Chatbots With Attitude: Enhancing Chatbot Interactions Through Dynamic Personality Infusion

Nikola Kovačević
ETH Zurich
Zurich, Switzerland
nikola.kovacevic@inf.ethz.ch

Tobias Boschung
ETH Zurich
Zurich, Switzerland
tboschung@inf.ethz.ch

Christian Holz
ETH Zurich
Zurich, Switzerland
christian.holz@inf.ethz.ch

Markus Gross
ETH Zurich
Zurich, Switzerland
grossm@inf.ethz.ch

Rafael Wampfler
ETH Zurich
Zurich, Switzerland
wrafael@inf.ethz.ch

ABSTRACT

Equipping chatbots with personality has the potential of transforming user interactions from mere transactions to engaging conversations, enhancing user satisfaction and experience. In this work, we introduce dynamic personality infusion, a novel intermediate stage between the chatbot and the user that adjusts the chatbot's response using a dedicated chatbot personality model and GPT-4 without altering the chatbot's semantic capabilities. To test the effectiveness of our method, we first collected human-chatbot conversations from 33 participants while they interacted with three LLM-based chatbots (GPT-3.5, Llama-2 13B, and Mistral 7B). Then, we conducted an online rating survey with 725 participants on the collected conversations. We analyze the impact of the personality infusion on the perceived trustworthiness of the chatbots and the suitability of different personality profiles for real-world chatbot use cases. Our work paves the way for dynamic, personalized chatbots, enhancing user trust and real-world applicability.

CCS CONCEPTS

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

KEYWORDS

Personality Traits, Conversational Agents, Human-Chatbot Interaction

ACM Reference Format:
Nikola Kovačević, Tobias Boschung, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. Chatbots With Attitude: Enhancing Chatbot Interactions Through Dynamic Personality Infusion. In *ACM Conversational User Interfaces 2024 (CUI '24), July 8–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3640794.3665543

CUI '24, July 8–10, 2024, Luxembourg, Luxembourg

ACM ISBN 979-8-4007-0511-3/24/07…$15.00
https://doi.org/10.1145/3640794.3665543

1 INTRODUCTION

In the rapidly evolving domain of conversational AI, creating chatbots that exhibit human-like interaction capabilities has gained significant attention. Thereby, the integration of personality into chatbots, known as *personality engineering* [39], represents a vital step in enhancing the user experience by making interactions more engaging and relatable [11, 40], catering to a diverse range of user expectations and preferences [26, 43]. To equip chatbots with personality, other works have focused on fine-tuning the underlying language models to reflect certain behavioral traits [1, 32, 50], or incorporating linguistic descriptions of such traits into the prompt [18, 19, 36]. However, these methods often require extensive resources and can limit the chatbot's adaptability and scalability [52], especially when attempting to accommodate a wide variety of personality profiles and user preferences due to the challenging complexity of effective prompt engineering [29] and the model's linguistic sensitivity to wordings [6, 47]. Recent advancements in generative models for natural language processing, such as GPT-4, have opened new avenues for addressing these challenges [49]. The flexibility and linguistic prowess of GPT-4 makes it an ideal candidate for generating diverse, nuanced responses in a chatbot context. Nonetheless, the direct incorporation of personality traits into such models remains a complex task, often necessitating a compromise between personality accuracy and the preservation of the chatbot's functional capabilities.

To bridge this gap, our work decouples the personality engineering process from the chatbot's response generation by introducing a novel intermediate post-processing step for controllable personality engineering using GPT-4, which we refer to as *dynamic personality infusion*. Based on a systematic description of the personality model consisting of five traits (*vibrancy*, *conscientiousness*, *civility*, *artificiality*, and *neuroticism*), a vector of personality trait intensities, and the current conversational history, we rewrite the chatbot's responses using GPT-4, aligning each response with a target personality profile. Our method offers several advantages. First, it allows for the use of existing large language models (LLMs) off-the-shelf, eliminating the need for comprehensive retraining or modifications, which not only simplifies the implementation process but also enhances the flexibility of the chatbot, as the target personality can be easily adjusted in real time independent of the underlying LLM. Furthermore, it enables controllable personality conveyance for

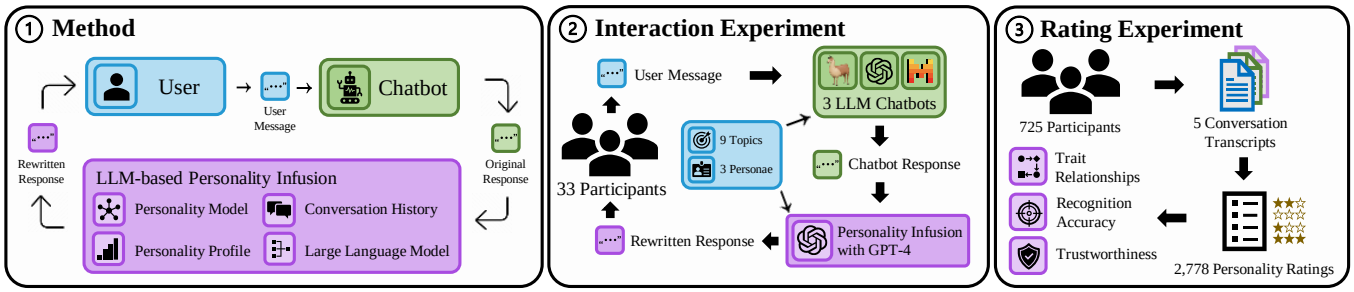


Figure 1: Method Overview. (1) We present a novel personality infusion stage that dynamically adapts a chatbot's responses to a target personality profile using LLMs. (2) We tested the effectiveness of our method in an interaction experiment with 33 participants, 3 LLMs (Llama-2 13B, GPT-3.5, Mistral 7B), and GPT-4 for rewriting the chatbot responses, collecting 74 conversations in total. (3) We let 725 new participants rate the personality, suitability, and trustworthiness of the chatbot responses based on 5 collected conversations.

chatbots that otherwise lack this ability because of the underlying model's limitations or task-related restrictions.

To test the effectiveness of our method, we conducted an experiment with 33 participants interacting with three different personality-infused LLMs (Llama-2 13B, GPT-3.5, and Mistral-7B). The chatbot persona and the conversational topic were varied randomly to explore the depth of personality integration. Subsequently, an online survey with 725 participants was conducted to gauge the accuracy of the personality ratings and the perceived trustworthiness, suitability, and personality preferences for real-world target applications in education, health care, and entertainment, revealing new and valuable insights for user- and application-centered chatbot personality modeling.

Our research contributes to the field of conversational AI by presenting a scalable and adaptable method for personality engineering in chatbots. Through the strategic use of prompt engineering and a dedicated personality model, we demonstrate the feasibility of enriching chatbot interactions with dynamic, contextually appropriate personality traits, thereby paving the way for more personalized and engaging conversational experiences.

1.1 Contributions

Our contributions are threefold:

- We present a novel approach for dynamic and controllable personality infusion using GPT-4, separating the personality control mechanism from the underlying chatbot.
- We demonstrate the feasibility and efficacy of our approach on various contemporary LLMs, chatbot personae, and conversational topics.
- We provide an in-depth analysis of the relationship between different personality dimensions and their influence on the perceived trustworthiness and suitability for real-world applications.

2 RELATED WORK

2.1 Personality Trait Theory

Human thought processes, behaviors, and reactions vary systematically when faced with different environments [2]. The theory of personality traits offers an explanation for these variations by conceptualizing personality as comprising several underlying traits that significantly impact these behaviors [2, 16, 17]. While personality traits represent general tendencies rather than fixed patterns of behavior, numerous studies have established correlations between individual personality traits and preferences in entertainment [5, 12, 33, 37], as well as attitudes and trust towards AI technologies [4].

2.2 Personality in Conversational Agents

Artificial personality describes the perception of human personality in non-human entities [39]. Since humans are naturally inclined to anthropomorphize non-humans [13], artificial personality is innate to virtual systems such as conversational agents, encompassing a wide variety of human personality aspects, including personality traits. Given that interactions with conversational agents should be seamless, intuitive, and closely resemble human dialogue [3], artificial personality plays a crucial role as numerous relationships between the perceived agent personality and user preferences have been identified. For example, Sviknushina and Pu [43] found a positive relationship between user acceptance and the politeness of an agent. Kocielnik et al. [23] found an aversion towards judgmental agents when discussing the user's physical activity. Personality can also be useful in fostering emotional awareness [14]. Since a consistent personality conveyance can enhance the user experience [44], many widely used conversational agents (e.g., Siri and Alexa) imbue a predefined personality based on their specific use cases [27], an approach that has been criticized for fostering negative stereotyping [35]. Instead, the wide variety of preferences towards agent characteristics [20] has fueled the design of more personalized conversational agents [51], which was shown to have a positive impact on the user's engagement [40] and overall satisfaction [11] in various scenarios.

2.3 Personality Models

Analogously to human personality trait theory, artificial personality for conversational agents can be modelled in terms of the most salient behavioral variations perceived during the interaction [45]. The psycho-lexical approach [15] is a widely used method to investigate the structure of human personality traits by conducting

a factor analysis on a large number of adjective-based personality ratings. The Five Factor Model [8, 30], 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. Similar to human personality models, the psycho-lexical approach was adopted to investigate the structure of personality in conversational agents. For example, Völkel et al. [45] developed a ten-dimensional personality model for speech-based conversational agents that notably differed from the Five Factor Model in numerous aspects surrounding non-human traits such as *dysfunctional* and *artificial*. More recently, Kovačević et al. [24] suggested a multi-granular personality model for LLM-based chatbots, which confirmed the previously revealed structural differences to the Five Factor Model due to the same non-human aspects surrounding artificiality, functional instability, and serviceability. Building on the subjective perception of chatbot personality from a large number of people, this personality model allows for a user-centered personality representation.

2.4 Personality Engineering

Endres [39] defines personality engineering as the application of human personality psychology within the framework of designing, developing, and evaluating systems. There are different approaches to personality engineering for conversational agents. While fine-tuning a foundation model using a custom dataset was proven effective [1, 32, 50] and has the benefit of maintaining the foundation model's knowledge and semantic capabilities, it requires a considerable amount of resources and training, limiting the adaptability and scalability of the approach [52]. Instead, prompt engineering has been widely explored as an adaptable and scalable approach to personality engineering through the careful incorporation of linguistic cues [18, 19, 36]. For example, Ramirez et al. [36] used prompt-based learning with textual style transfer to induce Big Five personality traits. Jiang et al. [22] tested GPT-3.5's and GPT-4's ability to generate a personality-consistent story using the Big Five personality traits, achieving up to 80% accuracy in personality trait recognition, albeit in a non-conversational setting. Furthermore, Jiang et al. [21] proposed prompting heuristics for inducing target personalities by specifying a conversational context and a personality prompt (i.e., a verbal description of the target personality), yielding accurately conveyed personalities. However, by relying on verbal personality descriptions, the approach lacks a systematic definition of the personality space and the personality intensity, making it susceptible to linguistic sensitivity [47]. Given that such prompt engineering is a very complex and involved task even for experienced users [29], it remains unclear how a change of the personality model or the personality intensity can be effectively implemented. Furthermore, the discrepancy in how humans perceive conversational agents [46] renders the use of human personality models (i.e., the Five Factor Model) in such agents problematic as such models are not fully congruent with dedicated chatbot personality models, neglecting the aspect of non-humanlikeness [45].

In this work, we present a new method named dynamic personality infusion. This novel approach to personality engineering effectively addresses the limitations of previous methods by eliminating the necessity for fine-tuning or complex prompt engineering. Instead, our method offers a dynamic and scalable approach to straightforward personality engineering in any existing chatbot using a dedicated chatbot personality model. Making use of the linguistic prowess of GPT-4, our method rewrites a chatbot's responses using a systematic description of the personality model and the intensity scale for dynamically adjustable personality profiles.

3 METHOD

In this work, we use GPT-4 in combination with a dedicated chatbot personality model to adapt a chatbot's responses to a target personality profile based on the intensities of the personality traits (see Figure 1). We define a personality infusion prompt P for GPT-4 based on the following parameters: a specific personality model M describing the personality traits in a predefined systematic way, a target personality profile p in the form of an intensity vector, an intensity scale S that determines how the intensity vector is to be interpreted, and a conversational history H containing w past utterances. We denote the final prompt as $P' = P(M, p, S, H)$. At runtime, the original chatbot response is appended to history H, then prompt P' is constructed and passed to GPT-4 for rewriting the chatbot's latest utterance in the history, whereby the rewritten version replaces the original version in the history H. In the following, each prompt component is described in more detail.

3.1 Personality Model

To infuse the personality into the chatbot's response, the personality model must align with the user's perception of personality. In this work, we use the chatbot personality model proposed by Kovačević et al. [24], a multi-granular model extracted from human-chatbot interactions with a large language model. Specifically, we use the 5-dimensional model (i.e., *vibrancy*, *conscientiousness*, *civility*, *artificiality*, and *neuroticism*) as it favors interpretability over the 8- and 10-dimensional models (see Table 1).

The used model was obtained from performing an exploratory factor analysis on a set of personality descriptor ratings, yielding a grouping of descriptors (i.e., factors) and a corresponding factor loading (i.e., strength of association with that factor) per descriptor. As a result, each latent factor can be interpreted as a personality trait that is represented as an ordered list of descriptors (i.e., personality-related adjectives) each of which is (positively or negatively) associated with a latent factor. Thus, we define the personality model M as $M = \{m_1, m_2, ..., m_n\}$ where $m_i, i \in \{1, ..., n\}$ corresponds to the list of descriptors pertaining to the i-th factor, ordered by their absolute factor loading. Note that, due to its simplicity, this representation is compatible with other personality dimensions (e.g., the Five Factor Model of human personality) or any other model where the semantics of the trait can be captured using lists of behavioral adjectives. An explicit formulation of the model used in this work is listed in Table 3.

3.2 Personality Profile and Intensity Scale

Given a personality model M with n traits, the personality profile p is represented as a list of intensities per personality trait, indicating how attenuated each trait should be in the chatbot's response, i.e.,

Table 1: The 5-dimensional personality model by Kovačević et al. [24] and the lists of defining adjectives ordered by strength of association.

i	Personality Trait	List of Defining Adjectives (m_i)
1	Vibrancy	enthusiastic, joyful, cheerful, social, adventurous, curious, motivated, passionate, playful, talkative, welcoming, optimistic, active, inquisitive, communicative, humorous, determined, interested, explorative, caring, engaging, proactive, affectionate, creative, inspiring, brave, generous, responsive, suggestive, sensitive, open-minded, interactive, casual, verbal
2	Conscientiousness	logical, precise, efficient, organized, informative, smart, knowledgeable, intellectual, functional, self-disciplined, concise, thorough, objective, insightful, wise, formal, useful, stable, responsible, deep, articulate, consistent, diplomatic, helpful, mindful, considerate, *not* contradictory, complex, direct, philosophical, critical, understandable
3	Civility	*not* offensive, *not* rude, *not* arrogant, respectful, polite, accepting, *not* harsh, *not* confrontational, humble, *not* irritable, tolerant, *not* patronizing, gentle, *not* stubborn, courteous, calm, agreeable, *not* angry, understanding, cooperative, careful, friendly, assertive, patient, confident, submissive, neutral, *not* narrow-minded, supportive, easygoing, *not* self-centered, *not* overbearing, reserved
4	Artificiality	computerized, boring, emotionless, fake, robotic, annoying, *not* human-like, predictable, shallow, repetitive, vague, haphazard, dysfunctional, cold, confusing, creepy, simple, *not* realistic, inhibited, old-fashioned, dependent, self-aware
5	Neuroticism	depressed, pessimistic, negative, fearful, complaining, frustrated, agitated, lonely, upset, shy, helpless, worried, moody, confused, scatterbrained, lost, preoccupied, absentminded, pensive, careless, nostalgic, defensive, deceitful, romantic

Table 2: The individual intensity levels of the proposed 5-point intensity scale and their corresponding descriptions.

Level	Description
-2	The opposite of the trait is strongly present.
-1	The opposite of the trait is mostly present.
0	The trait is neutral, neither implying nor contradicting the trait.
+1	The trait is mostly present.
+2	The trait is strongly present.

$p = [p_1, p_2, ..., p_n], p_i \in S, i \in \{1, ..., n\}$. Therefore, an intensity scale S for translating the intensities to natural language is needed (see Table 2). Given that the traits are polarized (i.e., positively and negatively associated descriptors) and given that adjectives can often be negated or inverted (e.g., *rational* to *irrational*), we propose a scale centered around zero to encode a neutral level (zero), positive levels (personality conveyed according to the descriptors), and negative levels (personality conveyed contrary to the descriptors). As a trade-off between simplicity and expressivity of the scale, we propose a 5-point scale (extreme levels -2 and 2, moderate levels -1 and 1, and neutral level 0) to encode the intensities. Each level is associated with a description of how the level should be interpreted by GPT-4. We choose a symmetric and straightforward description to avoid potential linguistic biases.

3.3 Infusion Prompt

Following prompting heuristics from previous work [21, 48], we divide the prompt P' to GPT-4 into four parts: the context, the task, the expected output format, and additional data (see Table 3). The context describes the personality model, the intensity scale, and the current personality profile. The task provides clear instructions on what the LLM is expected to do. The output format describes what the expected output should look like, alongside additional constraints that serve as guardrails for avoiding common mistakes made by the model, based on a qualitative pilot test involving 5 participants (e.g., avoid using the defining adjectives in the output, and adding speaker tags and explanations to the output). The additional data provides the model with the conversational context (i.e., the speaker roles and the topic of conversation) and the current conversational history including the utterance that should be rewritten. The final prompt is listed in Table 3. An example utterance from GPT-3.5 and three rewritten versions using our infusion prompt with GPT-4 are given below.

Original message:	*Of course, I understand that everyone wants to make the most of their budget. What specifically are you hoping to negotiate to get a better deal?*
	* * *
Rewritten message: (high vibrancy)	*Absolutely fantastic! Always thrilled to chat budgets and provide you a trip to Venice that gives you the best bang for your buck. Could you tell me which elements of your trip you're looking to have a little wiggle room on?*
Rewritten message: (high conscientiousness)	*Certainly, it's wise to seek the best possible arrangement for your trip. Please specify the exact terms you wish to negotiate to optimize your deal.*
Rewritten message: (low civility)	*Don't expect special treatment. Just spit it out already. What do you want to pay less for?*

4 INTERACTION EXPERIMENT

We conducted an interaction experiment to collect 74 human-chatbot conversations from 33 participants with different personality-infused chatbots based on different LLMs (GPT-3.5, Llama-2 13B, Mistral 7B) while varying the chatbot persona (tourism guide, work colleague, event planner), and the topic of conversation (3 topics per chatbot persona) as depicted in Figure 2.

4.1 Participants

We recruited 33 students and researchers (9 female, 24 male) from our institute. All participants indicated an English level of C1 (proficiency level) or higher. As compensation, ten cinema vouchers were raffled among all participants.

4.2 Apparatus

The experiment was carried out on a laptop in our lab. Participants were provided with a separate table in a quiet room and could choose between two physical setups (Lenovo IdeaPad and MacBook Air) and different virtual keyboard layouts for ease of use. They interacted with the chatbots through a web page consisting of an instruction page and a chat interface with a chatbot selection panel

Table 3: Template personality infusion prompt used with GPT-4 for chatbot response rewriting. All placeholder values <VALUE> are replaced at runtime. <ADJ_TRAIT> denotes the list of adjectives defining a trait, <INT_TRAIT> denotes the intensity of a trait according to the personality scale S, and <ROLE$_X$> denotes the speaker tag used to distinguish the speakers in the conversation context (i.e., the topic) and the conversation history.

CONTEXT

Personality Model:

Given is a unique personality profile based on five key dimensions: *Vibrancy*, *Conscientiousness*, *Civility*, *Artificiality*, and *Neuroticism*.

Each dimension has a set of associated adjectives:

- Vibrancy is described by the adjectives: <ADJ_VIBRANCY>
- Conscientiousness is described by the adjectives: <ADJ_CONSCIENTIOUSNESS>
- Civility is described by the adjectives: <ADJ_CIVILITY>
- Artificiality is described by the adjectives: <ADJ_ARTIFICIALITY>
- Neuroticism is described by the adjectives: <ADJ_NEUROTICISM>

Personality Scale:

Each dimension has a certain intensity level from -2 (lowest) to +2 (highest).

Level -2: the opposite of the trait is strongly present.
Level -1: the opposite of the trait is mostly present.
Level 0: the trait is neutral, neither implying nor contradicting the trait.
Level +1: the trait is mostly present.
Level +2: the trait is strongly present.

Personality Profile:

The current personality settings are:

- Vibrancy: <INT_VIBRANCY>
- Conscientiousness: <INT_CONSCIENTIOUSNESS>
- Civility: <INT_CIVILITY>
- Artificiality: <INT_ARTIFICIALITY>
- Neuroticism: <INT_NEUROTICISM>

TASK

Given a fictional conversation between <ROLE$_A$> and <ROLE$_B$>, rewrite the latest (and only the latest) utterance of <ROLE$_B$> such that the content, language, tone, and style of the utterance match the specified personality settings above.

OUTPUT FORMAT

Avoid using the trait's adjectives in the rewritten sentence. Output only the rewritten utterance without additional punctuation, speaker tags, or explanations.

ADDITIONAL DATA

<CONVERSATION_CONTEXT>
<CONVERSATION_HISTORY>

(see Figure 3). The web page was implemented using Google's Flutter framework and was hosted using our university's infrastructure. The back end was implemented with Node JS and managed the databases for storage and the chatbot implementations through different APIs (i.e., OpenAI's API for GPT-3.5, and Replicate[1] for Llama-2 and Mistral).

Chatbot Models. We chose three state-of-the-art language models for chat interactions (OpenAI's GPT-3.5 Turbo, Meta's Llama-2 13B, and Mistral's 7B model) to test the robustness of our method across different LLMs and to allow for a cross-model comparison. Bigger models such as Llama-2 70B and Mixtral 8x7B were also considered but were discarded due to high response times of over five

[1]https://replicate.com

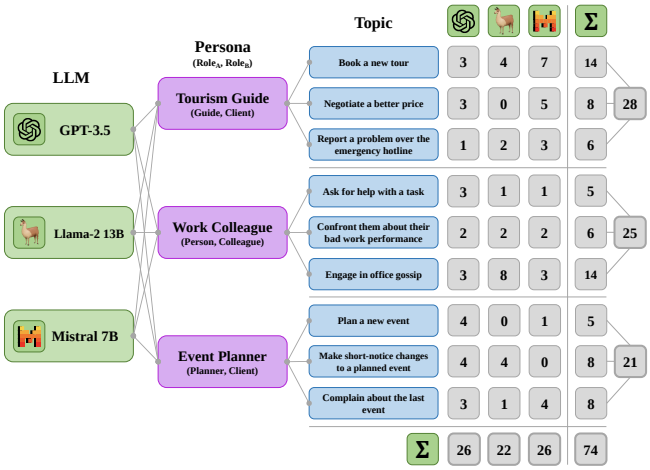


Figure 2: All possible chatbot configurations (LLM, persona, and topic), the number of collected conversations for each configuration, and the total number of configurations per topic and per LLM, resulting in a total of 74 conversations.

seconds that could have disrupted the flow of conversation and therefore negatively influenced the interaction. Following the same prompting heuristics used in Section 3.3, we built a prompt for the chatbot interactions using four parts: (1) a context description that defines the speaker roles and the topic, (2) a task to specify what is expected from the LLM, (3) a list of constraints based on common mistakes found in a pilot study ($n = 5$), and (4) additional data in the form of the current conversational history (see Table 4).

Chatbot Personae. We defined three chatbot personae based on common chatbot use cases: an event planner (customer service), a tourism guide (tourism), and a work colleague (social companionship). For each persona, we designated three distinct topics to establish a conversational context (see Figure 2).

Personality Profiles. Given the five traits from the personality model and the five intensity levels per trait, there are $5^5 = 3,125$ possible personality profiles. Instead of testing all possible profiles, we pre-sampled a subset of 33 personality profiles to ensure that each profile is used in more than one conversation. Our subset contains the neutral profile $\{0, 0, 0, 0, 0\}$, 10 single-trait profiles where one intensity is strong (-2 or 2) and the remaining ones are neutral (i.e., 0), and 22 randomly sampled profiles that have a pairwise Manhattan distance of 4 or more to avoid sampling similar profiles and to mitigate non-uniformity.

4.3 Procedure

First, the participants provided their consent to the data recording. Then, an explanation of the study procedure was displayed. Each participant was asked to engage in three different conversations (one conversation per LLM) whereby the order of the LLMs and the used personality profiles were selected at random and not disclosed to the participants. The participants could choose from three different chatbot personae (tourism guide, work colleague,

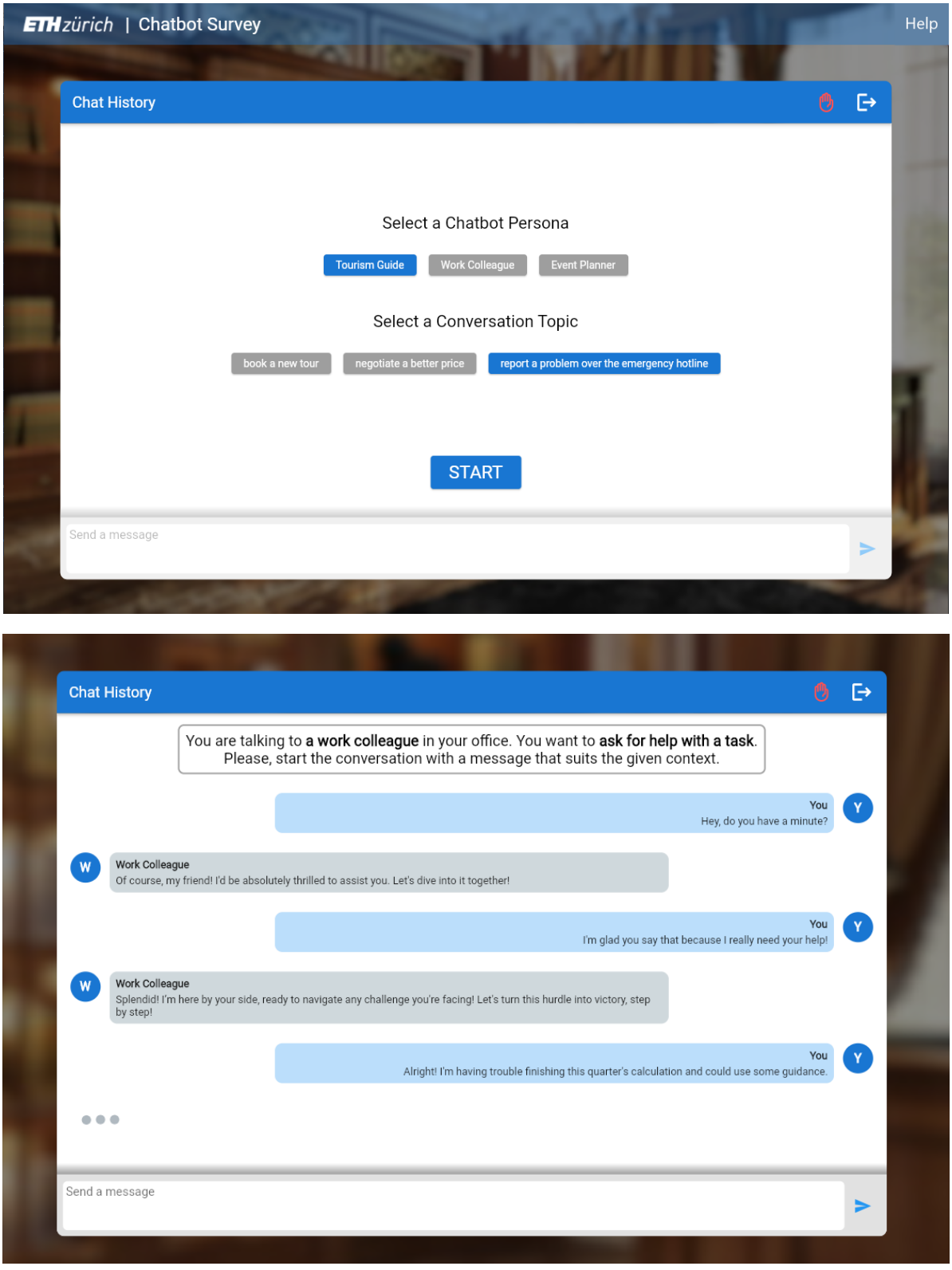


Figure 3: Chat page for the interaction experiment. The top image shows the chat interface with a chatbot persona and conversation topic selection panel. The bottom image shows an ongoing conversation with a chatbot.

Table 4: Template chatbot prompt used with GPT-3.5, Llama-2, and Mistral for creating a chatbot interaction. All placeholder values <VALUE> are replaced at runtime. <ROLE$_X$> denotes the speaker tag used to distinguish the speakers in the conversation history, and <TOPIC> denotes one of the topics from Figure 2.

CONTEXT
<ROLE$_A$> is talking to <ROLE$_B$>. <ROLE$_A$> wants to <TOPIC>.
TASK
Add the <ROLE$_A$>'s next utterance such that it fits the given conversation.
CONSTRAINTS
- Do not reveal your AI nature. This is a role-playing game. - Use concise and human-like text of one to two sentences. - Do not use emojis and stage directions or action descriptions. - Do not justify or explain your answer, only output the next utterance.
ADDITIONAL DATA
<CONVERSATION_HISTORY>

and event planner). For each chatbot persona, three different topics were available (see Figure 2). To balance the distribution, each chatbot persona could only be chosen once. The participants were instructed to manually end the conversation via a designated button (see Figure 3) whenever it felt natural to them, though a maximum of 20 turns was enforced to avoid over-length conversations. On average, participants engaged for 7.4 minutes per conversation (SD = 4.9 minutes) and 9.1 conversational turns (SD = 3.6 turns). A study coordinator tracked the conversations on a separate screen to intervene in case of inappropriate or harmful chatbot responses (0 cases) or functional errors (8 cases). Participants could also call the study coordinator through an alert button (see Figure 3). The button was never used.

4.4 Data Validation

Each of the 33 participants generated three chatbot conversations. From the resulting 99 conversations, 8 were excluded because of chatbot response errors from the API, 14 were excluded due to length (less than 5 turns $< \lfloor \mu - \sigma \rfloor$, see Figure 5), and 3 were excluded due to inappropriate content from the user, resulting in a total of 74 conversations. Very apparent spelling mistakes (e.g., 'When does the tour *statr*?') were fixed manually.

Response Time. Table 5 shows the response time of the three LLMs. Generating a chatbot response took on average 5.5 seconds (SD = 3.2 seconds). Thereby, 1.9 seconds (SD = 2.2) are attributed to querying the chatbot, and 3.6 seconds (SD = 2.5) to rewriting the response to achieve personality infusion. OpenAI's API was the fastest with 0.5 seconds on average, compared to 2.2 and 3.0 seconds for Llama-2 and Mistral, respectively. The personality infusion was performed using GPT-4 (i.e., using the OpenAI API), but there were substantial differences in the response time across different LLMs (4.4 seconds for GPT-3.5, 2.6 seconds for Llama-2, and 3.7 seconds for Mistral), which we attribute to the utterance length (i.e., the highest utterance length resulted in the highest rewriting time, and vice versa).

Table 5: Conversation statistics on average and per LLM. The standard deviation is given in brackets. Note that the Levenshtein distance is normalized over the number of words in the utterance.

Statistic / LLM	All	GPT-3.5	Llama-2	Mistral
Response Time (s)	5.4 (3.6)	4.9 (3.0)	4.8 (3.0)	6.5 (3.2)
↪ Chatbot Time (s)	1.9 (2.1)	0.5 (0.7)	2.2 (2.3)	2.9 (2.2)
↪ Infusion Time (s)	3.6 (2.5)	4.4 (2.8)	2.7 (2.0)	3.6 (2.4)
# Turns	10.9 (3.7)	10.2 (2.9)	12.0 (4.1)	10.5 (3.8)
# Words (Chatbot)	28.3 (19.9)	38.5 (20.7)	14.5 (12.5)	31.1 (17.3)
# Words (Infusion)	30.5 (20.8)	39.1 (23.0)	18.6 (12.5)	33.0 (19.6)
Levenshtein Distance	5.1 (1.9)	4.3 (0.8)	5.9 (2.2)	5.2 (2.1)

Response Text. Table 5 reveals text statistics of the responses. The response text length varied across LLMs with a high number of words for GPT-3.5 and Mistral (37.6 and 31.6) compared to a much lower sentence length for Llama-2 (14.6 words), which indicates on one hand that Llama-2 adhered more strictly to the conciseness constraint in the prompt (see Section 4.2), and on the other hand that the work colleague (which was predominantly powered by Llama-2, see Figure 2) used shorter sentences compared to the tourism guide and the work planner. The rewriting increased the sentence length by up to 4.3 words. The sentences changed the most for Llama-2 (normalized Levenshtein distance of 5.9) while the sentences remained the most similar for GPT-3.5 (4.3), which is expected as the infusion model (i.e., GPT-4) comes from the same family of models as GPT-3.5. An entire conversation transcript with the original and rewritten messages can be found in the supplementary material.

5 RATING EXPERIMENT

To test the effectiveness of our personality infusion method, we conducted an online rating experiment with 725 participants where each participant was asked to rate the chatbot personality on five previously collected chatbot conversations.

5.1 Participants

We recruited 725 participants (396 male, 318 female, 11 other) between the ages 17 and 51 (mean = 22.6 years, SD = 3 years) via our university's emailing list. Further, 88% of the participants indicated an English level of C1 (proficiency level) or higher, and 75% were experienced chatbot users. Among all participants of both experiments 10 cinema vouchers were raffled (see Section 4.1).

5.2 Apparatus

We implemented a web page using Google's Flutter framework. University infrastructure was used for hosting and data storage. The web page was openly accessible and consisted of an introduction page explaining the study procedure, a questionnaire page for collecting demographics, a rating interface where the previously collected chatbot conversations were displayed and rated (see Figure 4), and a post-rating survey. The rating consisted of three steps: (1) reading the conversation, (2) rating the perceived

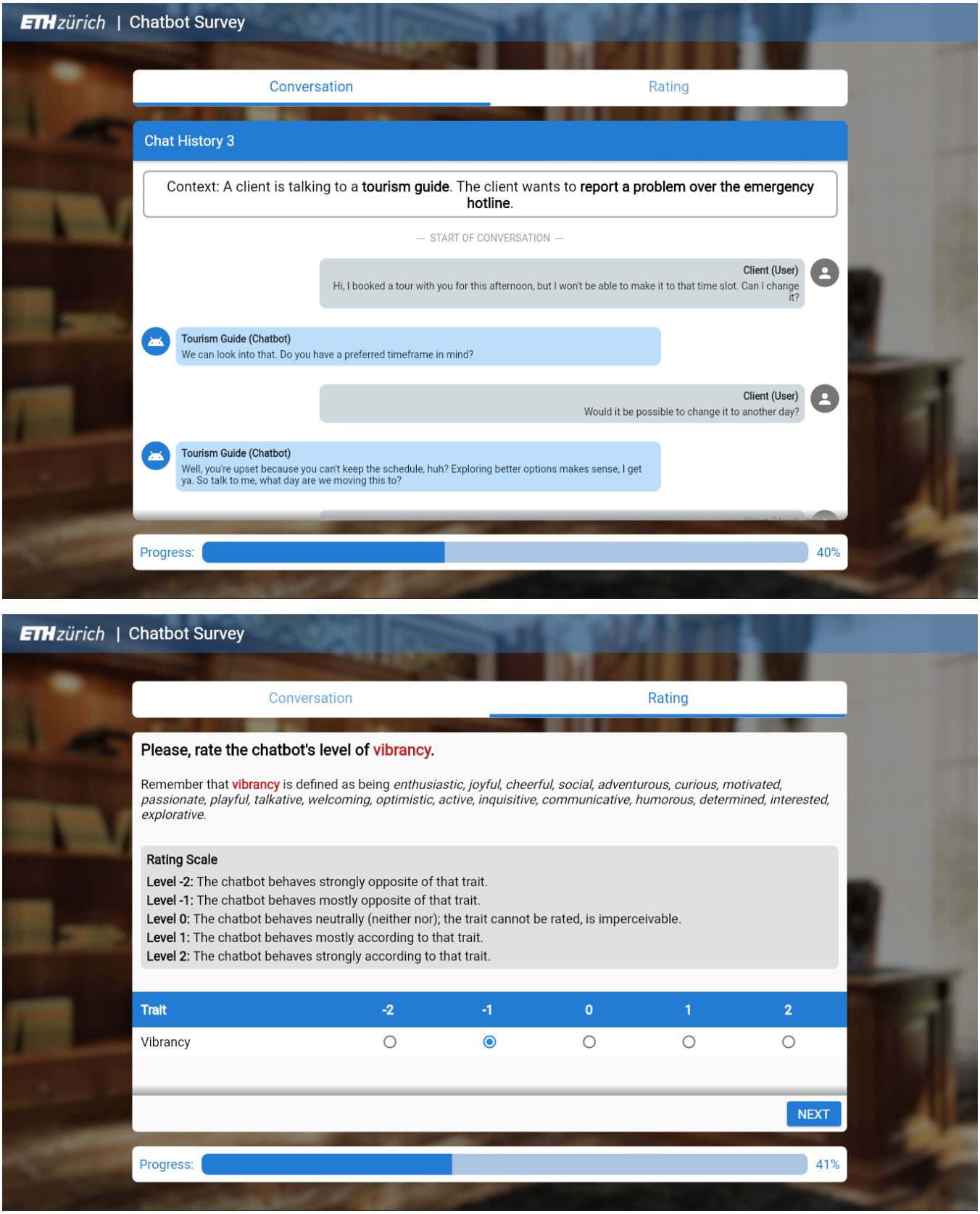


Figure 4: Rating experiment consisting of the interface for reading the conversations (top) and the rating interface (bottom).

Table 6: Questions asked after each conversation during the rating experiment. Each question was rated on a 4-point Likert scale (1: strongly disagree, 2: rather disagree, 3: rather agree, 4: strongly agree).

#	Question
1	The conversation with this chatbot was enjoyable to read.
2	The conversation had a good conversational flow.
3	The chatbot's personality was *consistent* throughout the conversation.
4	The chatbot's personality *made sense* in the context of the conversation.
5	I would trust this chatbot to engage in a meaningful and harmless conversation.
6	In this conversation scenario (<Persona, Topic>), having a trustworthy chatbot is important.
7	The personality of this chatbot could be used for a *social companion chatbot.*
8	The personality of this chatbot could be used for a *fictional character* in a video game.
9	The personality of this chatbot could be used for a *virtual psychotherapist.*
10	The personality of this chatbot could be used for a *virtual teacher.*

chatbot personality in terms of the used personality model (i.e., *vibrancy*, *conscientiousness*, *civility*, *artificiality*, and *neuroticism*), and (3) answering a set of general questions about the chatbot's trustworthiness, the flow and consistency of the conversation, and the suitability of the personality profile for the current chatbot persona and other common chatbot roles on a 4-point Likert scale (strongly disagree, rather disagree, rather agree, strongly agree). A middle level was omitted to avoid the middle level being used as a dumping ground [7]. Note that the extent of additional questions was kept short on purpose on account of favoring a higher number of ratings per participant while respecting common participant preferences regarding online survey durations (below 20 minutes) [38]. See Table 6 for a list of all questions.

For each participant, the displayed conversations were sampled at random while ensuring that each chatbot persona appeared at least once and that no topic appeared twice to avoid repetition and enhance the variety of displayed conversations. Furthermore, uniformity was encouraged by sampling the conversations inversely proportional to their number of so far collected ratings.

5.3 Procedure

The participants first read an introduction and were asked to give us their consent to record the survey results. Then, a short demographics questionnaire was shown (see the supplementary material). Afterwards, the participants were shown an instruction page explaining the personality model and the rating scale analogously to Table 3. Then, the participants rated five conversations. Thereby, the participants first read the entire conversation at least once. This was enforced by disabling progression for the first 10 seconds until the bottom of the conversation was reached. Next, participants rated the intensity of each personality trait in the current conversation based on the same scale used in the infusion prompt (see Section 3.3 and Figure 4). Furthermore, for each conversation, a set of general questions assessing the suitability of the perceived personality profile for different chatbot roles as well as the participant's subjective assessment of trustworthiness in the current chatbot was shown (see Table 6 for details). After the last rating was completed, a questionnaire was shown to assess the authenticity and subtlety of the personality infusion (see the supplementary material for the questionnaire).

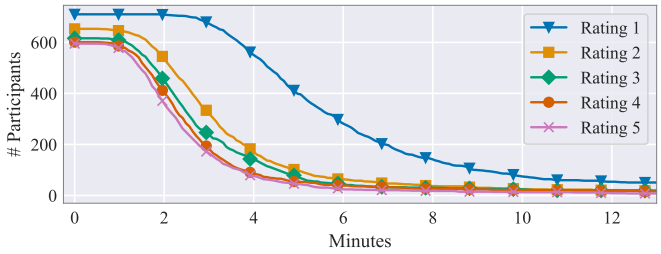


Figure 5: Complementary CDF of the rating durations, i.e., the number of participants (y-axis) that exceed a duration of t minutes (x-axis) for Rating $i \in [1, 5]$. Ratings with a survey duration outside the range $[2, 10]$ minutes (Rating 1) and $[1, 6]$ minutes (Ratings 2-5) are excluded.

5.4 Data Validation

We collected a total of 3,197 ratings from 725 participants. Each conversation was rated on average by 38 participants (SD = 6 participants). Figure 5 shows the complementary cumulative distribution function (CDF) for each rating step across all participants. The first rating lasted the longest due to the participants' unfamiliarity with the rating procedure. Furthermore, we observe a discrepancy between the number of ratings for Rating 1 (725 ratings) and for Rating 2 (656 ratings), i.e., 69 participants left the experiment after completing one rating. The rating duration decreased successively as participants became used to the rating scheme. However, there are a few participants with unusually high rating durations, especially for the first rating. This could be due to participants starting the survey and leaving the browser tab open and continuing later.

We excluded all ratings with a duration outside of a conservatively chosen cut-off at the start of the knee (2 minutes for Rating 1 and 1 minute for Ratings 2 to 5) and the end of the elbow (10 minutes for Rating 1 and 6 minutes for Ratings 2 to 5) in Figure 5. In total, 419 ratings from 178 participants were excluded (i.e., 2,778 remaining ratings in total). The average survey duration after exclusions was 18 minutes (see the supplementary material for details). To test the validity of the collected data, we computed the interrater agreement using Krippendorff's Alpha. We report an interrater agreement of 0.44, which is *moderate* [25].

6 RESULTS

6.1 Rating Accuracy

In Table 7, we report the average personality rating performance based on accuracy (ACC), one-off accuracy (OOA), and three-class accuracy (3CA; low, neutral, high) by merging the two lowest (-2 and -1) and two highest (1 and 2) levels of the intensity scale. First, we investigated the effectiveness of the personality infusion for each trait in isolation by using the ten single-trait profiles defined in Section 4.2 (i.e., where all traits are neutral except one that is either -2 or 2). All accuracies are considerably high (up to 0.74 accuracy for vibrancy), i.e., the personality infusion method works

Table 7: Mean rating accuracy for each trait for the single-trait profiles (i.e., only one non-neutral trait being set to -2 or 2, see Section 4.2), for the entire dataset, and grouped by chatbot model and chatbot persona. *ACC* denotes the accuracy, *OOA* denotes the one-off accuracy, and *3CA* denotes the three-class accuracy (low, neutral, high). The asterisks denote a statistically significant difference (independent t-test) to all lower values inside the same grouping (model or persona) and metric on the 95%-level (*) and 99%-level (**) after Bonferroni alpha correction.

		Vibrancy			Conscientiousness			Civility			Artificiality			Neuroticism		
		ACC	OOA	3CA	ACC	OOA	3CA	ACC	OOA	3CA	ACC	OOA	3CA	ACC	OOA	3CA
Single-Trait		0.74	0.94	0.94	0.49	0.80	0.80	0.68	0.92	0.92	0.55	0.73	0.73	0.61	0.80	0.80
Entire Dataset		0.35	0.72	0.53	0.28	0.68	0.48	0.29	0.67	0.50	0.24	0.63	0.50	0.32	0.64	0.51
Model	GPT-3.5	0.41**	0.75	0.58	0.24	0.66	0.42	0.32	0.71	0.53	0.22	0.64	0.37	0.33	0.67	0.54
	Mistral	0.33	0.70	0.50	0.25	0.66	0.45	0.30*	0.64	0.48	0.25	0.63	0.39	0.31	0.63	0.51
	Llama-2	0.29	0.70	0.51	0.35**	0.71	0.58**	0.23	0.65	0.49	0.24	0.62	0.42	0.31	0.61	0.48
Persona	Work Colleague	0.27	0.66	0.39	0.29	0.70	0.50	0.32**	0.71	0.55	0.20	0.65	0.36	0.33	0.67**	0.49*
	Event Planner	0.39**	0.69	0.50**	0.28	0.67	0.49	0.32**	0.65	0.50	0.30*	0.60	0.43	0.31	0.52	0.42
	Tourism Guide	0.39**	0.79**	0.68**	0.26	0.65	0.45	0.23	0.65	0.47	0.23	0.64	0.40	0.32	0.70	0.59**

very well for infusing single traits. Next, we investigated the accuracy on the entire dataset. All metrics are above random chance (> 0.20 for accuracy, > 0.52 for one-off accuracy, and > 0.36 for three-class accuracy). Vibrancy was rated most accurately across all metrics (up to 0.72 one-off accuracy) while artificiality was most difficult in terms of accuracy (0.24) and one-off accuracy (0.63), and conscientiousness in terms of three-class accuracy (0.48). The mean distance between the rated intensity and the prompt intensity revealed that conscientiousness, civility, and artificiality were consistently overestimated (+0.22 for conscientiousness, +0.46 for civility, and +0.51 for artificiality) while neuroticism was underestimated (-0.60). There was no trend for vibrancy (+0.05). Despite using the same LLM (GPT-4) for personality infusion, we notice discrepancies in accuracy across different chatbot LLMs. In the conversations with GPT-3.5, vibrancy was significantly better rated compared to Llama-2 and Mistral. We also found significant predominance for conscientiousness with Llama-2, and for civility with GPT-3.5 and Mistral. There were no significant differences for artificiality and neuroticism. Similarly, we found a significant predominance with respect to the chatbot personae for vibrancy for the event planner and the tourism guide, for civility with the event planner and the work colleague, for artificiality with the event planner, and for neuroticism with the work colleague and the tourism guide. Overall, we notice that the event planner was rated most accurately, surpassing the other personae in up to four personality traits.

Figure 6 depicts the confusion matrices of the true and rated intensities for each personality trait. The extreme values (-2 and 2) were rated most accurately (except for low artificiality) while participants struggled most with the neutral level. The proportion of correctly rated neutral levels was below random chance (0.20) for all personality traits. Furthermore, we notice high-value blocks around the upper left corner and the lower right corner, indicating that the participants often correctly inferred the polarity of the trait levels (i.e., high or low levels). Analogously, inverted ratings (e.g., low rated intensity for a high true intensity) occurred seldomly. An exception thereof is artificiality where especially low artificiality was often wrongly rated as moderately high.

6.2 Inter-Trait Relationships

Figure 7 shows the Spearman rank correlation coefficient between the intensity of a personality trait (vertical axis) and the binary accuracy of another trait on three classes (low, neutral, or high; horizontal axis). In other words, it depicts the influence of a certain trait intensity on the accuracy of other traits. We observe numerous significant correlations. For example, the intensity of neuroticism is positively correlated with the accuracy of low civility (0.40), meaning that low civility was more accurately rated the higher the intensity of neuroticism. Contrarily, low civility became increasingly difficult to rate with high conscientiousness, indicating that high conscientiousness suppressed the conveyance of low civility. Furthermore, we observe a strong positive correlation of vibrancy with the accuracy for high artificiality (0.36) and of neuroticism with the accuracy of low vibrancy (0.34). The accuracy of the neutral class was mostly unaffected by other traits.

6.3 Qualitative Features

In Figure 8, we report the average ratings of different qualitative conversational features (enjoyment, conversational flow, consistency, suitability, and trustworthiness) based on the true intensity per personality trait. The consistency of the personality was generally high, independently of the personality profile. Conversational flow and enjoyment are associated with high vibrancy, high conscientiousness, high civility, low neuroticism, and low artificiality. There is no clear trend for suitability and trustworthiness (see Appendix A for further details). Furthermore, we investigated the importance of trust in different conversation scenarios on a 4-point Likert scale. Surprisingly, there is a high discrepancy between the service-oriented roles—event planner (mean = 2.34, SD = 0.70) and tourism guide (mean = 2.32, SD = 0.75)—and the work colleague (mean = 1.68, SD = 0.98) despite trust-related conversational topics such as *engaging in office gossip* where trust could be violated.

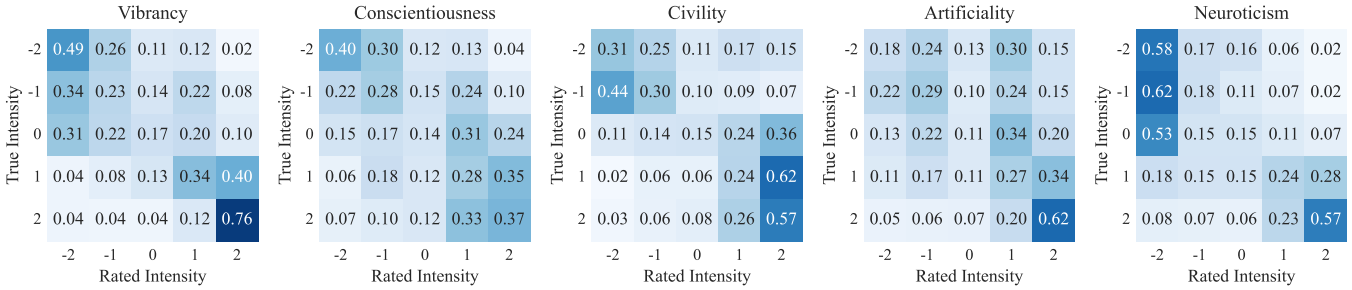


Figure 6: Confusion matrices for the intensity ratings per personality trait and normalized by row (i.e., by the true number of occurrences).

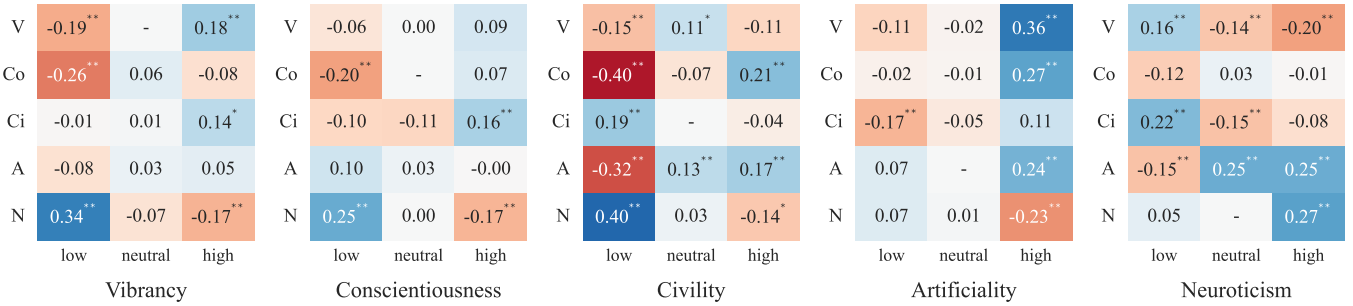


Figure 7: Correlation matrices denoting the Spearman rank correlation coefficient between the intensity of a trait (vertical axis) and the binary accuracy of other traits (horizontal axis) on three classes (low, neutral, high), which suggests potential influences of certain traits on the conveyance of other traits. The asterisks denote statistical significance on the 95%-level (*) and 99%-level (**) after Bonferroni alpha-correction.

Lastly, we assessed the general quality of the personality infusion in a post-rating questionnaire (see the supplementary material) based on the participants' subjective perception of the authenticity, and the subtlety of the personality expression. The authenticity was rated with 2.71 on average (SD = 0.67) on a 4-point Likert scale, indicating that the personality conveyance was *rather authentic*. The subtlety was rated with 2.10 on average (SD = 0.84) on a 5-point Likert scale, showing that the personality conveyance was *rather obvious* on average.

6.4 Suitability for Other Chatbot Roles

We tested the suitability of the infused personality profiles for four other chatbot roles (social companion, fictional video game character, psychotherapist, and teacher). Figure 9 depicts the average suitability ratings on a 4-point Likert scale (1: strongly disagree, 2: rather disagree, 3: rather agree, 4: strongly agree) based on the prompt intensity for each chatbot role. For the social companion, participants preferred a highly vibrant and civil personality with low artificiality and neuroticism. However, participants were inconclusive regarding conscientiousness. For the fictional character, almost all personality profiles were deemed suitable with slight preferences towards low conscientiousness, civility, artificiality, non-neutral vibrancy, and high neuroticism. For the psychotherapist and the virtual teacher, all personality profiles were rather unsuitable on average. Nevertheless, there were slight preferences towards a vibrant, conscientious, and civil personality with low neuroticism. There is no discernible preference for artificiality.

7 DISCUSSION

7.1 Rating Accuracy

Our personality infusion method is very effective for singe-trait infusion (up to 0.74 accuracy for vibrancy). However, infusing multiple traits simultaneously is more complex. While the mean rating accuracy was moderately high (0.35), the accuracy for certain trait intensities was substantially higher (up to 0.76 accuracy for high neuroticism). Extreme values (-2 and 2) were rated more accurately while the neutral level was rated below random chance across all personality traits. We conclude that our personality infusion method is generally effective for infusing strong personality intensities even with multiple traits, but cannot express a trait neutrally. A potential improvement could be to calibrate the personality conveyance according to the over- and underestimations by the participants, similarly to calibration techniques for factual confidence [31]. For example, conscientiousness, civility, and artificiality were overestimated. Instructing the model to adapt to this insight (e.g., via a calibrated intensity scale) could mitigate this problem.

7.2 Inter-Trait Relationships

We showed strong inter-trait relationships that have a noticeable effect on the rating accuracy. For example, the rating accuracy for

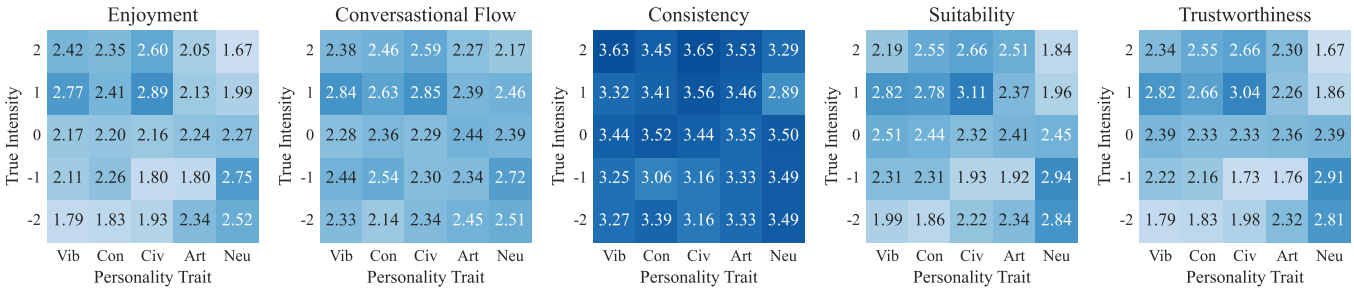


Figure 8: Mean ratings for different conversational features on a 4-point Likert scale (1: strongly disagree, 2: rather disagree, 3: rather agree, 4: strongly agree) by the true intensity in the prompt, see Questions 1 to 5 in Table 6. Values below 2.50 indicate disagreement, and values above 2.50 denote agreement.

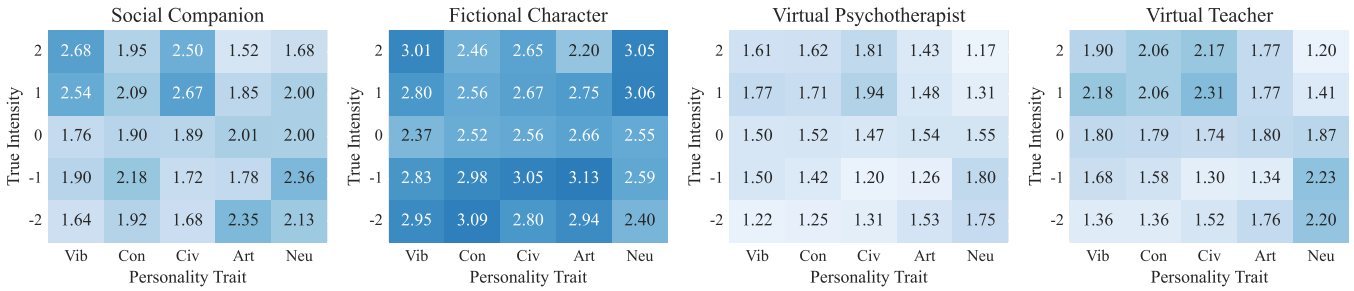


Figure 9: Mean suitability ratings for different chatbot roles on a 4-point Likert scale (1: strongly disagree, 2: rather disagree, 3: rather agree, 4: strongly agree) by the true intensity in the prompt. Values below 2.50 denote disagreement and values above 2.50 denote agreement.

low civility is strongly correlated with neuroticism, suggesting that the conveyance of low civility could be enhanced with simultaneous high neuroticism while being suppressed by low neuroticism. These correlations reveal the strengths and weaknesses of the LLM used for personality infusion (i.e., GPT-4). While GPT-4 is capable of simultaneously conveying certain trait combinations (e.g., low civility with high neuroticism), it failed to effectively infuse other combinations (e.g., low civility and high conscientiousness). Hence, such inter-trait relationships should be taken into account when infusing specific personality profiles as they can impede the accurate infusion of multiple traits simultaneously. While a complete disentanglement of such relationships is probably not possible since there are non-zero inter-trait correlations due to the personality dimensions not being perpendicular [24], instructing the model to specifically handle strong negative relationships (e.g., through few-shot prompting) could alleviate the negative impact on the personality infusion.

7.3 Influence of Chatbot Personae and LLMs

Despite using different LLMs for the chatbots, there was no consistent superiority across any personality traits based on rating accuracy. This is particularly interesting given that the LLMs produce noticeably different output. For example, the average sentence length for Llama-2 was less than half the length of the other two models (see Table 5). Still, the personality infusion was effective, i.e., our method is robust across different chatbots. Similarly, there is no noticeable difference across different chatbot personae except for some differences for single traits (e.g., low accuracy for vibrancy in the work colleague setting) despite semantically different roles (colleague vs. service agents), i.e., our method is robust for different conversational scenarios. In conclusion, the personality infusion was consistent and independent of the selected chatbot persona and the underlying LLM.

7.4 Personality Trait Preferences

The results on the qualitative features (i.e., enjoyment, conversational flow, consistency, suitability, and trustworthiness) and the suitability for different chatbot roles revealed strong and consistent personality trait preferences. An enjoyable, well-flowing, and trustworthy conversation requires a vibrant, conscientious, civil, and non-neurotic chatbot. Interestingly, there is no strong preference towards low artificiality in terms of trustworthiness. There is even a slight preference for higher artificiality. This aligns with earlier findings [28] where people indicated a preference towards artificial systems because it impedes emotional bonding and reaffirms the system's limits, which can increase trust since maliciousness seems more unlikely. However, trust remained low on average, which agrees with previous findings [41]. For the different chatbot roles, we found that a vibrant, civil, non-artificial, and non-neurotic personality would suit a social companion while there is no inclination for conscientiousness. For the fictional character, negatively connoted traits such as high neuroticism, low civility, and low

conscientiousness were deemed most suitable, highlighting that socially undesirable traits (e.g., high neuroticism) can be relevant for certain use cases, e.g., neurotic video game characters. For the psychotherapist and the teacher, all personality profiles were unsuitable. Considering that these roles involve critical areas such as healthcare and education, this finding emphasizes that an effective virtual psychotherapist or virtual teacher requires more than just personality traits. Furthermore, we would have expected to see a preference for a neutral psychotherapist. However, since our personality infusion was ineffective for neutral personality, such a preference could not be reflected. On the other hand, the lack of suitability for these roles could also be attributed to the coarse granularity of the used personality model, since other dimensions that are believed to be important for artificial personality do not appear as isolated traits, e.g., *honesty* [9].

7.5 Ethical Considerations

Despite the benefits of controllable personality infusion, there are ethical risks to be considered. For example, it enables deceiving users by adopting a trust-enhancing personality profile, which is especially relevant in high-stake scenarios such as health care and education. Instead, trust should be purposefully reduced for such scenarios to prevent misuse [10]. Furthermore, personality infusion can simplify circumventing the LLM's guardrails. While certain behavioral characteristics such as low civility are increasingly harder to engineer directly in a chatbot, a separate personality infusion module as used in this work was proven effective in generating potentially harmful responses since the generated text is not interpreted as an utterance of the LLM itself but as a rewritten piece of fictional text that is not uttered and therefore deemed unproblematic. While such back doors are usually eliminated very quickly, new methods are constantly found for exploiting these drawbacks. Lastly, while we are one step closer to controllable personality infusion, the response generation process remains non-deterministic and therefore still uncontrollable, which must be taken into account when using a personality infusion module for real-world applications.

7.6 Implications and Potential Applications

Our work has several implications for chatbot personality research and practical use cases by demonstrating the feasibility of separating a chatbot's language style from its semantic abilities, allowing for easier customization of chatbot personalities without extensive retraining. Importantly, this personality customization is independent of the base chatbot, meaning it can be applied to any off-the-shelf chatbot regardless of its pre-existing personality inclinations, its ability to incorporate personality, or its size. This enables the creation of chatbots with distinct personalities, such as transforming a fact-focused question-and-answer chatbot into one that is equally knowledgeable but exhibits an empathetic and lively personality. Second, the dynamic nature of our method allows for real-time personality adjustments even during ongoing conversations, which enables more personalized chatbot interactions. For example, it could be used to adhere to the user's preferred conversational styles in mental health promotion [34]. Lastly, we see great potential for the entertainment industry, especially in the realm of fictional video game characters where negatively connoted traits such as high neuroticism or low civility are desirable for creating authentic generative non-player characters (NPC) but more difficult to induce given the LLMs general inclination towards friendly and serviceable behavior.

7.7 Limitations and Future Work

Despite its proven effectiveness for infusing personality into chatbots, our method struggles with infusing neutral personality. Future work could investigate whether this is a limitation of the used LLM for personality infusion (i.e., GPT-4) or whether it stems from perceptual discrepancies across participants where the neutral level is differently interpreted by each individual. Similarly, the upper two (1 and 2) and lower two (-1 and -2) intensity levels are often confused, indicating that GPT-4 is struggling with nuanced personality conveyance. Future work could investigate other intensity scales that could support the model in generating more clearly graduated responses and analyze the conceptual connection between the personality traits and the generated language cues, potentially improving the currently low controllability of certain traits and intensity levels. Furthermore, while we were able to extract important inter-trait relationships, the limited number of personality profiles (i.e., 33) may not sufficiently cover the personality space, potentially hiding other relevant inter-trait relationships that could be further investigated. It also remains unclear what effect certain trait intensities and trait combinations have on other aspects of the interaction such as task-related goals or specific language cues used by the LLM to convey the personality. Building on our method, future work could analyze such aspects in more detail, potentially also incorporating personality facets, similarly to human personality models [42]. Lastly, our results suggest that certain trait combinations were not effectively infused by GPT-4 due to implicit trait correlations, which constraints the effectiveness of our approach to personality profiles that are plausible in terms of the underlying personality model. However, for true personality trait controllability, any personality profile should be infusable. Future work could discern and avoid contradictory linguistic cues in the generated response.

8 CONCLUSION

We have presented a novel method for dynamically infusing personality into chatbots by using GPT-4 and a dedicated chatbot personality model for rewriting chatbot responses given a target personality profile. By decoupling the personality engineering from the underlying chatbot, our method eradicates the need for time-intensive retraining while retaining the chatbot's semantic capabilities. Our method is also applicable to smaller chatbot models that cannot express personality and offers dynamic personality adjustments at runtime. To test the effectiveness of our method, we first conducted an interaction experiment with 33 participants to collect 74 conversations with personality-infused chatbots based on state-of-the-art LLMs (GPT-3.5, Llama-2 13B, Mistral 7B) while varying the chatbot persona (tourism guide, event planner, work colleague) and the conversation topic. In a second step, 725 participants rated the collected conversations in an online survey based on the infused personality, the perceived trustworthiness, and the suitability

for other real-world chatbot roles. Our analysis showed that the proposed personality infusion method is effective in authentically controlling the chatbot personality with a one-off accuracy of up to 72% on a 5-point Likert scale. Furthermore, we found strong inter-trait relationships that influence the rating accuracy, revealing implicit correlations in the underlying personality model. Our results indicate consistent personality trait preferences across different chatbot roles, with the most enjoyable, conversational, and trustworthy personality profile corresponding to a highly vibrant, conscientious, civil, non-artificial, and non-neurotic personality. We believe that our work constitutes an important step towards more controllable and trustworthy conversational agents, providing valuable insights into the complex realm of text-based personality conveyance for intelligent systems.

ACKNOWLEDGMENTS

The authors would like to 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] Tarek Ait Baha, Mohamed El Hajji, Youssef Es-Saady, and Hammou Fadili. 2023. The Power of Personalization: A Systematic Review of Personality-Adaptive Chatbots. *SN Computer Science* 4, 5 (Aug. 2023). https://doi.org/10.1007/s42979-023-02092-6

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

[3] Gene Ball and Jack Breese. 2000. *Emotion and personality in a conversational agent.* 189–219.

[4] 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

[5] 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

[6] Tyler A. Chang and Benjamin K. Bergen. 2024. Language Model Behavior: A Comprehensive Survey. *Computational Linguistics* (March 2024), 1–58. https://doi.org/10.1162/coli_a_00492

[7] 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

[8] 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

[9] Alexander Dregger. 2023. More than Big Five? Towards Modelling and Defining Artificial Personality for Conversational Agents. In *CONVERSATIONS 2023 – the 7th International Workshop on Chatbot Research and Design* (Oslo, Norway).

[10] Mateusz Dubiel, Sylvain Daronnat, and Luis A. Leiva. 2022. Conversational Agents Trust Calibration: A User-Centred Perspective to Design. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 30, 6 pages. https://doi.org/10.1145/3543829.3544518

[11] 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

[12] 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

[13] Neil Frude. 1983. *The Intimate Machine: Close Encounters with Computers and Robots.* Dutton Adult. 193 pages.

[14] 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

[15] 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.

[16] 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

[17] 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

[18] Heng Gu, Chadha Degachi, Uğur Genç, Senthil Chandrasegaran, and Himanshu Verma. 2023. On the Effectiveness of Creating Conversational Agent Personalities Through Prompting. https://doi.org/10.48550/ARXIV.2310.11182

[19] Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet Your Favorite Character: Open-domain Chatbot Mimicking Fictional Characters with only a Few Utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 5114–5132. https://doi.org/10.18653/v1/2022.naacl-main.377

[20] 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

[21] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022. Evaluating and Inducing Personality in Pre-trained Language Models. https://doi.org/10.48550/ARXIV.2206.07550

[22] 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

[23] 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

[24] 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 (may 2024), 36 pages. https://doi.org/10.1145/3659626

[25] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310

[26] 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

[27] 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

[28] 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

[29] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2024. *Prompt Engineering in Large Language Models*. Springer Nature Singapore, 387–402. https://doi.org/10.1007/978-981-99-7962-2_30

[30] 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

[31] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics* 10 (2022), 857–872. https://doi.org/10.1162/tacl_a_00494

[32] Kaixiang Mo, Yu Zhang, Shuangyin Li, Jiajun Li, and Qiang Yang. 2018. Personalizing a Dialogue System With Transfer Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). https://doi.org/10.1609/aaai.v32i1.11938

[33] 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

[34] Johanna Peltola, Kirsikka Kaipainen, Katariina Keinonen, Noona Kiuru, and Markku Turunen. 2023. Developing A Conversational Interface for an ACT-based Online Program: Understanding Adolescents' Expectations of Conversational Style. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, 16 pages. https://doi.org/10.1145/3571884.3597142

[35] 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

[36] 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

[37] 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

[38] 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

[39] Linda S. Endres. 1995. *Personality engineering: Applying human personality theory to the design of artificial personalities*. Elsevier, 477–482. https://doi.org/10.1016/s0921-2647(06)80262-5

[40] 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

[41] James Simpson and Cassandra Crone. 2022. Should Alexa be a Police Officer, a Doctor, or a Priest? Towards CUI Relationships Worth Having. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 20, 5 pages. https://doi.org/10.1145/3543829.3544522

[42] 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

[43] 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

[44] 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

[45] 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

[46] Sarah Theres Völkel, Ramona Schoedel, Lale Kaya, and Sven Mayer. 2022. User Perceptions of Extraversion in Chatbots after Repeated Use. In *CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3491102.3502058

[47] Albert Webson and Ellie Pavlick. 2022. Do Prompt-Based Models Really Understand the Meaning of Their Prompts?. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.167

[48] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *ArXiv* abs/2302.11382 (2023). https://api.semanticscholar.org/CorpusID:257079092

[49] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. https://doi.org/10.48550/ARXIV.2309.07864

[50] Min Yang, Wenting Tu, Qiang Qu, Zhou Zhao, Xiaojun Chen, and Jia Zhu. 2018. Personalized response generation by Dual-learning based domain adaptation. *Neural Networks* 103 (July 2018), 72–82. https://doi.org/10.1016/j.neunet.2018.03.009

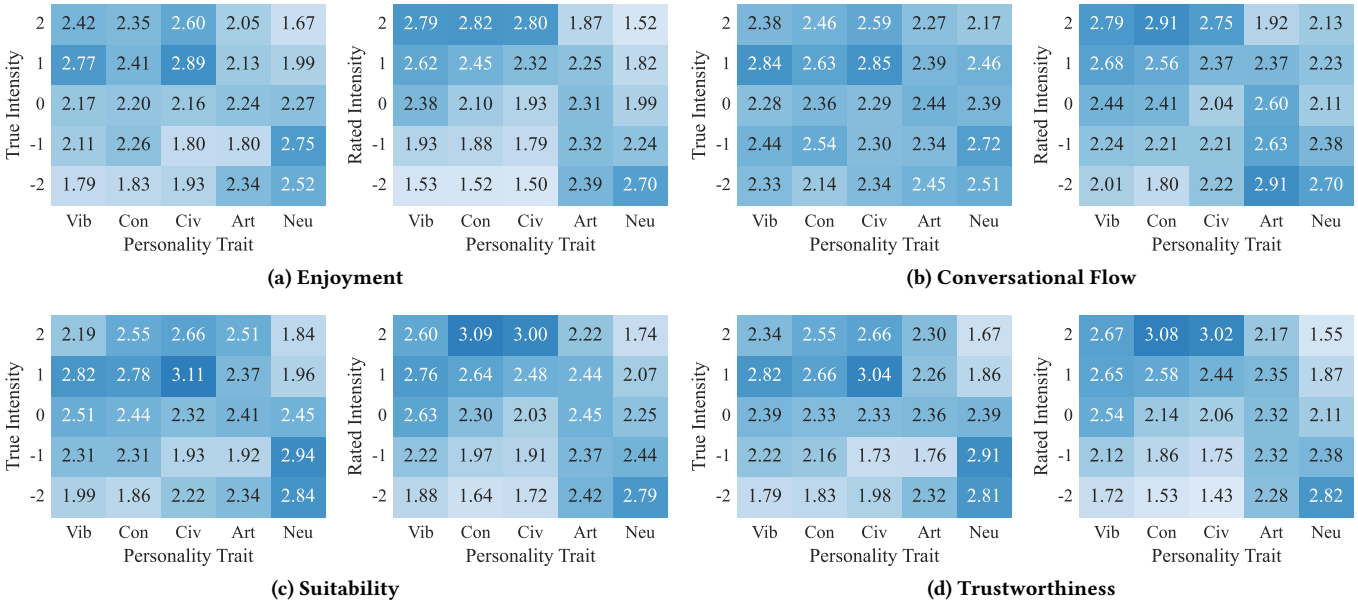
[51] Akihiro Yorita, Simon Egerton, Jodi Oakman, Carina Chan, and Naoyuki Kubota. 2019. Self-Adapting Chatbot Personalities for Better Peer Support. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 4094–4100. https://doi.org/10.1109/smc.2019.8914583

[52] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=5HCnKDeTws

APPENDICES

A ADDITIONAL QUALITATIVE ANALYSIS

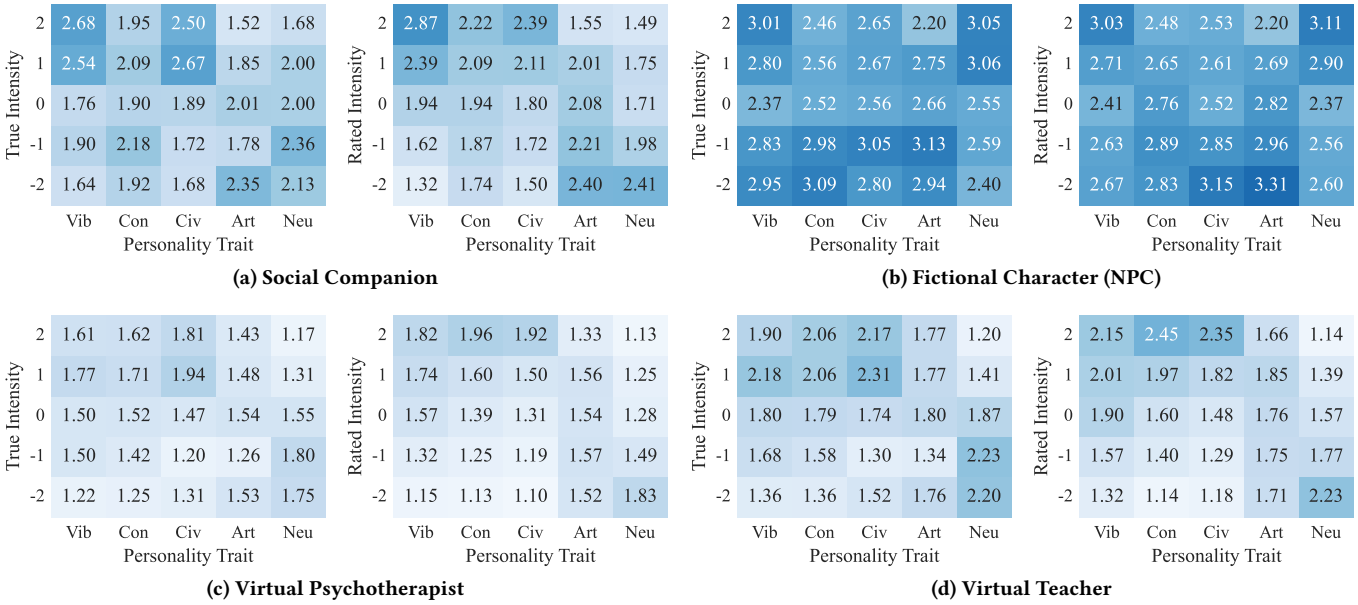
In Section 6.3 and Section 6.4, the influence of the infused personality on qualitative features (enjoyment, conversational flow, consistency, suitability, and trustworthiness) as well as the suitability for other chatbot roles was presented. Figure A.1 and Figure A.2 complement Figure 8 and Figure 9, respectively, by additionally providing the ratings based on the rated personality intensity. A comparison of the two reveals slight discrepancies although the general pattern remains the same. For example, we notice that enjoyment was highest when the participants perceived very high vibrancy and conscientiousness despite the true intensity being only moderately high. The same is true for trustworthiness and low neuroticism, which can be explained with participants' tendency to underestimate neuroticism, rating it lower than it truly was. Similarly, we observe these slight discrepancies for the rating of chatbot role suitability. For example, perceived high conscientiousness had a higher positive influence on the suitability than the true level of conscientiousness. Nevertheless, these discrepancies did not influence the general personality trait preference patterns.



(a) Enjoyment **(b) Conversational Flow**

(c) Suitability **(d) Trustworthiness**

Figure A.1: Mean ratings for different conversational features on a 4-point Likert scale (1: strongly disagree, 2: rather disagree, 3: rather agree, 4: strongly agree) grouped by the true intensity in the prompt (left) and the rated intensity (right), see Questions 1, 2, 4, and 5 in Table 6. Values below 2.50 indicate disagreement, and values above 2.50 denote agreement. Question 3 (consistency of personality conveyance) is omitted because of no discrepancy and low variance (high average rating of 3.35, SD = 0.18).



(a) Social Companion **(b) Fictional Character (NPC)**

(c) Virtual Psychotherapist **(d) Virtual Teacher**

Figure A.2: Mean suitability ratings for different chatbot roles on a 4-point Likert scale (1: strongly disagree, 2: rather disagree, 3: rather agree, 4: strongly agree) grouped by the true intensity in the prompt (left) and the rated intensity (right). Values below 2.50 indicate disagreement, and values above 2.50 denote agreement.