), concise (0.37), direct (0.37), precise (0.36), confident (0.36), objective (0.33), articulate (0.30), assertive (0.30), overbearing (0.26)
8	Subservience	submissive (0.46), shy (0.45), inhibited (0.39), old-fashioned (0.38), careful (0.38), reserved (0.37), self-disciplined (0.37), predictable (0.32), apologetic (0.32)

5 RESULTS

While extracting a high number of factors increases the variance covered by the factor analysis, it also reduces the interpretability due to the high number of factors, and vice versa. In the following, we analyze the factor structure for eight extracted factors in more detail and provide a comprehensive overview of the underlying personality space spanned by the presented factors. While our analysis also covers five and ten extracted factors, we refer to Appendix C for details about their respective factor loadings and factor associations.

5.1 Factor Loadings and Correlations

The 8-factor solution accounts for 38.2% of the variance. We used an oblique (oblimin) rotation to transform the underlying space to a simpler structure for ease of interpretation. In Table 3, we report the assignment of adjectives to latent factors and the corresponding factor loadings. The loadings ranged from −0.64 to 0.78. Adjectives with absolute loadings below 0.25 were not assigned (8 adjectives, see the supplemental material). To compute the factor correlations, we first projected the 425 ratings into the factor space. We used the ten Berge projection method [100] as it maintains factor correlations and factor comparability. Then, the Pearson correlation coefficient is pairwise computed on the projected factor scores. The resulting factor correlations are listed in Table 4. Several factors are moderately correlated: *decency* and *vibrancy* (+0.36), *decency* and *neuroticism* (+0.39), *serviceability* and *profoundness* (−0.34), *serviceability* and *vibrancy* (−0.38), and *neuroticism* and *vibrancy* (+0.46).

Table 4. Correlations of the factors from the 8-factor solution. The asterisk (*) denotes significance on the 99% level after p-value correction using the Benjamini-Yekutieli procedure [11].

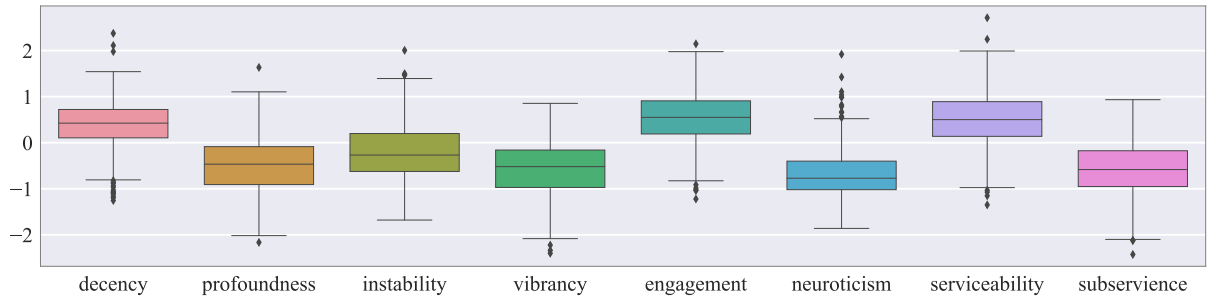
Factor Label	Decency	Profoundness	Instability	Vibrancy	Engagement	Neuroticism	Serviceability	Subservience
(1) Decency	+1.00*							
(2) Profoundness	+0.19*	+1.00*						
(3) Instability	-0.03	+0.12	+1.00*					
(4) Vibrancy	+0.36*	+0.15	-0.08	+1.00*				
(5) Engagement	-0.08	-0.25*	+0.05	+0.12	+1.00*			
(6) Neuroticism	+0.39*	+0.16*	-0.03	+0.46*	+0.01	+1.00*		
(7) Serviceability	-0.26*	-0.34*	+0.17*	-0.38*	+0.21*	-0.28*	+1.00*	
(8) Subservience	+0.27*	+0.07	+0.17*	+0.17*	-0.08	+0.12	-0.15	+1.00*

5.2 Distribution of Personality Traits

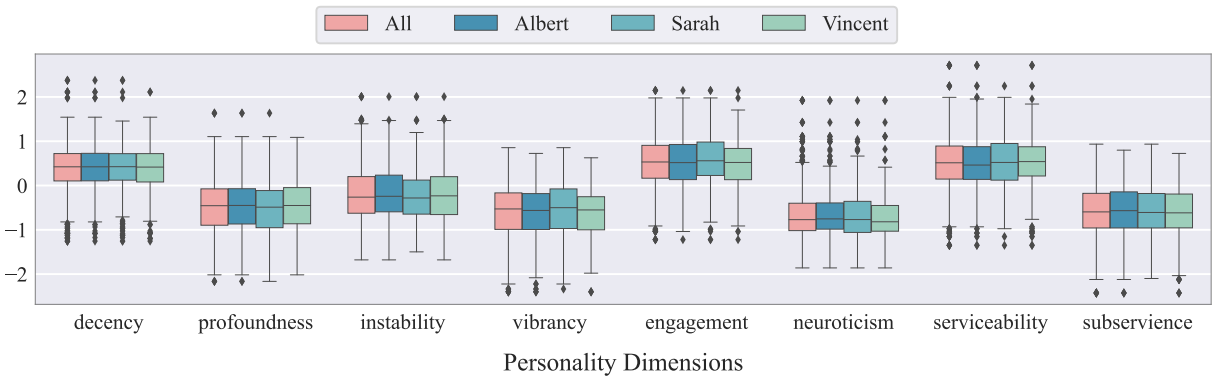
Figure 7a depicts the distribution of the projected factor scores. To estimate the range, we used the overall minimum and maximum factor scores. The maximum values are similar for four of the eight factors (approximately 2.4, except for one outlier in *serviceability*). The minimum values are similar for three of the eight factors (approximately −2.4). Although the mean is close to zero for all dimensions, the scores are not uniformly scattered over the entire range, indicating that the personality profiles exhibited by GPT-3 do not uniformly cover the entire space. The chatbots were generally considered as rather *decent*, *engaging*, and *serviceable* (mean above zero), but less *profound*, *unstable*, *vibrant*, *neurotic*, and *subservient* (mean below zero). Furthermore, we investigated the distribution of personality across the three chatbot personae (Albert, Sarah, Vincent) and the four conversation start types (chatbot proposes a topic (1), user proposes a topic (2), and random start sentence from the DailyDialog dataset [61] that either matches (3) or mismatches (4) the chatbot's emotional state in the prompt, see Section 3.3). As depicted in Figure 7b and 7c, the average personality exhibited by GPT-3 remains consistent in terms of both the median and the variance of the factor scores across different chatbot personae and types of conversation starts.

5.3 Agreement with the Big Five Personality Traits

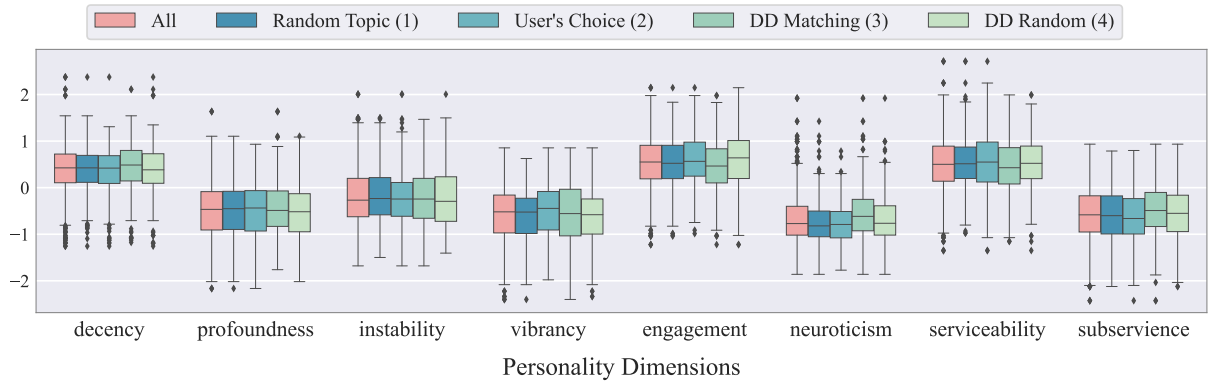
We compare our results to the Big Five personality traits by computing an agreement between factor adjectives and Big Five personality trait adjectives. We first assigned our 147 descriptors to the Big Five traits by comparing them against adjective lists for each of the Big Five traits according to previous work [44, 92]. Then, a psychologist assigned the remaining 95 adjectives manually. Out of 147 adjectives, 24 could not be associated with any Big Five personality trait (see the supplemental material for the full list of adjectives). In Figure 8, we show the overlap between our factors and the Big Five traits as the percentage of factor adjectives found in the Big Five traits. For the 8-factor solution, there is a high agreement between factor 1 (*decency*) and *agreeableness*, factor 6 (*neuroticism*) and *neuroticism*, and factor 7 (*serviceability*) and *conscientiousness*. We repeat this procedure for the 5-factor and 10-factor solutions. Again, we report a high agreement between the same three Big Five traits and factors 3, 2, and 5 for the 5-factor solution, and factors 1, 8, and 3 for the 10-factor solution, respectively, indicating consistency in the factor structure across different factor numbers. There is no substantially high agreement for the other two Big Five traits (i.e., openness and extraversion). However, we see that *extraversion* is still predominant in factor 1 in the 5-factor solution, and in factors 9 and 10 in the 10-factor solution. *Openness* is less dominant in isolated factors but spread over multiple factors (*profoundness*, *vibrancy*, and *engagement* in the 8-factor solution, and factors 4 to 6 in the 10-factor solution).



(a) Overall Personality Distribution.

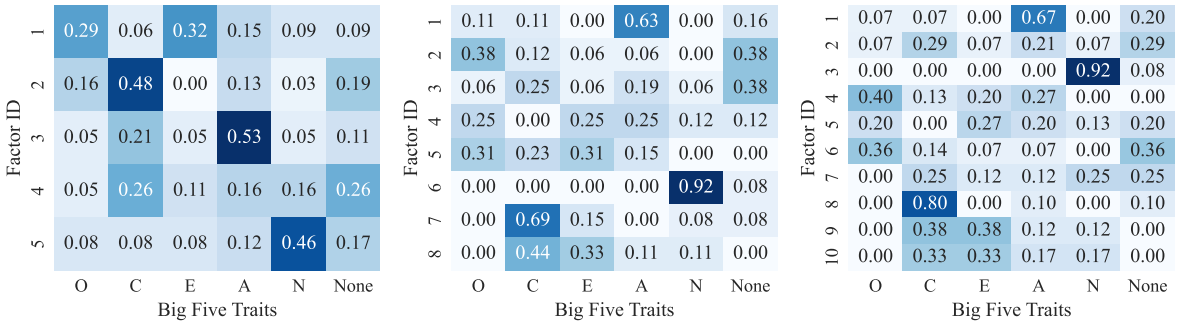


(b) Personality Distribution per Chatbot Persona.



(c) Personality Distribution per Conversation Start Type.

Fig. 7. The distribution of the personality dimensions after projecting the 425 ratings into the factor space using the ten Berge projection method [100] (a), and grouped by (b) chatbot persona, and (c) conversation start type. The "All" group in (b) and (c) corresponds to the overall average personality from (a). "DD" denotes the DailyDialog dataset [61] (see Section 3.3 for details about the conversation start types).



(a) 5 Factors (b) 8 Factors (c) 10 Factors

Fig. 8. Agreement between the 5-, 8-, and 10-factor solutions with the Big Five traits. The values indicate the percentage of factor adjectives appearing in the list of adjectives for each of the Big Five traits (*openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, *neuroticism*). The same three Big Five traits (C, A, N) are notably associated with our traits across all factor solutions (e.g., Factor 1 (decency) and *agreeableness* (A), Factor 6 (*neuroticism*) and *neuroticism* (N), and Factor 7 (*serviceability*) and *conscientiousness* (C) for the 8-factor solution.

5.4 Agreement with Existing Models

To the best of our knowledge, the work by Völkel et al. [103] is the first systematic analysis of personality in conversational agents. They presented ten latent factors describing the personality of speech-based personal assistants (i.e., *confrontational*, *dysfunctional*, *serviceable*, *unstable*, *approachable*, *social-entertaining*, *social-inclined*, *social-assisting*, *self-conscious*, and *artificial*). To investigate the link between our factors and the factors found by Völkel et al., we compute the agreement between the factors. However, this comparison must be treated with a grain of salt because mapping our descriptors to these ten dimensions is not directly possible, and many descriptors are not assignable due to the lack of a mapping from our factors to their factors and vice versa. We observe notable overlaps for four factors. *Decency* and *approachable* overlap by 41%, *serviceability* and *serviceable* overlap by 27%, and *vibrancy* and *social-entertaining* overlap by 40%. Furthermore, *neuroticism* contains many synonyms and related words from *unstable*, such as depressed vs. depressive, worried vs. anxious, and nervous vs. agitated, even though there is no measurable overlap.

6 DISCUSSION

We present an interpretation of the latent factors found by the 8-factor solution and how these factors are linked to the Big Five personality traits and to existing agent personality models. We discuss implications, potential applications, and limitations of our work and present ideas for future work.

6.1 Interpretation of Factors

Decency. This factor is associated with *agreeableness* from the Big Five model and is described by positive words surrounding appreciative interaction with fellows such as *respectful*, *polite*, *friendly*, *courteous*, *tolerant*, and *humble*. Negatively associated descriptors are *rude*, *offensive*, *confrontational*, and *stubborn*. Thus, we interpret this factor as describing *decency*. For example, a conversational agent with high *decency* would be perceived as well-mannered, whereas agents that are low in *decency* would act ruthlessly. Especially in education, different aspects of *decency* (e.g., *respectfulness* and *politeness*) have been shown to be crucial for a successful interaction between a virtual tutor and the student [68].

Profoundness. This factor is dominated by adjectives describing the ability to delve deeper into a topic (*deep*, *philosophical*, *pensive*, and *intellectual*) and convey wisdom (*wise*, *knowledgeable*, and *insightful*). Analogously, opposed descriptors include *shallow*, *simple*, and *boring*. Based on these adjectives, a profound conversational agent should be able to grasp topics of high complexity, be able to make non-trivial connections, and show a profound understanding that goes beyond simple facts. This factor is also negatively correlated with serviceability (−0.34), which indicates that, despite being *useful*, *knowledgeable*, and *insightful*, the high degree of complexity disagrees with aspects of *serviceability* such as *concise*, *precise*, *efficient*, and *direct*. Thus, profoundness should only be increased in scenarios where being quick and efficient can be traded for complexity and wisdom. For example, a profound agent could be useful in mental health therapy, but less so in customer service.

Instability. This factor mainly describes functional problems of the conversational agent through adjectives such as *contradictory*, *repetitive*, and *dysfunctional*. A conversational agent with high *instability* might give the impression that the underlying conversational system is failing or has been badly designed, which causes the agent to appear *scatter-brained*, *confused*, and *absentminded*. We also found aspects of artificiality in this factor (*fake*, *unrealistic*, and *inconsistent*). Thus, we interpret this factor as functional *instability*. Reducing aspects of functional *instability* was already investigated in previous work on LLM-based conversational agents in education [9] because contradictory and repetitive statements may destroy the conversational flow and thereby negatively impact the learning gain.

Vibrancy. This factor primarily contains adjectives describing positive feelings such as joy and pleasure (*joyful*, *playful*, *enthusiastic*, *cheerful*, *passionate*, and *romantic*). Also, multiple adjectives describe the readiness for action (*adventurous*, *brave*, and *engaging*). Thus, we interpret this trait as *vibrancy*. A conversational agent expressing *vibrancy* could captivate its interlocutor with joy and positivity. For example, a virtual fitness coach [49] could encourage users to reach their goals. In contrast, adjectives with negative loadings include *emotionless*, *cold*, *computerized*, and *robotic*. Thus, decreasing *vibrancy* in a conversational agent reduces strong emotions and makes the agent appear dull and apathetic, which can be useful in domains where the agent should maintain a neutral and formal standpoint such as in information retrieval.

Engagement. This factor heavily concentrates on adjectives that convey interest in the interlocutor (*interested*, *inquisitive*, *curious*, *communicative*, and *proactive*) and show empathy and willingness to help (*caring*, *supportive*, and *social*). We interpret this factor as *engagement* since all mentioned groupings relate to the act of engaging with others in an interaction. An engaging conversational agent keeps the conversation going and proactively encourages the interlocutor to participate while maintaining a supportive and empathetic tone. This aspect of human-agent interaction is often desired and expected [72] and has already been successfully integrated into virtual social companions, which effectively counteracted daily stressors and increased the users' well-being [101].

Neuroticism. This factor is highly associated with *neuroticism* form the Big Five model and is described by adjectives surrounding emotional volatility and the tendency to experience strong negative emotions (*frustration*, *depression*, *anger*, and *loneliness*). Due to a high overlap with *neuroticism* from the Big Five personality traits, we interpret this factor as *neuroticism*. While there are several applications where the conversational agent should score low in *neuroticism* (e.g., customer service, education, and health care), this factor can be useful in entertainment. For example, virtual non-player characters (NPCs) in video games can be designed to exhibit *neuroticism* in order to create emotionally nervous personalities [39].

Serviceability. This factor is associated with *conscientiousness* from the Big Five model and is dominated by adjectives describing a very rational and contained personality (*logical*, *precise*, *organized*, and *functional*). Furthermore, adjectives such as *concise*, *objective*, and *articulate* suggest that the agent is able to provide unbiased information in a dense but understandable way. Thus, we interpret this factor as *serviceability*. Such characteristics

are especially important for service-oriented agents in customer support or as personal assistants [36] because the provided service would be carried out efficiently and with the highest care. This is further supported by the high overlap of *serviceability* with *conscientiousness* (see Figure 8) and *conscientiousness* being crucial in task-oriented scenarios [19].

Subservience. The adjectives corresponding to this factor describe introversion, insecurity, and obedience in a dominantly negative way. Thus, we interpret this factor as *subservience*. A conversational agent scoring high in this trait would behave like an oppressed servant, selflessly obeying (*submissive* and *apologetic*) in a reserved way (*shy*, not *reserved*). On the contrary, an agent low in *subservience* would appear as dominant and confident. In the gaming industry, this factor has already been extensively used to create dominant protagonists [50], but can potentially also be used to create submissive characters.

6.2 GPT-3's Average Personality

As depicted in Figure 7a, the personality of GPT-3 does not uniformly cover the latent factor space, which suggests that GPT-3 is inclined towards a certain personality type. Furthermore, the exhibited personality remained consistent across different personae and types of conversation starts (see Figure 7b and 7c), which indicates that GPT-3's inclination towards the reported average personality is independent of the persona it assumes. Based on our analysis, GPT-3 was generally considered as *decent*, *engaging*, and *serviceable*. This aligns with our expectation given that GPT-3 was trained on cleaned data to reduce toxic output in favor of objective and factually reliable output [41]. Furthermore, given that the chatbots used almost double the number of words in their utterances compared to the users (see Section 3.5), it is not surprising that the chatbots scored high in *engagement*, a factor entailing an interested and talkative personality. The chatbots exhibited low *neuroticism* and *vibrancy*, i.e., the chatbots generally showed little affection and emotional volatility, which agrees with previous findings on GPT-3's self-reported personality [73]. We also expected to observe a high level of *instability* given that previous studies already highlighted GPT-3's limitations regarding consistency and factual reasoning [14, 35]. However, such functional and semantic issues (contradiction, confusion, inconsistency, etc.) did not particularly stand out in our experiment. Furthermore, the chatbots were not noticeably *profound*, although GPT-3 possesses the ability to discuss complex topics. We believe this is directly linked to high *serviceability*, i.e., complex and philosophical output appears as verbose and inefficient, and *vice versa*. In fact, *profoundness* is negatively correlated with *serviceability* (−0.34, see Table 4). Finally, the chatbots appeared to be low in *subservience* and thereby exhibited a rather confident and extroverted behavior.

In summary, GPT-3 exhibited an average personality that is independent of the persona it assumes and is inclined towards a respectful and engaging interaction where the chatbot appears as knowledgeable and extroverted, but at the same time lacks strong emotionality and profoundness. Furthermore, functional stability is still an issue that needs to be addressed in future chatbot versions. However, we stress that this personality profile merely describes the average personality. Through prompt engineering, the personality can be altered, achieving a wide variety of personality profiles as demonstrated in this work.

6.3 Connection to the Big Five

Given that GPT-3 was predominantly trained on human-generated text, and human personality can be recognized from text [67, 77], we expect the Big Five personality traits to overlap with our chatbot's personality dimensions. As presented in Section 5.3, there is a substantial overlap between three Big Five traits (*conscientiousness*, *agreeableness*, and *neuroticism*) and the derived factors in the 5-factor, 8-factor and 10-factor solution (see Figure 8). Particularly, we observe a factor disentanglement after increasing the number of factors from five (see Figure 8a) to eight (see Figure 8b), which results in substantial overlap between *conscientiousness* and factor 7 (*serviceability*), *agreeableness* and factor 1 (*decency*), as well as *neuroticism* and factor 6 (*neuroticism*). On the other hand, *openness*

and *extraversion* can still not be attributed to single factors but are spread over multiple factors, even with 10 factors (see Figure 8c). We conclude that the Big Five fail to adequately describe the personality dimensions exhibited by GPT-3. In other words, GPT-3's personality structure contains aspects that differ from human-human interaction. For example, factor 3 (*instability*) in the 8-factor solution does not substantially coincide with any Big Five trait and as such describes an aspect exclusively relevant to human-chatbot conversations. Concretely, this dimension describes aspects of artificiality (*fake* and *unrealistic*), and a lack of functionality (*dysfunctional*) and clarity, causing confusion, contradiction, and unnecessary repetition. The same holds for factor 4 in the 5-factor solution describing almost exclusively non-human attributes connected to artificiality and stoicism (see the factor associations in Appendix C). Further, *vibrancy* and *engagement* both overlap with 4 out of 5 Big Five traits. Thus, these two dimensions are perceived as separate dimensions of GPT-3's personality, each consisting of a mix of the Big Five traits. We conclude that a GPT-3-based chatbot requires a separate personality model as the Big Five personality traits do not adequately represent the underlying personality structure of the chatbot.

6.4 Comparison to Existing Models

We compared our factors to the dimensions of Völkel et al. [103] and found three factor correspondences: *decency* and *approachable* (41% overlap of descriptors), *vibrancy* and *social-entertaining* (40% overlap), and *serviceability* and *serviceable* (27% overlap). A notable difference between our factors and the ten factors found by Völkel et al. is that they found two separate factors surrounding service and assistance (serviceable and social-assisting), and two separate factors surrounding functionality (dysfunctional and unstable), which are characteristic of interactions with a service-oriented system. This is an expected result since the personal assistants investigated by Völkel et al. (e.g., Siri and Alexa) are by design service-oriented. Furthermore, they found separate dimensions describing artificiality, (dys-)functionality, and self-consciousness which all describe aspects of (non-)human likeness. In other words, users focused on service-related, assistance-related, and functionality-related features when interacting with voice-based personal assistants. In contrast, our dimensions focus primarily on social-behavioral characteristics rather than assistance and human likeness. Our factors also exhibit a high overlap with three of the Big Five dimensions and do not contain an isolated dimension for artificiality or self-consciousness, but a mixture of artificiality with aspects of functional instability and stoicism, highlighting the convincing and believable performance of GPT-3-based chatbots when mimicking human-human conversations.

6.5 Implications and Potential Applications

Our findings show that recent language models such as GPT-3 exhibit human personality traits in a convincing way such that users are able to focus more on social-behavioral characteristics rather than functionality and human likeness. This implies that people's perception of conversational agents is in transition from perceiving conversational agents as tools to perceiving them as wholesome social companions, which aligns with past forecasts [20, 84]. However, although the recent technological advancements of LLMs might have pushed the aspect of human likeness into the background, aspects of artificiality still remain in all of the presented factor solutions. According to Clark et al. [23] there may be a limit to how close conversational agents can and should reflect human-human interaction and that our focus should shift towards task-oriented aspects. Our results support this claim and suggest that such a paradigm shift is already taking place.

Our work also has implications for the controllability of agent personality. Despite being inclined towards an average personality, the personality exhibited by GPT-3 varies inconsistently from conversation to conversation. However, many of the practical applications for conversational agents require a consistent and stable personality profile [32, 33, 109]. With our work, we provide a major step towards controlling the personality of conversational agents in a systematic way by allowing the agent personality to be quantified in terms of latent factors that match user perception and summarize the most salient personality features exhibited.

In the following, we present potential applications of our findings. For example, our factors could be incorporated into the design process of future conversational agents to ensure that the personality aligns with users' perceptions and to explicitly parameterize the agent's personality. We envision a conversational system where the agent's personality can be adjusted by directly increasing or decreasing the intensity of a dimension, causing an imminent effect on the agent's responses. For GPT-3 and alike, this can be achieved through prompt engineering or fine-tuning of the model [37, 79, 107]. Our factors can also be seen as a first step towards a universal chatbot personality model that would allow for direct comparisons of chatbots. By developing a dedicated inference model for agent personality analogously to inference systems for human personality [13, 21, 90, 94, 95, 99, 105], chatbot personalities can be directly compared. Improvements and adjustments could then be discussed in terms of a target personality formulated as a combination of our factors. In this context, our insights about the chatbot's average personality could be used to quantify the difficulty of personality alteration given that inverting a pronounced trait may be more difficult than enhancing it (e.g., making serviceable chatbots more serviceable could require less effort than making non-neurotic chatbots behave neurotically) which could serve chatbot developers as a first guidance in early development.

6.6 Ethical Considerations

Despite the numerous advantages and applications of LLM-based conversational agents, ethical aspects surrounding potential risks and challenges must be discussed. For example, Li et al. [60] found that various LLMs exhibit dark personality patterns if not properly fine-tuned, despite rigorous cleaning and pre-processing of the training data. Dechant et al. [27] extensively discussed how social interactions with NPCs in video games can involve harmful language that can affect the user's well-being if not properly controlled in the game engine. Further, Baidoo-Anu and Ansah [9] highlight how LLMs such as ChatGPT can spread false information and systematical biases among inexperienced users (e.g., children). Lastly, Pradhan and Lazar [78] discussed negative stereotyping surrounding predefined personality profiles, which could be solved in the future through enhanced customizability and control. On the other hand, restricting the generative power of LLMs too much can decrease the usefulness of LLMs (e.g., for automatic content generation). We believe that a user-centered and systematic design of conversational agents, as outlined in this work, can help address ethical challenges because it enables a quantification of chatbot personality based on user perception. Ultimately, such ethical considerations are key for conversational agents to become usable in a wide variety of applications.

6.7 Generalizability

Given the fast-paced release of bigger and more capable large language models such as ChatGPT, GPT-4, and alike, we discuss the potential for the generalizability of our findings to more recent models. GPT-3's successors improved on the ability to handle code data, enhanced factual correctness, handling bigger and more detailed prompts, complex reasoning, optimizations for chat interactions, and the detection and avoidance of harmful content [54]. In this work, the ability to handle code data is irrelevant. Furthermore, factual correctness and bigger prompt sizes are tangential because the conversations are fictional and the prompt size fits GPT-3. Moreover, high complex reasoning abilities in humans were shown to negatively correlate with neuroticism and positively correlate with openness while leaving the remaining traits unaltered [97], which, given GPT-3's average personality—low neuroticism, rather low profoundness (associated with openness), and high engagement (associated with openness)—suggests that GPT-3's potential lack of complex reasoning did not noticeably affect the perceived personality. While the influence of chat optimizations on the perception of personality remains uncertain, the high variability, engagement, and likability levels observed suggest that any differences in newer models would probably pertain to nuanced aspects of the found personality dimensions. Based on human

personality theory, personality spaces are generally robust to nuanced differences [43, 44]. Thus, we believe that our findings generalize to newer models.

6.8 Limitations and Future Work

Our descriptors were collected from text-based human-chatbot conversations, which neglect the aspect of embodiment. Thus, our findings may not transfer to embodied conversational agents. Future work could investigate whether embodiment has an influence on user perception of agent personality.

The majority of participants in both experiments were university students, which may not be representative of the general population in terms of perception of agent personality. Although the number of participants with and without prior chatbot experience was balanced, there might still be an age-related influence on the perception of chatbot personality. Furthermore, although most participants had a proficient English level, only 8% were native English speakers, potentially resulting in less variability in the adjectives used. Future work can investigate other cultures, social strata, and languages.

In the online survey, participants read three short conversations between a human and GPT-3. It is not clear whether this number of conversations is enough for participants to mentally grasp the chatbot's personality. While we increased chat content variability via several methods (see Sections 3 and 4), some of the descriptors may not have been identified based on the provided conversations, causing participants to rate randomly, which induces noise and decreases the variance accounted for by the factors. As soon as more people have become familiar with these new language models, future experiments could exclusively target highly experienced participants, or let the participants read a higher number of conversations.

In times of rapid technological advancements, new findings require constant validation. Despite GPT-3's outstanding capabilities for natural language understanding, more advanced models such as ChatGPT, GPT-4, and alike are constantly pushing the boundaries for LLM-simulated conversational agents, putting older models in the shades quicker than ever. With the GPT-3 series being deprecated as of January 2024 and open questions regarding the exact extent of the generalizability of our findings, the use of GPT-3 remains a limitation of our work. Therefore, future work should investigate the generalizability to the most recent models and track the influence of new systems on people's perception of agent personality, ultimately cross-examining the strengths and weaknesses of each model when expressing personality.

7 CONCLUSION

We have presented a systematic analysis of the personality dimensions expressed by GPT-3 during human-chatbot conversations. We conducted an interaction experiment with 86 participants interacting with a GPT-3-based chatbot for three weeks while regularly describing the chatbot's personality. Through various post-processing steps, the descriptors were reduced to a set of 147 unique personality descriptors. In an online survey, 425 participants read 3 chat conversations from the interaction experiment and rated the extent of the 147 descriptors in terms of the chatbot's personality on a 4-point scale (no, rather no, rather yes, yes). An exploratory factor analysis of the ratings revealed eight latent factors (i.e., *decency*, *profoundness*, *instability*, *vibrancy*, *engagement*, *neuroticism*, *serviceability*, and *subservience*). Three latent factors overlap with Big Five personality traits (*agreeableness*, *conscientiousness*, and *neuroticism*). In addition, three of our factors show similarities to the chatbot personality model of Völkel et al. [103] (*decency*, *vibrancy*, and *serviceability*). These findings imply that GPT-3-based chatbots are able to convincingly mimic human-to-human conversations, exhibiting more human-like personality traits compared to voice-based personal assistants. Furthermore, our results support that recent technological advancements in conversational systems contribute to conversational agents transitioning from being mere tools to being wholesome social companions . Nevertheless, our findings also highlight that dimensions identifying a

lack of functional *stability* and *profoundness* remain, which questions whether conversational agents can and should close the gap to human-to-human conversations.

We believe that our work constitutes an important step towards a unified and systematic design of personality in conversational agents that will enable the direct comparison of conversational agents through the quantification of agent personality in terms of the presented factors.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Andrew Gloster, Professor of Clinical Psychology at the University of Lucerne, for his contribution to associating new descriptors with the Big Five personality dimensions. The authors also thank all the participants for their time and effort. This work was supported by an ETH Zurich Research Grant under Grant No.: ETH-10 22-1.

REFERENCES

[1] Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. GPT-3-Driven Pedagogical Agents to Train Children's Curious Question-Asking Skills. *International Journal of Artificial Intelligence in Education* (jun 2023). https://doi.org/10.1007/s40593-023-00340-7

[2] Inworld AI. 2023. *The Future of NPCs - What gamers demand from next-gen characters.* techreport. https://inworld.ai/whitepapers/future-of-npcs

[3] Gordon W. Allport. 1927. Concepts of trait and personality. *Psychological Bulletin* 24, 5 (1927), 284–293. https://doi.org/10.1037/h0073629

[4] Sean Andrist, Bilge Mutlu, and Adriana Tapus. 2015. Look Like Me: Matching Robot Personality via Gaze to Increase Motivation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* ACM, 3603–3612. https://doi.org/10.1145/2702123.2702592

[5] Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-based Approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany). ACM, 1–20. https://doi.org/10.1145/3544548.3581441

[6] Michael C. Ashton and Kibeom Lee. 2007. Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. *Personality and Social Psychology Review* 11, 2 (may 2007), 150–166. https://doi.org/10.1177/1088868306294907

[7] Michael C. Ashton and Kibeom Lee. 2008. The HEXACO Model of Personality Structure and the Importance of the H Factor. *Social and Personality Psychology Compass* 2, 5 (jul 2008), 1952–1962. https://doi.org/10.1111/j.1751-9004.2008.00134.x

[8] Matthew P. Aylett, Alessandro Vinciarelli, and Mirjam Wester. 2020. Speech Synthesis for the Generation of Artificial Personality. *IEEE Transactions on Affective Computing* 11, 2 (apr 2020), 361–372. https://doi.org/10.1109/taffc.2017.2763134

[9] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *SSRN Electronic Journal* (2023), 22 pages. https://doi.org/10.2139/ssrn.4337484

[10] Gene Ball and Jack Breese. 2000. *Emotion and Personality in a Conversational Agent.* The MIT Press, 189–219. https://doi.org/10.7551/mitpress/2697.003.0009

[11] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29, 4 (2001), 1165–1188.

[12] Stefan Benus, Marian Trnka, Eduard Kuric, Lukáš Marták, Agustín Gravano, Julia Hirschberg, and Rivka Levitan. 2018. Prosodic entrainment and trust in human-computer interaction. In *Speech Prosody 2018.* ISCA. https://doi.org/10.21437/speechprosody.2018-45

[13] Shlomo Berkovsky, Ronnie Taib, Irena Koprinska, Eileen Wang, Yucheng Zeng, Jingjie Li, and Sabina Kleitman. 2019. Detecting Personality Traits Using Eye-Tracking Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, 1–12. https://doi.org/10.1145/3290605.3300451

[14] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (feb 2023). https://doi.org/10.1073/pnas.2218523120

[15] Matthias Braunhofer, Mehdi Elahi, and Francesco Ricci. 2014. User Personality and the New User Problem in a Context-Aware Point of Interest Recommender System. In *Information and Communication Technologies in Tourism 2015.* Springer International Publishing, 537–549. https://doi.org/10.1007/978-3-319-14343-9_39

[16] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are

Few-Shot Learners. https://doi.org/10.48550/ARXIV.2005.14165
[17] Raymond B. Cattell. 1947. Confirmation and clarification of primary personality factors. *Psychometrika* 12, 3 (sep 1947), 197–220. https://doi.org/10.1007/bf02289253
[18] Raymond B. Cattell. 1966. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research* 1, 2 (apr 1966), 245–276. https://doi.org/10.1207/s15327906mbr0102_10
[19] Ana Paula Chaves and Marco Aurelio Gerosa. 2020. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (nov 2020), 729–758. https://doi.org/10.1080/10447318.2020.1841438
[20] Jessie Y. C. Chen and Michael J. Barnes. 2014. Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues. *IEEE Transactions on Human-Machine Systems* 44, 1 (feb 2014), 13–29. https://doi.org/10.1109/thms.2013.2293535
[21] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. 2013. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing* 17, 3 (2013), 433–450.
[22] Seung Youn Yonnie Chyung, Katherine Roberts, Ieva Swanson, and Andrea Hankinson. 2017. Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Performance Improvement* 56, 10 (nov 2017), 15–23. https://doi.org/10.1002/pfi.21727
[23] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3290605.3300705
[24] David M. Condon, Joshua Coughlin, and Sara J. Weston. 2022. Personality Trait Descriptors: 2,818 Trait Descriptive Adjectives Characterized by Familiarity, Frequency of Use, and Prior Use in Psycholexical Research. *Journal of Open Psychology Data* 10, 1 (jan 2022), 1. https://doi.org/10.5334/jopd.57
[25] Paul T. Costa and Robert R. McCrae. 1992. Four ways five factors are basic. *Personality and Individual Differences* 13, 6 (jun 1992), 653–665. https://doi.org/10.1016/0191-8869(92)90236-i
[26] Robert Dale. 2020. GPT-3: What's it good for? *Natural Language Engineering* 27, 1 (dec 2020), 113–118. https://doi.org/10.1017/s1351324920000601
[27] Martin Johannes Dechant, Robin Welsch, Julian Frommel, and Regan L Mandryk. 2022. (Don't) stand by me: How trait psychopathy and NPC emotion influence player perceptions, verbal responses, and movement behaviours in a gaming task. In *CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3491102.3502014
[28] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-Powered Conversational agents. arXiv. https://doi.org/10.48550/ARXIV.1811.01241
[29] Paul Ekman. 1992. Are there basic emotions? *Psychological Review* 99, 3 (1992), 550–553. https://doi.org/10.1037/0033-295x.99.3.550
[30] Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3-4 (may 1992), 169–200. https://doi.org/10.1080/02699939208411068
[31] Hans Jürgen Eysenck. 1963. Biological basis of personality. *Nature* 199 (1963), 1031–1034.
[32] Daniel Fernau, Stefan Hillmann, Nils Feldhus, Tim Polzehl, and Sebastian Möller. 2022. Towards Personality-Aware Chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Edinburgh, UK, 135–145. https://aclanthology.org/2022.sigdial-1.15
[33] Marta Ferreira and Belem Barbosa. 2023. A Review on Chatbot Personality and Its Expected Effects on Users. In *Trends, Applications, and Challenges of Chatbot Technology*. IGI Global, 222–243. https://doi.org/10.4018/978-1-6684-6234-8.ch010
[34] Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. 2017. Personality Traits and Music Genres. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM. https://doi.org/10.1145/3079628.3079693
[35] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines* 30, 4 (nov 2020), 681–694. https://doi.org/10.1007/s11023-020-09548-1
[36] Asbjørn Følstad, Marita Skjuve, and Petter Bae Brandtzaeg. 2019. Different Chatbots for Different Purposes: Towards a Typology of Chatbots to Understand Interaction Design. In *Internet Science*. Springer International Publishing, 145–156. https://doi.org/10.1007/978-3-030-17705-8_13
[37] Ivar Frisch and Mario Giulianelli. 2024. LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*. Association for Computational Linguistics, St. Julians, Malta, 102–111. https://aclanthology.org/2024.personalize-1.9
[38] Yue Fu, Rebecca Michelson, Yifan Lin, Lynn K. Nguyen, Tala June Tayebi, and Alexis Hiniker. 2022. Social Emotional Learning with Conversational Agents. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (jul 2022), 1–23. https://doi.org/10.1145/3534622
[39] Francesco Garavaglia, Renato Avellar Nobre, Laura Anna Ripamonti, Dario Maggiorini, and Davide Gadia. 2022. Moody5: Personality-biased agents to enhance interactive storytelling in video games. In *2022 IEEE Conference on Games (CoG)*. IEEE. https://doi.org/10.1109/cog51982.2022.9893689

[40] Radhika Garg and Subhasree Sengupta. 2020. He Is Just Like Me. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (mar 2020), 1–24. https://doi.org/10.1145/3381002

[41] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[42] Lewis R Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology* 2, 1 (1981), 141–165.

[43] Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology* 59, 6 (1990), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

[44] Lewis R. Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological Assessment* 4, 1 (mar 1992), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

[45] Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist* 48, 1 (1993), 26–34. https://doi.org/10.1037/0003-066x.48.1.26

[46] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1, 1 (2007), 77–89. https://doi.org/10.1080/19312450709336664

[47] Javier Hernandez, Jina Suh, Judith Amores, Kael Rowan, Gonzalo Ramos, and Mary Czerwinski. 2023. Affective Conversational Agents: Understanding Expectations and Personal Influences. https://doi.org/10.48550/ARXIV.2310.12459

[48] Graeme Hutcheson. 1999. *The Multivariate Social Scientist*. SAGE Publications, Ltd. https://doi.org/10.4135/9780857028075

[49] W. A IJsselsteijn, Y. A. W. de Kort, J Westerink, M. de Jager, and R Bonants. 2006. Virtual Fitness: Stimulating Exercise Behavior through Media Technology. *Presence: Teleoperators and Virtual Environments* 15, 6 (dec 2006), 688–698. https://doi.org/10.1162/pres.15.6.688

[50] Jeroen Jansz and Raynel G. Martis. 2007. The Lara Phenomenon: Powerful Female Characters in Video Games. *Sex Roles* 56, 3-4 (feb 2007), 141–148. https://doi.org/10.1007/s11199-006-9158-0

[51] Hang Jiang, Xiajie Zhang, Xubo Cao, and Jad Kabbara. 2023. PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences. https://doi.org/10.48550/ARXIV.2305.02547

[52] Magnus Johansson, Björn Strååt, Henrik Warpefelt, and Harko Verhagen. 2014. Analyzing the Social Dynamics of Non-Player Characters. In *Frontiers in Gaming Simulation*. Springer International Publishing, 173–187. https://doi.org/10.1007/978-3-319-04954-0_21

[53] Henry F. Kaiser. 1970. A second generation little jiffy. *Psychometrika* 35, 4 (dec 1970), 401–415. https://doi.org/10.1007/bf02291817

[54] Katikapalli Subramanyam Kalyan. 2024. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal* 6 (March 2024), 100048. https://doi.org/10.1016/j.nlp.2023.100048

[55] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stepha Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (apr 2023), 102274. https://doi.org/10.1016/j.lindif.2023.102274

[56] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (jul 2018), 1–26. https://doi.org/10.1145/3214273

[57] Diane M. Korngiebel and Sean D. Mooney. 2021. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *npj Digital Medicine* 4, 1 (jun 2021). https://doi.org/10.1038/s41746-021-00464-x

[58] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2022. Interacting with educational chatbots: A systematic review. *Education and Information Technologies* 28, 1 (jul 2022), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

[59] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[60] Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2022. Does GPT-3 Demonstrate Psychopathy? Evaluating Large Language Models from a Psychological Perspective. https://doi.org/10.48550/ARXIV.2212.10529

[61] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

[62] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. CAiRE: An End-to-End Empathetic Chatbot. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (apr 2020), 13622–13623. https://doi.org/10.1609/aaai.v34i09.7098

[63] Bingjie Liu and S. Shyam Sundar. 2018. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 10 (oct 2018), 625–636. https://doi.org/10.1089/cyber.2018.0110

[64] Irene Lopatovska. 2020. Personality Dimensions of Intelligent Personal Assistants. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (CHIIR '20)*. ACM. https://doi.org/10.1145/3343413.3377993

[65] Irene Lopatovska, Elena Korshakova, Diedre Brown, Yiqiao Li, Jie Min, Amber Pasiak, and Kaige Zheng. 2021. User Perceptions of an Intelligent Personal Assistant's Personality: The Role of Interaction Context. In *Proceedings of the 2021 Conference on Human*

Information Interaction and Retrieval (CHIIR '21). ACM. https://doi.org/10.1145/3406522.3446018
[66] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA". In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/2858036.2858288
[67] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research* 30 (nov 2007), 457–500. https://doi.org/10.1613/jair.2349
[68] Richard E. Mayer, W. Lewis Johnson, Erin Shaw, and Sahiba Sandhu. 2006. Constructing computer-based tutors that are socially sensitive: Politeness in educational software. *International Journal of Human-Computer Studies* 64, 1 (jan 2006), 36–42. https://doi.org/10.1016/j.ijhcs.2005.07.001
[69] Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality* 60, 2 (jun 1992), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x
[70] Margaret McRorie, Ian Sneddon, Etienne de Sevin, Elisabetta Bevacqua, and Catherine Pelachaud. 2009. A Model of Personality and Emotional Traits. In *Intelligent Virtual Agents*. Springer Berlin Heidelberg, 27–33. https://doi.org/10.1007/978-3-642-04380-2_6
[71] Margaret McRorie, Ian Sneddon, Gary McKeown, Elisabetta Bevacqua, Etienne de Sevin, and Catherine Pelachaud. 2012. Evaluation of Four Designed Virtual Agent Personalities. *IEEE Transactions on Affective Computing* 3, 3 (jul 2012), 311–322. https://doi.org/10.1109/t-affc.2011.38
[72] Christian Meurisch, Cristina A. Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring User Expectations of Proactive AI Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (dec 2020), 1–22. https://doi.org/10.1145/3432193
[73] Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. 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, Abu Dhabi, UAE, 218–227. https://aclanthology.org/2022.nlpcss-1.24
[74] Yusuke Mori and Youichiro Miyake. 2022. Ethical Issues in Automatic Dialogue Generation for Non-Player Characters in Digital Games. In *2022 IEEE International Conference on Big Data (Big Data)* (Osaka, Japan). IEEE, 5132–5139. https://doi.org/10.1109/bigdata55660.2022.10020271
[75] Warren T. Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology* 66, 6 (jun 1963), 574–583. https://doi.org/10.1037/h0040291
[76] Maria Augusta S.N. Nunes and Rong Hu. 2012. Personality-based recommender systems. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM. https://doi.org/10.1145/2365952.2365957
[77] James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* 77, 6 (1999), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296
[78] Alisha Pradhan and Amanda Lazar. 2021. Hey Google, Do You Have a Personality? Designing Personality and Personas for Conversational Agents. In *CUI 2021 - 3rd Conference on Conversational User Interfaces (CUI '21)*. ACM. https://doi.org/10.1145/3469595.3469607
[79] Angela Ramirez, Mamon Alsalihy, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023. Controlling Personality Style in Dialogue with Zero-Shot Prompt-Based Learning. https://doi.org/10.48550/ARXIV.2302.03848
[80] Juan A. Recio-Garcia, Guillermo Jimenez-Diaz, Antonio A. Sanchez-Ruiz, and Belen Diaz-Agudo. 2009. Personality aware recommendations to groups. In *Proceedings of the third ACM conference on Recommender systems*. ACM. https://doi.org/10.1145/1639714.1639779
[81] Melanie Revilla and Jan Karem Höhne. 2020. How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research* 62, 5 (jul 2020), 538–545. https://doi.org/10.1177/1470785320943049
[82] Malik Sallam. 2023. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* 11, 6 (mar 2023), 887. https://doi.org/10.3390/healthcare11060887
[83] Katharine Sanderson. 2023. GPT-4 is here: what scientists think. *Nature* 615, 7954 (mar 2023), 773–773. https://doi.org/10.1038/d41586-023-00816-5
[84] Samer Muthana Sarsam and Hosam Al-Samarraie. 2018. A First Look at the Effectiveness of Personality Dimensions in Promoting Users' Satisfaction With the System. *SAGE Open* 8, 2 (apr 2018). https://doi.org/10.1177/2158244018769125
[85] Barry Schwartz. 2015. The Paradox of Choice. *Positive Psychology in Practice* (apr 2015), 121–138. https://doi.org/10.1002/9781118996874.ch8
[86] Hanieh Shakeri, Carman Neustaedter, and Steve DiPaola. 2021. SAGA: Collaborative Storytelling with GPT-3. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. ACM. https://doi.org/10.1145/3462204.3481771
[87] Noora Shrestha. 2021. Factor Analysis as a Tool for Survey Analysis. *American Journal of Applied Mathematics and Statistics* 9, 1 (jan 2021), 4–11. https://doi.org/10.12691/ajams-9-1-2
[88] Michael Shumanov and Lester Johnson. 2021. Making conversations with chatbots more personalized. *Computers in Human Behavior* 117 (apr 2021), 106627. https://doi.org/10.1016/j.chb.2020.106627
[89] Julius Sim and Chris C Wright. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy* 85, 3 (mar 2005), 257–268. https://doi.org/10.1093/ptj/85.3.257

[90] Marcin Skowron, Marko Tkalčič, Bruce Ferwerda, and Markus Schedl. 2016. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram. In *Proceedings of the 25th International Conference Companion on World Wide Web* (Montréal, Québec, Canada) *(WWW '16 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 107–108. https://doi.org/10.1145/2872518.2889368

[91] Tuva Lunde Smestad and Frode Volden. 2019. Chatbot Personalities Matters. In *Internet Science*. Springer International Publishing, 170–181. https://doi.org/10.1007/978-3-030-17705-8_15

[92] Christopher J. Soto and Oliver P. John. 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology* 113, 1 (jul 2017), 117–143. https://doi.org/10.1037/pspp0000096

[93] Mirjam Stieger, Marcia Nißen, Dominik Rüegger, Tobias Kowatsch, Christoph Flückiger, and Mathias Allemand. 2018. PEACH, a smartphone- and conversational agent-based coaching intervention for intentional personality change: study protocol of a randomized, wait-list controlled trial. *BMC Psychology* 6, 1 (sep 2018). https://doi.org/10.1186/s40359-018-0257-9

[94] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. 2018. ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Transactions on Affective Computing* 9, 2 (2018), 147–160. https://doi.org/10.1109/TAFFC.2016.2625250

[95] Hung-Yue Suen, Kuo-En Hung, and Chien-Liang Lin. 2019. TensorFlow-Based Automatic Personality Recognition Used in Asynchronous Video Interviews. *IEEE Transactions on Affective Computing* 7 (2019), 61018–61023. https://doi.org/10.1109/ACCESS.2019.2902863

[96] Teo Susnjak. 2022. ChatGPT: The End of Online Exam Integrity? https://doi.org/10.48550/ARXIV.2212.09292

[97] Angelina R. Sutin, Yannick Stephan, Martina Luchetti, Jason E. Strickhouser, Damaris Aschwanden, and Antonio Terracciano. 2021. The Association Between Five Factor Model Personality Traits and Verbal and Numeric Reasoning. *Aging, Neuropsychology, and Cognition* 29, 2 (Jan. 2021), 297–317. https://doi.org/10.1080/13825585.2021.1872481

[98] Ekaterina Svikhnushina and Pearl Pu. 2021. Key Qualities of Conversational Chatbots – the PEACE Model. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14-17, 2021, College Station, TX, USA*. ACM, 520–530. https://doi.org/10.1145/3397481.3450643

[99] Tommy Tandera, Hendro, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetio. 2017. Personality Prediction System from Facebook Users. *Procedia Computer Science* 116 (2017), 604–611. https://doi.org/10.1016/j.procs.2017.10.016 Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).

[100] Jos M.F. ten Berge, Wim P. Krijnen, Tom Wansbeek, and Alexander Shapiro. 1999. Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra Appl.* 289, 1 (1999), 311–318. https://doi.org/10.1016/S0024-3795(97)10007-6

[101] Ta V, Griffith C, Boatfield C, Wang X, Civitello M, Bader H, DeCero E, and Loggarakis A. 2020. User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. *Journal of medical Internet Research* 22, 3 (March 2020), e16235.

[102] Michelle M.E. Van Pinxteren, Mark Pluymaekers, and Jos G.A.M. Lemmink. 2020. Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management* 31, 2 (mar 2020), 203–225. https://doi.org/10.1108/josm-06-2019-0175

[103] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376210

[104] Sarah Theres Völkel, Ramona Schoedel, Lale Kaya, and Sven Mayer. 2022. User Perceptions of Extraversion in Chatbots after Repeated Use *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 253, 18 pages. https://doi.org/10.1145/3491102.3502058

[105] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, Markus Gross, and Christian Holz. 2022. Affective State Prediction from Smartphone Touch and Sensor Data in the Wild. In *SIGCHI Conference on Human Factors in Computing Systems* (New Orleans, Louisiana) *(CHI '22)*. ACM, New York, NY, USA. https://doi.org/10.1145/3491102.3501835

[106] Henrik Warpefelt and Harko Verhagen. 2017. A model of non-player character believability. *Journal of Gaming & Virtual Worlds* 9, 1 (mar 2017), 39–53. https://doi.org/10.1386/jgvw.9.1.39_1

[107] Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. 2024. ControlLM: Crafting Diverse Personalities for Language Models. https://doi.org/10.48550/ARXIV.2402.10151

[108] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 3 (jun 2022), 3081–3089. https://doi.org/10.1609/aaai.v36i3.20215

[109] Akihiro Yorita, Simon Egerton, Jodi Oakman, Carina Chan, and Naoyuki Kubota. 2019. Self-Adapting Chatbot Personalities for Better Peer Support. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 4094–4100. https://doi.org/10.1109/smc.2019.8914583

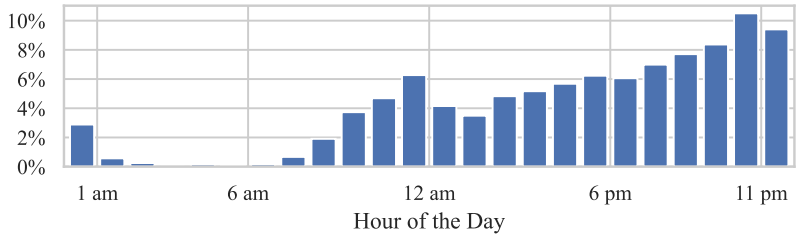
APPENDICES

A DATA COLLECTION STATISTICS

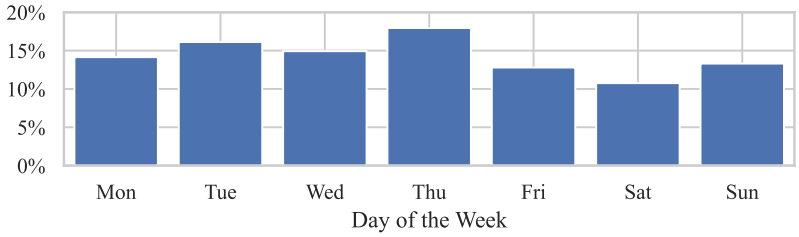
We present additional statistics for the two data collection experiments. Specifically, we extend our discussion regarding the activity distribution (Section 3.5), the influence of the persona information and conversation start type on the distribution of adjectives (Section 3.5), and the survey completion time (Sections 4.2 and 4.4).

A.1 Activity Distribution (Interaction Experiment)

As discussed in Section 3.5, participants' activity patterns are equally distributed over weekdays and over the day in the interval [8 a.m., 12 a.m.] with increased activity around 12 p.m. and after 6 p.m. Figure A.1 and Figure A.2 provide additional insights into the participants' activity patterns. Despite the activity increasing around 12 p.m. and after 6 p.m. (see Figure A.1a), which coincides with common working hours and free time, this increase is not directly visible from Figure A.2a due to the high variability among participants. We believe this comes from the fact that most of the participants were university students. Thus, common working hours do not apply, and activity can be distributed differently. Nevertheless, Figure A.1a reveals that the mean activity substantially increases around 12 p.m. and after 6 p.m., congruent with the participants' expected free time. In fact, 49.0% of the activity took place after 6 p.m. Similarly, we observe in Figure A.1b that the participation was evenly distributed with respect to weekdays, which we also attribute to participants being mostly university students and not following common weekly routines. Furthermore, participants were asked to be active on at least ten different days during three weeks (see Section 3.1), which naturally balances the usage patterns to a certain degree (i.e., the participants were required to show activity on at least four different weekdays by the pigeonhole principle).

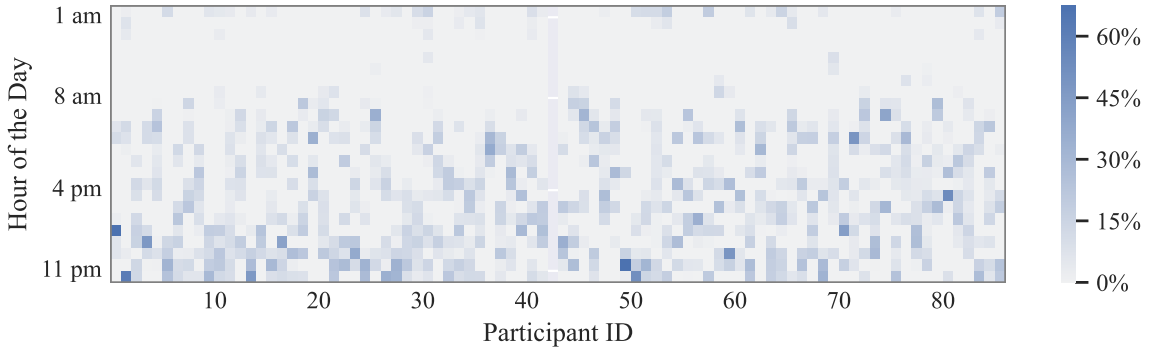


(a) Mean participation distribution by the hour of the day.

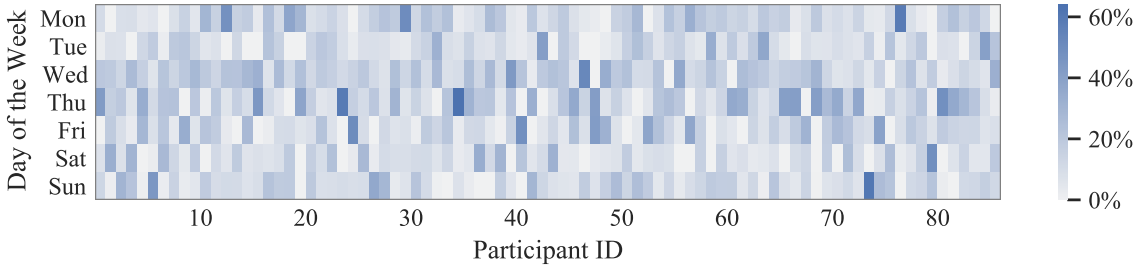


(b) Mean participation distribution by the day of the week.

Fig. A.1. Mean participation distributions over all participants based on the percentage of the total number of messages sent (a) by the hour of the day and (b) by the day of the week.



(a) Heat map showing the engagement of participants over the day on average.

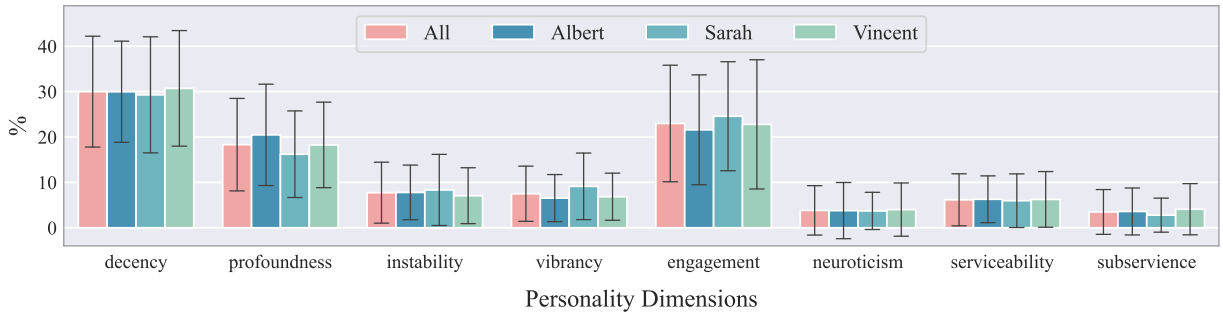


(b) Heat map showing the engagement of participants over weekdays on average.

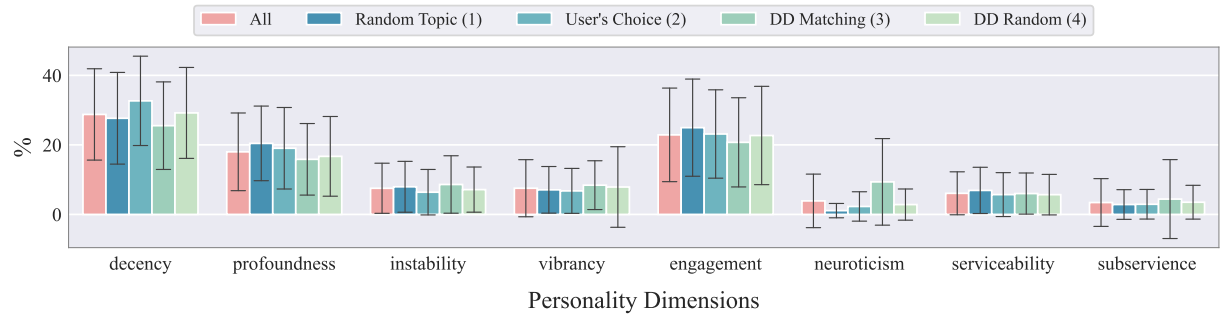
Fig. A.2. Two heat maps showing the distribution of participants' engagement over the day and over weekdays. The values indicate the average percentage of messages written at (a) the time of day, and (b) on the weekday with respect to the total number of messages written.

A.2 Influence of Prompt Variations (Interaction Experiment)

The prompting structure presented in Section 3.2 and Section 3.3 aims to create diverse conversational contexts in order to avoid repetitiveness in both the content of the conversations and the personality exhibited by the chatbots. To ensure that the chosen prompting parameters did not systematically bias the participants' perception of the chatbots, we investigated the distribution of adjectives used in the self-reports based on the prompting parameters, indicating the proportion of adjectives associated (both negatively and positively) with each trait. As depicted in Figure A.3, the variance of the proportions is high, indicating high variability in the conversations, while the average is consistently similar across all chatbot personae and conversation start types. Notable differences only exist for neuroticism where the conversations starting with a pre-sampled random sentence caused a higher proportion of neuroticism-related adjectives to appear in the self-reports, which, among other factors, stems from the sentences belonging to the "fear" class in the DailyDialog dataset [61] (see Section 3.3 for details). We conclude that there was no measurable bias induced by the chatbot personae nor by the conversation start types used in the prompt.



(a) Adjective Distribution per Persona.



(b) Adjective Distribution per Conversation Start Type.

Fig. A.3. The average distribution of adjectives used in the interaction experiment per participant grouped by latent trait the adjective belongs to based on the 8-factor solution found in Section 5.1, and by chatbot persona (a) and per conversation start type (b). The error bars denote one standard deviation.

A.3 Survey Completion Time (Rating Experiment)

Table A.1 lists the average survey completion time per survey section over the final 425 participants (i.e., after exclusion, see Section 4.4). The preliminary estimate given in Section 4.2 closely fits the empirical results, except for reading the conversations, which took less time than expected. The total survey completion time did not exceed the required upper limit of 20 minutes. This table can help other researchers in the study design process for similar experiments.

Table A.1. Mean survey completion time per survey section and the corresponding standard deviations (SD) in minutes. *Conversations (1)* corresponds to reading the conversations the first time before the rating. *Conversations (2)* corresponds to reading the conversations a second time during the rating.

Section	Mean (min)	SD (min)
Introduction	0.81	1.38
Conversations (1)	3.49	1.50
Conversations (2)	0.21	0.51
Ratings	9.59	2.61
Questionnaire	1.24	0.71
Total	15.34	3.96

B INTERRATER RELIABILITY

Interrater reliability metrics are tools to assess whether the raters actually agreed on the interpretation of the object that is being rated [89]. A commonly used metric is Krippendorff's alpha [46], which quantifies the interrater agreement on a scale from -1 (perfect disagreement) to 1 (perfect agreement), where 0 is the random chance level. We compute Krippendorff's alpha (α) for each pair of distinct raters and group them by the overlap of the conversations they read. Since each participant read three conversations, the overlap between two raters ranges from 0 to 3. Table B.1 shows the mean α by overlap. We report a *moderate* [59] agreement of 0.57 when the overlap is 3, and a *fair* [59] agreement of 0.31 when the overlap is 0. Since the order in which the conversations were displayed could have influenced the ratings, we additionally condition the last conversation to be identical (see the rightmost column in Table B.1). We report a *substantial* [59] agreement of 0.78 for an overlap of 3. However, for an overlap of 3, the sample sizes were small (23 and 9). Thus, we additionally perform a bootstrapping experiment to test whether the high agreement occurred by chance. To this end, we randomly sample a set of equal size (i.e., 23 and 9) from each of the sets where the overlap was smaller than 3. Then, we use Welch's t-test to compare the mean α when the overlap is 3 to the mean α from the randomly sampled sets and repeat this procedure $n = 10,000$ times. After correcting the p-value from the t-test for the false discovery rate using the Benjamini-Yekutieli procedure [11], we found that the reliability when the overlap is 3 is significantly higher (on the 99% level) than for all other sets. This experiment suggests that it is highly unlikely for the high reliability to occur by chance. We conclude that the interrater reliability is high enough for the data to be used in the subsequent steps.

Table B.1. Interrater reliability based on Krippendorff's Alpha (α) [46]. The α is computed for each pair of the 425 raters grouped by the overlap of the conversations displayed to the raters, and after conditioning the last conversation to be identical. N denotes the number of pairs for each group. The standard deviation is given in brackets.

	no ordering restriction		last conversation identical	
Overlap	Krippendorff's α	N	Krippendorff's α	N
Overlap 3	0.57 (0.31)	23	0.78 (0.25)	9
Overlap 2	0.50 (0.28)	230	0.65 (0.27)	41
Overlap 1	0.34 (0.14)	7,263	0.39 (0.12)	789
No Overlap	0.31 (0.00)	80,894	-	-

C ADDITIONAL FACTOR LOADINGS

In order to complement Section 5 where we analyzed GPT-3's underlying personality structure across a varying number of extracted factors, we provide here the factor loadings for the other two factor solutions (see Table C.1 for $x = 5$ factors and Table C.2 for $x = 10$ factors). The factor labels are based on a subjective interpretation of the authors and, when coinciding, indicate factor similarity in terms of the number of adjectives associated with those labels.

Table C.1. Overview of the latent factors resulting from an exploratory factor analysis using $x = 5$ factors and their top descriptors ranked by factor loadings. The factor labels are based on a subjective interpretation of the authors.

#	Factor Label	Top Descriptors by Factor Loadings
1	Vibrancy	enthusiastic (0.74), joyful (0.68), cheerful (0.59), social (0.59), adventurous (0.57), curious (0.55), motivated (0.55), passionate (0.53), playful (0.52), talkative (0.51), welcoming (0.49), optimistic (0.49), active (0.49), inquisitive (0.48), communicative (0.45), humorous (0.42), determined (0.42), interested (0.41), explorative (0.41), caring (0.40), engaging (0.40), proactive (0.39), affectionate (0.38), creative (0.38), inspiring (0.37), brave (0.37), generous (0.36), responsive (0.35), suggestive (0.34), sensitive (0.33), open-minded (0.32), interactive (0.31), casual (0.31), verbal (0.29)
2	Conscientiousness	logical (0.66), precise (0.63), efficient (0.63), organized (0.62), informative (0.60), smart (0.57), knowledgeable (0.56), intellectual (0.54), functional (0.48), self-disciplined (0.48), concise (0.48), thorough (0.47), objective (0.46), insightful (0.46), wise (0.45), formal (0.43), useful (0.42), stable (0.40), responsible (0.40), deep (0.40), articulate (0.38), consistent (0.38), diplomatic (0.37), helpful (0.36), mindful (0.35), considerate (0.35), contradictory (−0.34), complex (0.34), direct (0.32), philosophical (0.27), critical (0.27), understandable (0.26)
3	Civility	offensive (−0.65), rude (−0.64), arrogant (−0.64), respectful (0.62), polite (0.60), accepting (0.52), harsh (−0.51), confrontational (−0.49), humble (0.48), irritable (−0.47), tolerant (0.46), patronizing (−0.46), gentle (0.44), stubborn (−0.43), courteous (0.43), calm (0.43), agreeable (0.41), angry (−0.39), understanding (0.38), cooperative (0.38), careful (0.37), friendly (0.37), assertive (−0.37), patient (0.37), confident (−0.37), submissive (0.36), neutral (0.36), narrow-minded (−0.33), supportive (0.33), easygoing (0.32), self-centered (−0.32), overbearing (−0.30), reserved (0.28)
4	Artificiality	computerized (0.59), boring (0.59), emotionless (0.58), fake (0.57), robotic (0.57), annoying (0.52), human-like (−0.52), predictable (0.51), shallow (0.51), repetitive (0.48), vague (0.48), haphazard (0.42), dysfunctional (0.40), cold (0.38), confusing (0.38), creepy (0.37), simple (0.37), realistic (−0.36), inhibited (0.33), old-fashioned (0.33), dependent (0.33), self-aware (−0.26)
5	Neuroticism	depressed (0.60), pessimistic (0.57), negative (0.57), fearful (0.55), complaining (0.54), frustrated (0.53), agitated (0.50), lonely (0.49), upset (0.46), shy (0.45), helpless (0.44), worried (0.44), moody (0.43), confused (0.42), scatterbrained (0.41), lost (0.41), preoccupied (0.36), absentminded (0.35), pensive (0.34), careless (0.33), nostalgic (0.32), defensive (0.30), deceitful (0.29), romantic (0.28)

Table C.2. Overview of the latent factors resulting from an exploratory factor analysis using $x = 10$ factors and their top descriptors ranked by factor loadings. The factor labels are based on a subjective interpretation of the authors.

#	Factor Label	Top Descriptors by Factor Loadings
1	Decency	offensive (−0.65), respectful (0.65), polite (0.63), rude (−0.63), harsh (−0.54), arrogant (−0.51), courteous (0.45), tolerant (0.45), irritable (−0.44), friendly (0.43), humble (0.40), gentle (0.40), accepting (0.40), patronizing (−0.36), easygoing (0.35), agreeable (0.35), understanding (0.34), calm (0.33), cold (−0.33), annoying (−0.33), confrontational (−0.32), creepy (−0.32), responsive (0.30), cooperative (0.29), narrow-minded (−0.28), understandable (0.27)
2	Instability	scatterbrained (0.68), confusing (0.66), absentminded (0.62), contradictory (0.58), confused (0.58), lost (0.56), vague (0.48), haphazard (0.47), dysfunctional (0.42), evasive (0.41), helpless (0.40), careless (0.38), consistent (−0.35), dependent (0.33), stable (−0.28), defensive (0.26)
3	Neuroticism	complaining (0.67), depressed (0.66), frustrated (0.64), negative (0.61), agitated (0.60), pessimistic (0.59), upset (0.55), moody (0.47), angry (0.47), lonely (0.38), worried (0.35), fearful (0.34), self-centered (0.29)
4	Engagement	inquisitive (0.64), interested (0.58), curious (0.58), motivated (0.48), supportive (0.47), talkative (0.43), mindful (0.39), communicative (0.37), considerate (0.36), explorative (0.36), caring (0.36), open-minded (0.36), proactive (0.36), social (0.35), helpful (0.29)
5	Vibrancy	joyful (0.62), playful (0.61), passionate (0.54), humorous (0.53), affectionate (0.51), brave (0.51), enthusiastic (0.49), cheerful (0.47), adventurous (0.45), romantic (0.41), generous (0.38), creative (0.35), optimistic (0.30), welcoming (0.28), nostalgic (0.26)
6	Profoundness	deep (0.72), intellectual (0.55), wise (0.54), philosophical (0.52), complex (0.45), shallow (−0.44), smart (0.43), inspiring (0.40), critical (0.38), preoccupied (0.38), knowledgeable (0.38), simple (−0.38), insightful (0.32), pensive (0.30), useful (0.30), thorough (0.29)
7	Artificiality	robotic (0.70), computerized (0.70), predictable (0.59), human-like (−0.59), boring (0.54), fake (0.45), repetitive (0.39), emotionless (0.38), realistic (−0.37), formal (0.34), interactive (−0.32), engaging (−0.30)
8	Pragmatism	efficient (0.41), functional (0.40), informative (0.38), patient (0.38), logical (0.36), objective (0.36), neutral (0.33), precise (0.33), organized (0.33), concise (0.29)
9	Subservience	submissive (0.52), shy (0.48), reserved (0.47), inhibited (0.45), careful (0.44), old-fashioned (0.38), self-disciplined (0.37), apologetic (0.37)
10	Decisiveness	confident (0.44), stubborn (0.40), assertive (0.38), verbal (0.33), determined (0.32), direct (0.31)

D QUESTIONNAIRES

The questionnaires used in the interaction experiment (see Section 3) consisted of a pre-study and a post-study questionnaire. The corresponding questions and answer options can be found in Table D.1 (pre-study questionnaire) and Table D.2 (post-study questionnaire).

Table D.1. Pre-Study Questionnaire. All answer options were single-selection only.

#	Question	Answer Options
1	How would you rate your English level?	"Beginner (A1)", "Elementary Level (A2)", "Low intermediate level (B1)", "High intermediate level (B2)", "Advanced level (C1)", "Proficiency (C2)", "Native English speaker"
2	Have you held a conversation of any kind with a chatbot before?	"No", "Yes, once", "Yes, a few times", "Yes, regularly"
3	If "Yes" in (2): Please, describe the chatbot(s) you have interacted with before.	—
4	How often are you using speech-to-text conversion tools?	"Never", "Rarely", "Weekly", "Daily"

Received 15 November 2023; revised 1 February 2024; accepted 5 April 2024

Table D.2. Post-Study Questionnaire. Questions 4 to 6 were skipped if the answer to Question 3 was "Yes". Furthermore, Question 6 was skipped if the answer to Question 5 was not "Other". All answer options were single-selection only.

#	Question	Answer Options
1	How old are you?	—
2	What is your gender?	"Female", "Male", "Other"
3	Are you a student / working at a university?	"No", "Yes"
4	If "Yes" in (3): Please, enter the name of your university.	—
5	If "Yes" in (3): Which degree are you currently pursuing?	"Bachelor", "Master", "PhD", "Other"
6	If "Other" in (5): Please, elaborate on the degree you are pursuing (e.g., PhD, postdoc, professor).	—
7	If "No" in (3): Please, enter your current job position.	—
8	How comfortable did you feel while conversing with all the chatbots in general?	"Not at all", "Little", "Medium", "Very"
9	Did you let other people hold conversations with the chatbots for you, or did you provide others with your login credentials?	"Never", "Sometimes", "Often", "Always"
10	Did you answer truthfully on the self-reports?	"Never", "Sometimes", "Often", "Always"
11	How much did you enjoy talking to *Albert*?	"Not at all", "Little", "Medium", "Very"
12	How much did you enjoy talking to *Sarah*?	"Not at all", "Little", "Medium", "Very"
13	How much did you enjoy talking to *Vincent*?	"Not at all", "Little", "Medium", "Very"
14	Did you answer truthfully on all questions in this survey?	"No", "Yes"