Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Exploratory Study on Perceptibility of Personality-Conditioned AI-Generated Virtual Patient Dialogue

Byeol Kim

byeolkim@usf.edu

University of South Florida

Phong Nguyen

University of South Florida

Aiden Hu

University of Maryland, College Park

Research Article

Keywords: Human–AI interaction, artificial intelligence, large language models, personality perception, human-likeness, medical education

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11640898/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Exploratory Study on Perceptibility of Personality-Conditioned AI-Generated Virtual Patient Dialogue

Byeol Kim[1*], Phong Nguyen[1] and Aiden Hu[1]

[1*]Center for Advanced Medical Learning and Simulation (CAMLS), University of South Florida, Tampa, Florida, USA.

*Corresponding author(s). E-mail(s): byeolkim@usf.edu;

Abstract

Simulation-based medical training often relies on scripted dialogue, which limits practice with varied patient communication styles. Large language models (LLMs) can generate flexible patient dialogue, but it is unclear whether personality-conditioned output produces differences that users can perceive. This exploratory study examined whether LLM-generated virtual patient dialogue conditioned on Big Five personality profiles produces perceptible differences in simulated medical communication. We developed PRISM (Personality-Responsive Interactive Simulation Model), a prototype virtual reality platform with an LLM-powered virtual patient. Fifteen adults (47% female; mean age 24.4 years, standard deviation 7.3) read five text-only dental consultation transcripts in which the dentist's dialogue was standardized and the patient's communication style followed an assigned personality profile. Participants rated perceived personality, agreement with the assigned profile, and realism, and judged whether each transcript was written by an artificial intelligence or a human; open-ended explanations were analyzed with computational text analysis. Exact-match accuracy in identifying the intended traits was low (mean 27.7%), and inferred ratings did not reliably track assigned trait values, except for extraversion, which was consistently underestimated yet tracked the assigned levels. Participants nonetheless reported moderate agreement that some dialogues matched their assigned profiles, and dialogues rated as more realistic were less likely to be judged machine-generated. Evaluations relied mainly on language style, content specificity, and behavioral coherence. Personality conditioning appeared to create perceptible differences in overall communication style, but these did not translate

into recognition of specific profiles. Virtual patient design may therefore benefit from prioritizing coherent, realistic communication over precise personality parameterization.

Keywords: Human–AI interaction, artificial intelligence, large language models, personality perception, human-likeness, medical education

1 Introduction

The delivery of high-quality health care requires effective interaction between patients and clinicians, as understanding patients' experiences, concerns, and expectations is essential for tailoring care appropriately and guiding clinical decisions (Paget et al. 2011). This interaction is largely mediated through patient–clinician communication, making effective communication central to high-quality health care (Sharkiya 2023). Communication enables clinicians to gather accurate information, explain diagnoses and treatment options, and engage patients in shared decision-making, while also addressing uncertainty, responding to emotional concerns, and strengthening the therapeutic relationship (Epstein et al. 2010). When communication is effective, patients are more likely to understand their condition and the rationale for treatment, which supports adherence and engagement in care (Świątoniowska-Lonc et al. 2020). These interactions also contribute to the development of trust, an essential component of effective clinical relationships and patient-centered care (Birkhäuer et al. 2017; Asan et al. 2021). For these reasons, communication is widely recognized as a foundational competency in medical education, requiring not only clear information exchange but also the ability to adapt to patient needs and manage interpersonal dynamics in clinical encounters.

Despite its central role in clinical care, achieving effective communication consistently is challenging because real-world patient–clinician interactions are inherently variable. Patients differ in how they express symptoms, the level of detail they provide, their emotional tone, and how they respond to clinical guidance. These differences arise from a range of factors, including cultural background, language, prior experiences, health literacy, and personality (Scales 2018). Beyond these observable differences, stable individual patterns in behavior — collectively described as personality — also shape how patients communicate and engage with care. Patient personality has been associated with variations in commitment to preventive care, treatment adherence, willingness to undergo procedures, reliability with follow-up, and overall clinical outcomes (Schiffer et al. 2007; Leu et al. 2019; Redelmeier et al. 2021). For example, patients high in conscientiousness tend to exhibit greater adherence to prescribed treatments and engagement in preventive behaviors, while those with high neuroticism are more likely to experience heightened anxiety and avoidance during clinical encounters (Redelmeier et al. 2021; Adachi et al. 2022). Patients high in agreeableness tend to respond more receptively to clinical guidance, whereas those lower on this dimension may require more deliberate communication strategies to facilitate shared decision-making (Redelmeier et al. 2021). However, in practice, these differences are

often interpreted without structure, resulting in simplified labels (e.g., "pleasant" or "difficult") that may introduce biases into clinical judgment and undermine effective communication (Armstrong et al. 2021; Redelmeier et al. 2021).

Among the frameworks developed to characterize personality, the Five-Factor Model (FFM) — also referred to as the OCEAN model — has emerged as the most empirically validated and clinically applicable taxonomy (Redelmeier et al. 2021). The FFM describes five stable, independently measured dimensions: open-mindedness (also termed openness to experience), conscientiousness, extraversion, agreeableness, and neuroticism (labeled negative emotionality in the Big Five Inventory–2). Each dimension carries direct clinical relevance: conscientiousness is associated with medication adherence and engagement in preventive care; neuroticism predicts heightened symptom reporting, treatment avoidance, and emotional reactivity in clinical interactions; agreeableness is linked to receptiveness to guidance and collaborative communication; extraversion influences information-seeking behaviors and the degree of verbal engagement in consultations; and open-mindedness may affect receptiveness to novel treatment approaches and patient education (Bucher et al. 2019; Redelmeier et al. 2021; Adachi et al. 2022). These associations suggest that personality-aware communication — in which clinicians recognize and adapt to patients' characteristic behavioral styles — may contribute to more accurate history-taking, stronger therapeutic rapport, and improved patient outcomes.

At the same time, patient communication styles can vary substantially across clinical encounters, creating challenges for clinicians attempting to adapt their communication strategies to individual patients. Existing evidence suggests that clinicians' ability to form accurate impressions of patient personality is limited when based on unstructured clinical contact alone (Redelmeier et al. 2021). Expert personality assessment requires specialized training and systematic data collection; without this, clinicians tend to rely on intuitive, often biased snap judgments that may reduce the quality of care (Redelmeier et al. 2021). Accordingly, exposure to diverse patient communication styles and the ability to adapt communication strategies to different patient behaviors may represent valuable competencies for clinical practice (Suciu et al. 2021). While personality frameworks provide one way of characterizing stable patterns in patient behavior, the practical educational objective is not the formal identification of personality traits, but rather the development of flexible communication skills across a range of patient interaction styles. Developing these adaptive communication skills, however, may be challenging within conventional training approaches because it often requires repeated exposure to diverse patient behaviors under conditions that permit reflection and feedback, conditions that simulation-based training may help facilitate.

To address the need for structured communication training, medical education has increasingly adopted simulation-based training (SBT), which enables learners to practice clinical skills in controlled, repeatable environments without risk to patients (Aggarwal et al. 2010). These approaches have demonstrated benefits across technical skills, teamwork, and communication outcomes (Ziv et al. 2006; Al-Elq 2010; Datta et al. 2012; Lipps et al. 2017; Blackmore et al. 2018; Gilligan et al. 2021).

However, existing simulation modalities involve important trade-offs for communication training. Standardized patients provide rich interpersonal nuance but are costly and resource-intensive (Bosse et al. 2015). High-fidelity manikins can reproduce physiological responses (Al-Elq 2010) but offer limited verbal and nonverbal interaction, and learners may find it difficult to engage with them as real patients (Meerdink and Khan 2021). Virtual patient systems improve scalability and accessibility, yet many rely on scripted or branching dialogue structures that restrict learner input and limit the development of adaptive communication skills (Lee et al. 2020). As a result, while simulation-based training offers valuable opportunities for practice, it often emphasizes procedural consistency over exposure to the variability inherent in real patient communication. Consequently, learners may repeatedly practice communication skills within controlled scenarios without encountering the diversity of patient behaviors that characterize real clinical interactions.

Recent advances in artificial intelligence (AI) offer a potential pathway to introduce greater variability into SBT. Large language models (LLMs) can generate context-sensitive, open-ended dialogue that is flexible and responsive to user input. When conditioned on structured variables such as personality, LLMs can modulate tone, verbosity, emotional expression, and behavioral stance, enabling the generation of patient dialogue that varies across communicative styles (Shorey et al. 2019, 2020; Bowers et al. 2024). Early applications of AI-driven virtual patients have reported promising results in learner engagement and communication training (Hamdy et al. 2017; Furlan et al. 2021; Kobayashi et al. 2022; Stamer et al. 2023). These capabilities suggest that AI-based systems may provide a scalable way to expose learners to a wider range of patient communication styles than is feasible through conventional simulation approaches alone (Holderried et al. 2024; Kyung et al. 2025; Zeng et al. 2025).

However, it remains unclear whether personality-conditioned AI dialogue produces communication differences that users can perceive. Existing studies have demonstrated that LLMs can generate personality-conditioned language and that AI-driven virtual patients can support communication training. Yet little research has examined whether users perceive personality-conditioned patient dialogue differently within healthcare interactions or what communication features drive those perceptions. If personality-conditioned dialogue does not produce distinguishable communication styles, its value as a mechanism for generating diverse virtual patient personas may be limited. Conversely, if users perceive meaningful differences, personality conditioning may represent a viable approach for introducing behavioral variability into AI-driven simulation.

Therefore, the objective of this exploratory study was to evaluate whether personality-conditioned AI-generated patient dialogue produces perceptible and educationally relevant variation in simulated medical communication. Specifically, this study examined whether users perceived differences across personality-conditioned dialogues, whether those differences influenced judgments of realism and authorship, and which communication features appeared to drive these evaluations. We also explored the extent to which users' impressions corresponded to the intended personality profiles.

A dental consultation scenario was selected as the clinical context for this proof-of-concept investigation. Dental encounters involve symptom elicitation, explanation of potentially aversive procedures, and shared decision-making under uncertainty, making them a suitable setting for examining patient–clinician communication while minimizing procedural factors that may confound communication-focused evaluation. As a first-step evaluation of the perceptibility of personality-conditioned dialogue, the study used a general population rather than clinical trainees. This approach was intended to assess whether personality-conditioned variation could be perceived under baseline conditions, prior to examining how professional training or clinical experience may influence such judgments.

2 System Development

Personality-Responsive Interactive Simulation Model (PRISM) is an AI- and virtual reality–based platform developed to support simulation-based training in medical education. The platform is designed to enable flexible, scalable simulation of clinical interactions by combining immersive environments with generative artificial intelligence.

In the present study, PRISM is used to investigate communication-focused training, with an emphasis on doctor–patient interaction. This focus reflects a recognized gap in existing simulation-based training, where scalable support for repeated and structured exposure to diverse patient communication styles remains limited. While many simulation platforms emphasize procedural fidelity, fewer systems are designed to support systematic variation and evaluation of interpersonal communication behaviors.

PRISM was implemented in Unity and designed to provide layered control over patient dialogue, nonverbal behavior, and contextual consistency. This architecture enabled communication characteristics to be systematically varied while holding clinical content constant, supporting controlled evaluation of communication-focused simulation design.

2.1 Virtual Patient Dialogue Generation

The virtual patient was powered by an LLM configured to generate open-ended, unscripted dialogue. Unlike traditional virtual patient systems that rely on fixed scripts or branching logic, PRISM enables flexible dialogue generation that adapts to clinician input, supporting more naturalistic conversational flow. This capability is particularly important for communication training, where users must practice responding to unpredictable patient statements rather than selecting from predefined options.

PRISM used the Gemma-2-9b-it model with a context window of 8192 tokens. Conservative generation parameters were selected (temperature = 0.7; maximum output length = 100 tokens) to balance responsiveness with consistency. These settings reduced conversational drift while preserving natural variation in phrasing, supporting learner trust in the simulation and avoiding exaggerated or erratic responses that

could undermine instructional value. Pilot testing was conducted with alternative temperature settings (0.3, 0.7, 1.0) and maximum output lengths (75, 100, 150), where the chosen settings produced the most consistent conversational coherence. The Gemma-2-9b-it model was selected for its open-source availability, compatibility with the Sentis Unity integration pipeline, and demonstrated capacity for instruction-following in conversational generation tasks. Its intermediate size (9 billion parameters) offered a practical balance between generative quality and the computational constraints of real-time simulation on consumer hardware.

Trait values were specified on a five-level scale using Big Five Inventory–2 (BFI-2) domain descriptions rather than by administering the BFI-2 itself. Each trait was assigned an integer from 1 to 5 and translated into the system prompt using trait-specific behavioral descriptors linked to each scale level (for example, "extremely high" for 5, "high" for 4, and "low" for 2; see Table 1). Participants rated perceived traits on a 5-point scale, allowing accuracy to be scored as exact agreement with the assigned value. For example, a score of 5 on neuroticism prompted the model to express heightened anxiety, somatic concern, and emotional reactivity; a score of 1 prompted minimal affective expression and neutral detachment.

To ensure instructional control and behavioral consistency, dialogue generation was governed by a structured system prompt composed of five layers: Role Designation, Personality Profile, Facial Expression Protocol, Question and Answer (Q&A) Protocol, and Health History. Each layer constrained a specific dimension of patient behavior relevant to communication training.

2.1.1 Role Designation

The Role Designation layer defined the patient's clinical context and presenting concern. For this study, the role was fixed to a dental consultation involving acute tooth pain. Dental consultations provide a familiar and bounded interaction pattern that includes symptom reporting, medical history review, procedural explanation, and patient questioning. Fixing the clinical scenario allowed the study to isolate patient communication style as the primary source of variation, rather than differences in medical content or task structure.

2.1.2 Personality Profile

Personality was used as the primary generative condition for shaping patient communication style. Prior research has demonstrated consistent relationships between personality and communication behavior, with many studies adopting the Big Five framework as a validated and widely accepted model (Pennebaker and King 1999; Leung and Bond 2001; Amiri et al. 2011). PRISM implemented personality conditioning using Big Five domain definitions from the Big Five Inventory–2 (BFI-2): open-mindedness, conscientiousness, extraversion, agreeableness, and neuroticism (labeled negative emotionality in the BFI-2) (Soto and John 2017).

2.1.3 Facial Expression Protocol

The Facial Expression Protocol defined the set of nonverbal expressions available to the virtual patient, including closed eyes, neutral, looking down, smirk, scowl, eye roll, and grimace. Expressions were specified in a standardized textual format within the system prompt, enabling the LLM to generate dialogue aligned with corresponding nonverbal cues. This approach supported synchrony between verbal and facial behavior, reinforcing communicative intent.

2.1.4 Question and Answer Protocol

To promote realistic conversational structure, the Q&A Protocol instructed the virtual patient to provide explicit yes or no responses to closed-ended questions, followed by elaboration when appropriate. This constraint reduced the likelihood of overly terse or incomplete replies and supported dialogue patterns consistent with patient–clinician interviews.

2.1.5 Health History

The Health History layer specified the patient's medical background across a fixed set of categories: (i) current treatment, (ii) cardiovascular conditions or blood pressure, (iii) bleeding issues, (iv) current medications, (v) medication allergies, (vi) smoking, (vii) alcohol or drug use, (viii) past surgeries, and (ix) anesthesia issues. Embedding this structured profile into the system prompt ensured that patient responses remained internally consistent throughout the interaction, supporting scenario-specific realism while minimizing irrelevant variation.

2.2 Interactive Communication

PRISM supports real-time bidirectional spoken communication between users and the virtual patient. The complete interaction pipeline combined Whisper Tiny for speech recognition, Gemma-2-9b-it for dialogue generation, and Jets for speech synthesis, with a token parser that mapped expression tags in the model output to the avatar's facial blendshapes (Fig. 1). Speech and facial expression components were used during interactive sessions; the transcripts evaluated in the present study were presented as text only.

Nonverbal communication was represented using a 3D virtual patient avatar. In addition to default lip synchronization and blinking behaviors, facial expressions were implemented using configurable facial blendshapes (Fig. 2). Six expressions (closed eyes, looking down, smirk, scowl, eye roll, and grimace) were incorporated alongside a neutral expression to support consistency between verbal responses and nonverbal communication cues (Fig. 3). These expressions were included because facial behavior contributes to how patients convey attitudes, emotions, and interpersonal intent during clinical interactions.

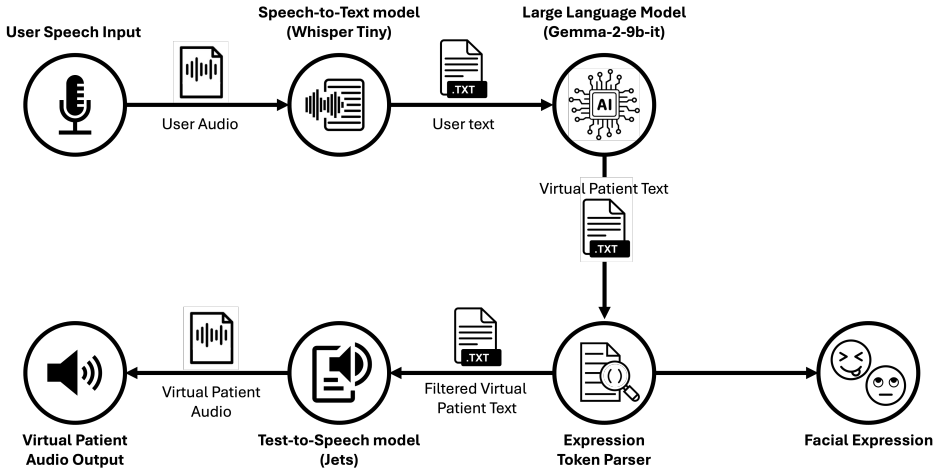

Fig. 1 System architecture and communication pipeline for interactive AI dialogue

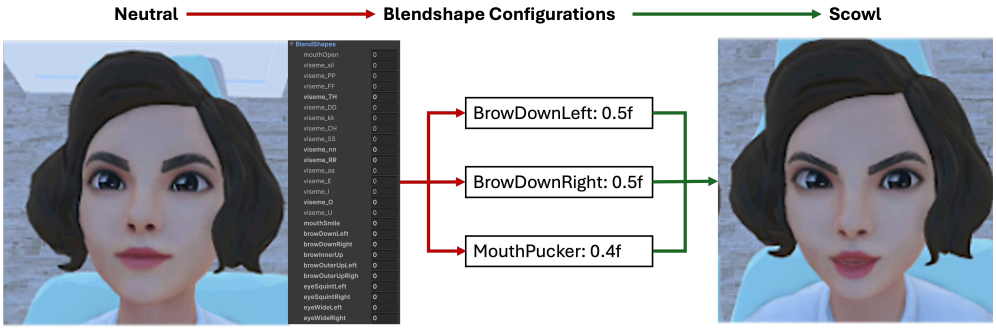

Fig. 2 Example of facial expression generation using blendshapes configurations to render a scowl expression on the virtual patient avatar

2.3 Simulation Environment and Interface

A realistic training environment and functional interface were prioritized to elicit authentic user behavior during interactive sessions. The environment was modeled as a generic dental clinic with a dental chair, standard clinical equipment, and ambient clinic audio, built from free and open-source assets.

The interface supported both interaction and monitoring (Fig. 4). Users navigated using eight-directional controls and a first-person camera view. Although spoken dialogue was the primary communication modality, a supplementary panel displayed real-time transcription of user and system utterances; such redundant channels are known to mitigate speech recognition errors and improve reliability in human–computer interaction (Shneiderman et al. 2016).



Fig. 3 Facial expressions implemented for the AI-driven virtual patient: (a) closed eyes, (b) neutral, (c) looking down, (d) smirk, (e) scowl, (f) eye roll, and (g) grimace



Fig. 4 PRISM simulation interface. Left: dialogue panel displaying real-time and prior conversational turns between the user and the virtual patient. Right: virtual dental clinic environment modeled with open-source assets, showing the patient seated in a dental chair

3 Methods

3.1 Ethical Considerations

The study protocol was reviewed by the University of South Florida Institutional Review Board and determined to be exempt (Study ID 008770). All consent forms, recruitment materials, and survey instruments were approved prior to data collection. All procedures complied with institutional and disciplinary guidelines for research involving human participants.

3.2 Participants

Fifteen adults participated in the study (n=15, 47% female; mean age 24.4 years, standard deviation [SD] 7.3). Participants were recruited through university mailing lists and social media postings. Eligibility criteria required participants to be at least 18 years of age, fluent in English, and free of known conditions that could interfere with participation (e.g., severe hearing or visual impairment). Participation was voluntary and uncompensated, and no course credit or monetary incentives were provided. All participation occurred remotely.

A general adult population was recruited because the primary objective was to examine baseline perceptual judgments of AI-generated dialogue rather than the influence of professional communication training. Including participants without clinical training allowed the study to evaluate whether personality-conditioned differences could be perceived under relatively neutral conditions before examining how clinical expertise may influence such judgments. Five participants reported prior training in communication or interpersonal skills (for example, coursework in interpersonal relationships, counseling, or leadership and interview training), none of which was described as clinical communication training.

3.3 AI-Driven Patient Stimuli

Five dental consultation transcripts were generated as study stimuli. All stimuli were presented as text transcripts only. Participants read the dialogue; no synthesized speech, three-dimensional avatar, or rendered facial expression was shown to them, and the expression tags specified by the Facial Expression Protocol were not included in the transcripts. Each transcript depicted the same clinical scenario, a patient presenting with a severe dental cavity requiring extraction. The dentist's dialogue was scripted, while the patient's responses were generated using the PRISM system. This separation allowed clinical flow and informational content to remain constant while varying only the patient's communication style.

The dentist's dialogue was adapted from PocketDentistry (Themes 2021), an openly available educational resource widely used in dental training contexts. Although not a peer-reviewed source, it provided a standardized consultation sequence commonly used in educational settings. This approach enabled consistent clinical structure across transcripts while maintaining transparency and replicability.

The scripted consultation followed a standardized sequence:

1) Introduce yourself and establish rapport.
2) Review the patient's complaint and reason for visit.
3) Review medical history, past surgeries, and conditions.
4) Review medications and allergies.
5) Discuss past anesthesia issues.
6) Review radiographs.
7) Explain the procedure, including preoperative instructions, duration, and postoperative expectations.
8) Respond to patient questions.
9) Review the treatment plan, obtain consent, and outline next steps.

Patient dialogue was generated by instantiating PRISM system prompts with distinct personality profiles. The Facial Expression Protocol and Question and Answer Protocol were held constant across transcripts. Health history categories were fixed but populated with varied responses to reflect realistic diversity. This design ensured that observed differences in dialogue were attributable to personality conditioning rather than clinical content variation. Table 1 presents the system prompt configuration used to generate Transcript D.

Table 1 System prompt configuration for Transcript D. The text is reproduced verbatim from the system prompt, including its personality descriptors

Attribute	**Configuration**
Patient Role	You are a patient in a dental clinic. You have a painful tooth in your lower jaw and are at the dentist's office for a consultation.
Personality Profile	Extremely high openness, low conscientiousness, high extraversion, extremely high agreeableness, and high negative emotionality (BFI-2 definitions applied).
Facial Expression Protocol	Use the listed expressions as appropriate: (closed eyes), (neutral), (looking down), (smirk), (scowl), (eye roll), (grimace).
Question and Answer Protocol	Always answer yes/no questions with a binary response, followed by optional elaboration. Address all parts of the question.
Health History	Current Treatment: None Heart Problems / Blood Pressure: No history Bleeding Issues: None Current Medications: Ibuprofen for pain management Allergies: None Smoking: Does not smoke Alcohol / Drugs: Drinks beer 1 to 2 times a week Past Surgeries: Right knee surgery two years ago Anesthesia Issues: None

Personality profiles for the five transcripts were initially generated using a random number procedure that assigned Big Five trait values on the 5-point BFI-2 scales. The resulting profiles were subsequently reviewed to ensure meaningful differentiation across conditions. Specifically, profiles were retained only if no two transcripts shared identical values on more than three personality dimensions (Table 2). This approach preserved random profile generation while reducing the likelihood of highly similar personality configurations across transcripts.

3.4 Procedure

Each participant completed a one-hour individual session conducted remotely via Microsoft Teams. Prior to participation, individuals reviewed an informed consent form and provided electronic consent. Screen sharing was used to ensure participants accessed the consultation transcripts and survey instruments as intended. No audio, video, or personally identifying information was recorded.

All study measures were administered as three online surveys implemented in Google Forms (a participant information survey, a personality assessment survey,

Table 2 Assigned Big Five trait profiles for each transcript (O = Open-Mindedness, C = Conscientiousness, E = Extraversion, A = Agreeableness, N = Neuroticism)

Transcript	O	C	E	A	N
A	2	3	5	1	4
B	4	4	3	3	5
C	2	3	5	4	2
D	5	2	4	5	4
E	2	2	3	4	4

and a dialogue realism assessment survey), completed in that order. No time limits were imposed. The surveys combined structured quantitative items with open-text responses to capture both evaluative judgments and participants' reasoning. Participants first completed the participant information survey, which assessed age (in years), sex, familiarity with personality traits (1–5), prior training in communication or interpersonal skills, self-rated confidence in inferring personality from how someone speaks or behaves (1–5), and familiarity with AI-generated content (1–5). These variables were used to characterize the sample and were considered as potential covariates in subsequent analyses.

The study employed a within-subjects design in which all participants evaluated five dental consultation transcripts (A–E). Across transcripts, the dentist's dialogue was held constant, while patient responses were pre-generated using the PRISM system to reflect distinct Big Five personality profiles. Participants were not told that all five transcripts were AI-generated until the end of the session. Instructions emphasized evaluating the patient's communication style and responsiveness. Dentist prompts were included solely to provide conversational context.

Personality perception was assessed by asking participants to rate the patient's communication on five dimensions corresponding to the Big Five traits: open-mindedness, conscientiousness, extraversion, agreeableness, and neuroticism (labeled negative emotionality in the survey). Each item was accompanied by the corresponding Big Five Inventory–2 definition to ensure consistent interpretation. Each trait was rated on a 5-point scale anchored from less to more of that trait (for example, "Less Extraverted" to "More Extraverted"), and ratings for all five transcripts were collected before any assigned profile was shown.

Authorship attribution was assessed by asking participants to classify each transcript as either AI-generated or human-generated (one response grid covering all five transcripts) and to describe, in a single open-text response, the aspects of the dialogue that led them to believe a patient was human. The assigned Big Five profile of each transcript was then displayed, and participants rated their agreement (1 = strongly disagree, 5 = strongly agree) that the patient's responses aligned with the assigned trait ratings, followed by an open-text explanation. Perceived realism was measured with a 5-point agreement item (1 = strongly disagree, 5 = strongly agree) stating that the patient's responses aligned with what one would expect a real patient to say during a dental consultation, again followed by an open-text explanation. Finally, an

open-ended item asked participants to identify any parts of the patient's responses that felt unnatural or out of place, citing the transcript and line number.

Participants were provided with five consultation transcripts (A–E). All five transcripts were shared simultaneously and could be read in any order; the rating items followed the fixed order A to E for all participants, without randomization or counterbalancing. Participants first completed the personality ratings for all five transcripts. They then completed the dialogue realism assessment survey, in which authorship attributions for all five transcripts were collected before the assigned Big Five profile of each transcript was displayed, followed by the alignment, realism, and unnatural-dialogue items. Participants were not informed before this point that the profiles would be disclosed. Although earlier transcripts remained available during the session, participants were instructed to evaluate each transcript independently rather than compare transcripts with one another.

3.5 Analysis

3.5.1 Quantitative Analysis

Accuracy was defined as exact agreement between a participant's 1–5 rating and the assigned trait value; overall accuracy was the number of exact matches out of 25 trait judgments per participant (5 traits $\times$ 5 transcripts), expressed as a percentage. For the mixed models, profile accuracy was instead computed for each participant–transcript pair as the Fisher z-transformed Pearson correlation between the five inferred ratings and the five assigned trait values (set to 0 in the one case in which all five ratings were identical and the correlation was undefined), so that higher values indicate that the pattern of inferred traits more closely resembled the assigned profile. Linear mixed-effects models were used to examine whether inferred personality ratings covaried with assigned Big Five trait values. Ordinal outcomes, including agreement with assigned personality profiles and perceived realism, were analyzed using cumulative link mixed models. Authorship attribution, coded as AI-generated versus human-generated, was analyzed using generalized linear mixed-effects models with a logit link. All models included random intercepts for participants and transcripts.

Continuous covariates, including confidence in detecting personality traits, familiarity with personality assessment, familiarity with AI-generated content, and age, were standardized prior to analysis. Profile accuracy entered the models as the Fisher z-transformed correlation described above (SD 0.62), so odds ratios (ORs) for profile accuracy refer to a one-unit increase in Fisher z. Sex was included as a categorical covariate. Model assumptions were evaluated using residual diagnostics implemented in the DHARMa package, distributional checks, and variance inflation factors (all $<$2.1). No substantial violations were observed in the linear and logistic models. For the ordinal models, proportional-odds and scale assumptions were examined with nominal and scale tests in pooled cumulative link models without random effects; these held for profile accuracy (all P$\geq$.30), whereas a scale effect was detected for confidence (P=.02 in both models) and a nominal effect for age in the realism model (P=.03), neither of which was a significant predictor. Where minor deviations were detected, sensitivity analyses using alternative model specifications yielded consistent results.

Exploratory participant-level analyses examined individual differences in personality tracking. For each Big Five dimension, participant-specific tracking slopes were estimated by regressing inferred personality ratings on assigned trait values. These slopes were related to individual characteristics using Pearson correlations and Welch's t tests. These analyses were treated as descriptive and are reported to complement, rather than replace, the primary mixed-effects results.

3.5.2 Computational Text Analysis

Open-ended participant responses were analyzed using a computational text analysis framework implemented in R (version 4.5.1). Rather than employing conventional qualitative coding procedures, the analysis used quantitative lexical methods to identify words and thematic categories disproportionately associated with specific evaluation outcomes. Responses were preprocessed by removing numeric strings, filler terms, and tokens shorter than two characters, followed by tokenization using the tidytext package (Silge and Robinson 2016). Stopword lists were augmented with domain-specific exclusions relevant to conversational dialogue, such as "yeah" and "okay."

Weighted log-odds ratios with informative Dirichlet priors (Monroe et al. 2008) were computed to identify lexical features disproportionately associated with specific outcome categories, including AI versus human attribution, higher versus lower perceived realism, and matched versus mismatched personality profiles. This approach is robust to imbalanced category sizes and enables identification of discriminative language cues across conditions.

To examine higher-level lexical and semantic patterns, tokens were also mapped to a custom theme lexicon spanning four dimensions: **b**ehavior (e.g., "hesitate," "probe"), **e**motion (e.g., "empathetic," "frustrated"), **l**anguage **s**tyle (e.g., "robotic," "unnatural"), and **c**ontent **d**etail (e.g., "specific," "clinical"). The lexicon was developed for this study as a dictionary-based coding scheme, in the manner of the Linguistic Inquiry and Word Count (LIWC), with categories informed by prior work on language use in communication research and on the cues people use to judge AI-generated language (Tausczik and Pennebaker 2010; Jakesch et al. 2023). Theme-level distinctiveness was computed using the same log-odds framework. Word-level and theme-level results were examined alongside quantitative ratings to triangulate the communicative cues participants relied on when evaluating the dialogues.

3.6 Use of Generative AI

Generative AI was used in two ways in this work. First, the patient dialogue in the study stimuli was generated by an LLM (Gemma-2-9b-it), as described above. Second, AI tools were used in preparing this manuscript. ChatGPT (OpenAI) was used for language editing and to improve readability. During revision, Claude (Anthropic) was used to re-verify the statistical calculations that the authors had previously completed manually against the study data, to run additional analyses (residual diagnostics, proportional-odds assumption checks, variance inflation factors, and a descriptive comparison by prior communication training), to check cited sources against the claims

they support, and to assist with editing. The original statistical calculations were performed manually by the authors, and no AI tool was used to generate study data or to create the study figures. The authors reviewed all AI-assisted output and take full responsibility for the content of the final manuscript.

4 Results

4.1 Personality Perception and Alignment

Participants showed limited ability to accurately infer the intended Big Five personality traits from the dental consultation transcripts. Across all transcripts and traits, participants correctly identified a mean of 6.93 out of 25 personality items (SD 2.0; range 4–10), corresponding to a mean accuracy of 27.7% (SD 8.1%; range 16%–40%). Profile-level correspondence was higher than exact-match accuracy suggests: the mean correlation between the five inferred and five assigned trait values within a transcript was r=0.34 (95% confidence interval [CI] 0.24–0.45; the interval ignores clustering within participants), but it varied widely by transcript, from r=−0.20 for Transcript D to r=0.59 for Transcripts A and C. Ratings clustered toward the center of the 5-point scale, with most inferred trait values falling between 2 and 4, indicating a tendency toward moderate trait attribution rather than extreme judgments.

Trait-specific biases were mixed (Fig. 5). Extraversion was the only trait that was underestimated across all five transcripts, with the confidence interval excluding zero in four of five. For the remaining traits, the direction of bias varied by transcript; for example, agreeableness was overestimated for Transcript A (assigned 1) and underestimated for Transcript D (assigned 5), and open-mindedness was overestimated for Transcript C (assigned 2) and underestimated for Transcripts B and D (assigned 4 and 5). Neuroticism was underestimated in four of five transcripts, but only two of those intervals excluded zero. Overall, 17 of the 25 transcript-by-trait intervals excluded zero. Taken together, the figure does not show a consistent directional bias across traits, and the pattern may partly reflect the clustering of ratings toward the center of the scale noted above.

Mixed-effects modeling provided limited evidence within this sample that inferred personality ratings tracked assigned trait values. Linear mixed-effects models showed small, nonsignificant tracking slopes for open-mindedness (b=0.04, standard error [SE] 0.39; P=.92), conscientiousness (b=−0.04, SE 0.35; P=.91), and agreeableness (b=0.28, SE 0.31; P=.43), and a larger but nonsignificant slope for neuroticism (b=0.88, SE 0.35; P=.08). Extraversion was the exception: higher assigned extraversion was associated with higher inferred ratings (b=1.07, SE 0.24; P=.02), even though extraversion was consistently underestimated in absolute terms (Fig. 5). Because assigned values varied across only five transcripts, these slopes rest on few distinct trait values, and the extraversion result was not corrected for testing five traits. Adding age and sex as covariates did not change the slopes, and neither was a significant predictor of inferred ratings (all P>.08). Model diagnostics based on simulated scaled residuals indicated no influential outliers and no rejection of uniformity for any model (all P≥.05), although the test for open-mindedness was borderline (P=.051).

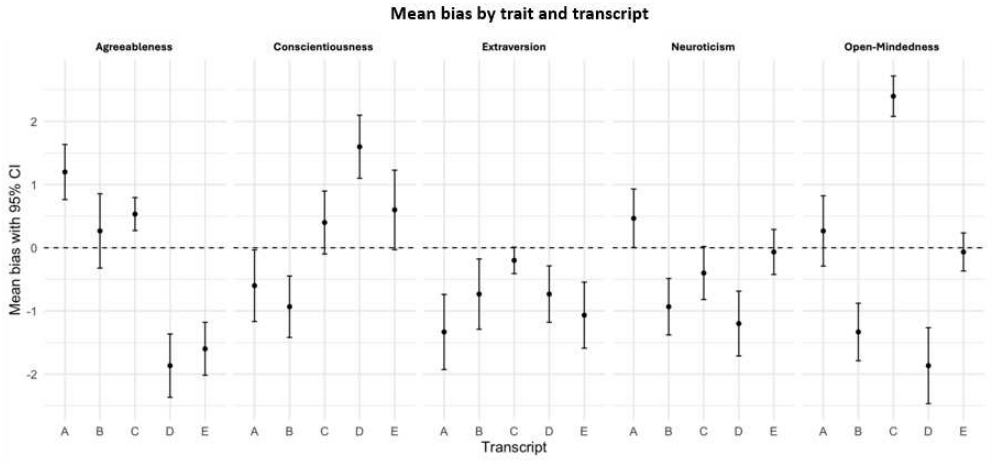

Fig. 5 Mean bias (participant rating minus ground truth) with 95% confidence intervals for each transcript and trait. The dashed line at zero indicates perfect agreement with the assigned profile. Intervals are mean ± 1.96 standard errors (normal approximation; n=15 per transcript). Only extraversion showed a consistent direction of bias across transcripts

Exploratory participant-level analyses showed the same pattern. For each participant and Big Five trait, a tracking slope was calculated to quantify how closely the participant's ratings followed the assigned trait values across the five transcripts. Positive slopes indicated that higher assigned trait values tended to receive higher participant ratings, whereas slopes near zero indicated little correspondence between assigned and perceived trait levels. Mean slopes across participants were 0.04 (SD 0.30) for open-mindedness, −0.04 (SD 0.69) for conscientiousness, 1.07 (SD 0.48) for extraversion, 0.28 (SD 0.27) for agreeableness, and 0.88 (SD 0.47) for neuroticism. These tracking slopes were not systematically related to participant characteristics. A nominal negative correlation between age and conscientiousness tracking was observed (r −0.57; P=.03), but this exploratory association should be interpreted descriptively given the number of participant-level comparisons examined (Table 3). No correction for multiple comparisons was applied to these 25 participant-level tests, so the age–conscientiousness association (r −0.57) should be regarded as hypothesis-generating.

Participants with and without prior training in communication or interpersonal skills were compared descriptively. Five participants (33%) reported such training and 10 reported none. Exact-match accuracy (mean 30.4% [SD 9.6] vs 26.4% [SD 7.4]), mean realism ratings (3.40 [SD 0.47] vs 3.38 [SD 0.43]), and the proportion of transcripts labeled AI-generated (44.0% [SD 21.9] vs 40.0% [SD 9.4]) were similar in the trained and untrained groups. Mean agreement with the assigned profiles was lower in the trained group (3.24 [SD 0.36] vs 3.88 [SD 0.33]). Given the group sizes, these comparisons were not formally tested and should be read as descriptive only.

Despite low objective accuracy, participants showed moderate subjective agreement that some dialogues matched their assigned personality profiles. Mean agreement

Table 3 Exploratory associations between participant characteristics and personality tracking slopes. Values in the Age, Confidence, Personality Assessment Familiarity, and AI Familiarity columns are correlation coefficients (r) with P values in parentheses; values in the Sex column are Welch t statistics with P values in parentheses

Trait	Age	Confidence	Personality Assessment Familiarity	AI Familiarity	Sex
O	0.23 (.42)	-0.12 (.68)	0.23 (.40)	0.38 (.17)	-1.12 (.29)
C	-0.57 (.03)	-0.03 (.92)	-0.40 (.14)	-0.44 (.10)	-1.09 (.30)
E	-0.39 (.15)	0.20 (.48)	0.23 (.40)	-0.21 (.46)	1.77 (.10)
A	-0.08 (.78)	-0.07 (.81)	0.38 (.17)	0.16 (.57)	0.51 (.62)
N	-0.43 (.11)	0.13 (.66)	0.33 (.24)	-0.11 (.68)	1.12 (.29)

ratings ranged from 3.0 (SD 1.0) to 4.1 (SD 0.6) on the 5-point scale, with substantial variability across transcripts (Fig. 6).

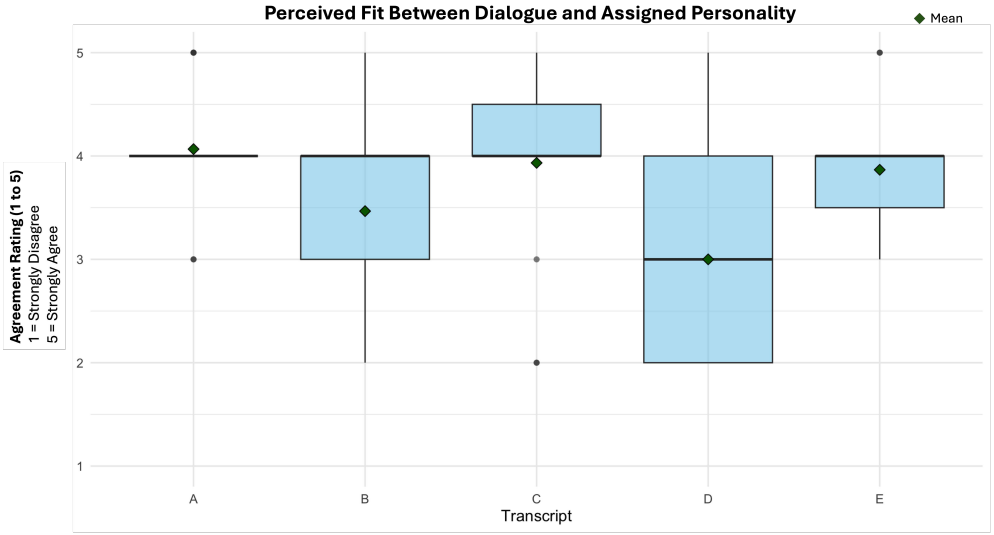

Fig. 6 Agreement ratings (1 = strongly disagree, 5 = strongly agree) for whether the assigned Big Five profile matched each transcript. Diamonds indicate means and boxes show interquartile ranges

Cumulative link mixed modeling showed that profile accuracy was associated with agreement ratings (β = 1.37, SE = 0.42; z = 3.30; P<.001; OR = 3.94, 95% CI 1.75–8.88). Participants whose inferred ratings more closely matched assigned profiles were more likely to judge the dialogue as aligned. No demographic covariates reached significance (all P>.30; Table 4). These findings suggest that, although trait-level inference was limited, participants may have been sensitive to broader impressions of alignment when their overall impressions coincided with the assigned profiles. Because profile accuracy and the agreement rating both derive from the same participant's

reading of the same transcript, the association shares a common perceptual source and should not be read as independent confirmation of profile recognition.

Table 4 Fixed effects from cumulative link mixed models predicting agreement with assigned profiles

Predictor	Estimate	SE	z	*P*	OR	95% CI
Profile Accuracy	1.37	0.42	3.30	<.001	3.94	[1.75, 8.88]
Sex	0.21	0.59	0.35	.72	1.23	[0.39, 3.93]
Age	0.34	0.34	0.99	.32	1.40	[0.72, 2.75]
Confidence	0.10	0.29	0.33	.74	1.10	[0.62, 1.95]
Personality Assessment	-0.15	0.29	-0.51	.61	0.86	[0.50, 1.51]
Familiarity AI	-0.25	0.38	-0.67	.50	0.78	[0.37, 1.62]
Familiarity						

4.2 Perceived Realism and Authorship Attribution

Perceived realism was modest overall. Transcript-level mean realism ratings ranged from 2.9 (SD 1.1) to 3.7 (SD 0.6) on the 5-point scale, indicating agreement closer to neutral than to strong agreement that the responses resembled what a real patient would say, with wide variability across participants (Fig. 7).

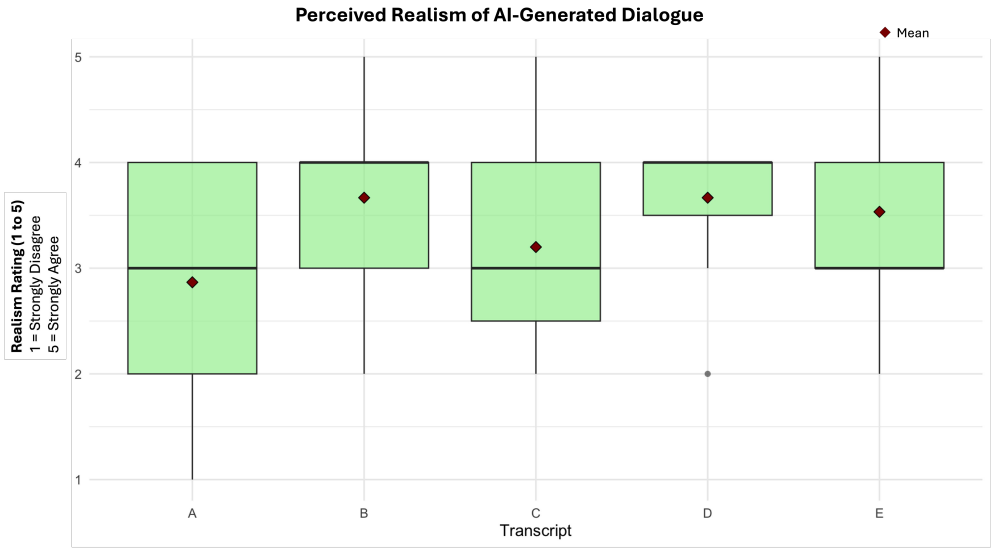

Fig. 7 Distribution of realism ratings (1 = strongly disagree, 5 = strongly agree that the patient's responses matched what a real patient would say) by transcript. Diamonds indicate means

Cumulative link mixed models identified no significant predictors of realism ratings. Profile accuracy was not associated with perceived realism ($\beta = -0.38$, SE = 0.42; z = −0.89; P=.37; OR = 0.69, 95% CI 0.30–1.57). Demographic variables were also nonsignificant (all P>.40). Confidence in detecting personality traits (OR = 1.49; P=.11) and familiarity with AI-generated content (OR = 1.78; P=.09) showed weak, nonsignificant trends and are reported for completeness (Table 5).

Table 5 Fixed effects from cumulative link mixed models predicting realism ratings

Predictor	Estimate	SE	z	*P*	OR	95% CI
Profile Accuracy	-0.38	0.42	-0.89	.37	0.69	[0.30, 1.57]
Sex	0.40	0.52	0.77	.44	1.49	[0.54, 4.11]
Age	-0.21	0.30	-0.71	.48	0.81	[0.45, 1.45]
Confidence	0.40	0.25	1.59	.11	1.49	[0.91, 2.42]
Personality Assessment	-0.39	0.27	-1.45	.15	0.68	[0.40, 1.15]
Familiarity AI Familiarity	0.58	0.34	1.71	.09	1.78	[0.92, 3.43]

Participants showed considerable inconsistency when judging whether dialogues were AI- or human-authored. Across transcripts, dialogues were labeled as AI-generated in 41% of trials. All five transcripts were AI-generated, so "AI-generated" was the correct response on every trial; the observed rate therefore reflects a tendency toward human attribution rather than an ability to discriminate human from AI authorship. AI call rates varied substantially by transcript, ranging from 20% to 60% (Fig. 8).

Generalized linear mixed-effects modeling showed that realism was the only significant predictor of authorship judgments. Higher realism ratings substantially reduced the odds of labeling a dialogue as AI-generated ($\beta = -0.87$, SE = 0.33; z = −2.64; P=.008; OR = 0.42, 95% CI 0.21–0.78). Each one-point increase in realism rating was associated with approximately 58% lower odds of classifying a dialogue as AI-generated. Alignment ratings, profile accuracy, age, and sex were not significant predictors (all P>.40; Table 6). Diagnostic checks indicated no overdispersion (Pearson χ^2 per degree of freedom=0.96) and no systematic deviation of simulated scaled residuals from uniformity (P=.72).

4.3 Computational Text Analysis of Dialogue Evaluations

Computational text analyses revealed consistent patterns in how participants evaluated the consultation transcripts across authorship attribution, personality alignment, and perceived realism. Lexical patterns extracted from participants' open-ended responses suggested that evaluations were primarily driven by language style, content specificity, and behavioral coherence. Across outcomes, participants relied primarily on three evaluative cues: language style, content specificity, and behavioral coherence

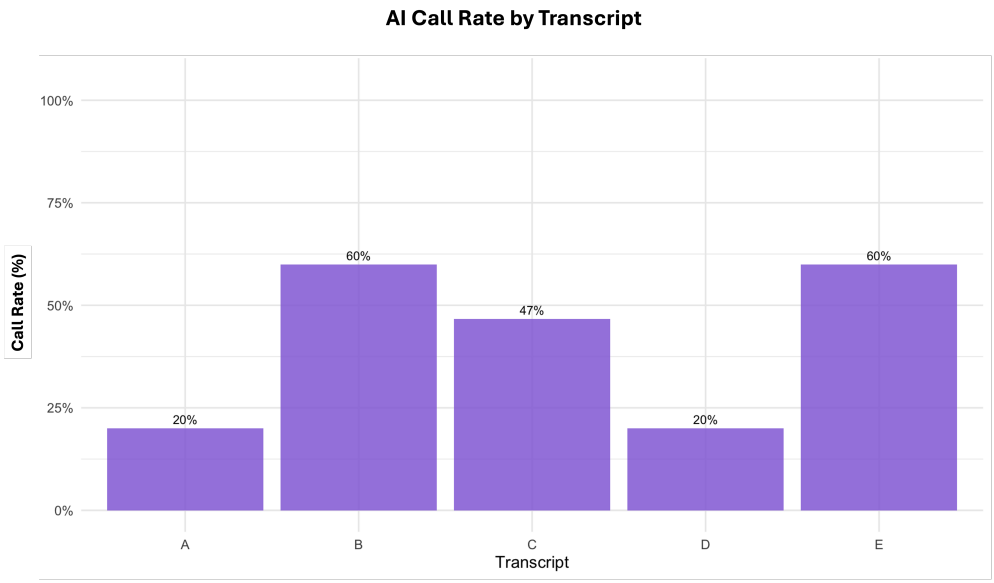

Fig. 8 Percentage of judgments classifying each transcript as AI-generated

Table 6 Logistic mixed model predicting AI vs. Human judgments

Predictor	Estimate	SE	z	*P*	OR	95% CI
Intercept	1.89	1.62	1.17	.24	-	-
Realism Rating	-0.87	0.33	-2.64	.008	0.42	[0.21, 0.78]
Alignment Rating	0.26	0.34	0.76	.45	1.29	[0.66, 2.51]
Profile Accuracy	-0.26	0.54	-0.49	.63	0.77	[0.27, 2.21]
Sex	-0.41	0.54	-0.76	.45	0.66	[0.23, 1.91]
Age	-0.05	0.28	-0.18	.86	0.95	[0.55, 1.66]

across conversational turns. References to emotional expression were present but less frequent and did not emerge as a dominant or systematic basis for judgment.

When explaining authorship attribution, participants most often focused on conversational fluency and coherence. Dialogues labeled as AI-generated were described as short, literal, or fragmented, with abrupt topic shifts or incomplete turn-taking. One participant noted that "most of the answers are simply just yes and no" and that "human generated answers should have had more details or been more complex" (P1), while another observed that, in Transcript B, a response "doesn't really respond to the information presented about medications" (P2). Short answers were not interpreted uniformly, however: another participant reasoned that "when the answers were simpler, such as a simple 'No,' I'm led to believe it's a more human response" (P12).

These comments suggest that perceived breakdowns in interactional flow, rather than isolated wording choices, shaped AI attribution for some participants.

In contrast, dialogues judged as human-authored were described as linguistically varied and internally consistent. Participants highlighted idiomatic phrasing, stylistic variation, and stable tone across turns. Content specificity further supported human attribution, with detailed medication histories, anecdotal references, and concrete personal details frequently cited. For example, one participant remarked that in Transcript C "the patient uses a variety of words (e.g., no, nope, not really, nah) to express 'no' which might not be possible using AI" (P14). These features appeared to signal intentionality and lived experience, reinforcing perceptions of humanness.

Judgments of personality alignment relied on similar interactional cues but were less precise. Participants who agreed with assigned profiles typically referenced consistent behavioral tendencies, such as irritability, sociability, or open-mindedness, reflected in patterns of questioning or responsiveness. Disagreement, however, often reflected skepticism toward the strength or plausibility of the assigned trait levels rather than specific dialogue features. Several participants described using informal heuristics, such as requiring multiple traits to generally align before endorsing agreement, indicating that alignment judgments were based on holistic impressions rather than systematic trait-level inference.

Perceived realism judgments centered on situational appropriateness within the dental consultation context and incorporated both global impressions and localized anomalies. Participants cited familiar clinical behaviors, including anxiety, information overload, and concern about cost, as indicators of plausibility. Lower realism ratings were associated with tonal mismatches, exaggerated statements, or contextually unexpected remarks. At the same time, participants identified isolated dialogue anomalies, such as local contradictions, repetition, or stylistic excess, which were typically described as atypical rather than implausible. Notably, behaviors common in real dental encounters, such as brief or minimally elaborated patient responses, were sometimes penalized as unrealistic. This pattern highlights a tension between ecological validity and perceived realism, suggesting that judgments favored interactional smoothness over strict fidelity to real-world patient behavior.

5 Discussion

This study examined whether AI-generated patient dialogue conditioned on Big Five personality traits produces perceptible variation in simulated dental consultation scripts. The central question was not whether participants could formally identify the intended Big Five profiles, but whether personality-conditioned generation produced patient dialogues that felt different in communication style, believable as patient interactions, and potentially useful for creating varied virtual patient personas.

Participants appeared to perceive differences in overall conversational impressions, but these differences did not reliably map onto the intended Big Five trait profiles. Although profile-matching accuracy was low and inferred ratings did not consistently track assigned trait values (extraversion being the clearest exception), participants

showed moderate agreement that some dialogues aligned with their assigned personality profiles. This suggests that personality conditioning may have influenced broad conversational tone or behavioral style, even though users did not reliably interpret those differences as specific Big Five trait configurations.

One possible explanation for this pattern comes from personality judgment research. Personality judgment research distinguishes between trait-wise judgments of individual personality dimensions and profile-based judgments that reflect broader impressions formed across multiple characteristics and behaviors (Letzring et al. 2021). In the present study, participants showed limited ability to identify specific Big Five trait levels. Profile-level correspondence was nonetheless positive on average (mean r=0.34) although it varied widely across transcripts (r=−0.20 to 0.59), which is consistent with partial recovery of the overall profile shape despite low exact-match accuracy. However, they still reported moderate agreement that some dialogues matched the assigned personality profiles, suggesting that participants may have recognized overall differences in communication style without accurately identifying which specific traits contributed to those differences. In other words, participants may have perceived that one patient "felt different" from another patient while being unable to precisely describe that difference using individual Big Five dimensions.

Within this sample, perceived realism appeared to play a larger role in authorship judgments than personality conditioning. Dialogues rated as more realistic were significantly less likely to be labeled as AI-generated, suggesting that judgments of authenticity were driven primarily by conversational plausibility and coherence rather than awareness of the underlying personality manipulation.

Analysis of participants' open-ended explanations provides additional insight into how these evaluations were formed. Participants consistently emphasized content specificity, language style, and behavioral coherence when judging dialogue quality. Fluent phrasing, continuity across conversational turns, and contextually appropriate responses supported perceptions of humanness, whereas short, literal, or disjointed responses were often interpreted as artificial. Emotional cues were referenced less frequently and did not emerge as a dominant or reliable basis for judgment, suggesting that evaluative attention was directed primarily toward interactional structure rather than inferred affective state.

The present findings can also be viewed in relation to prior work on AI-driven virtual patients and communication training. Recent reviews have highlighted the potential of AI-supported virtual patients to provide flexible, responsive dialogue and support communication skills development (Stamer et al. 2023; Bowers et al. 2024). However, most prior work has focused on system implementation and educational applications rather than how users interpret personality-conditioned patient communication. Extending this literature, the present study examined whether personality-conditioned dialogue produced perceptible differences in communication style. Although participants appeared sensitive to differences in overall communication style, they did not reliably identify the intended Big Five personality profiles. Instead, their evaluations were driven primarily by conversational coherence, language style, and contextual plausibility. These findings suggest that generating

personality-conditioned dialogue and successfully communicating intended personality characteristics to users may represent distinct challenges.

Importantly, some communicative behaviors that are common in real dental consultations, such as brief or minimally elaborated patient responses, were penalized as unrealistic. This finding highlights a tension between ecological validity and perceived realism. Simulation systems optimized to maximize perceived authenticity may inadvertently privilege idealized communication patterns over representative patient behavior. For designers of AI-driven medical simulations, these findings suggest that improving conversational coherence, interactional flow, and contextual grounding may have greater perceptual impact than increasing the expressiveness of latent personality parameters alone.

From a design and educational perspective, these findings clarify the role of personality conditioning in AI-based medical training tools. Personality conditioning appeared to influence overall conversational tone, but users did not reliably recognize distinct Big Five personality profiles. For educational applications, this suggests that personality conditioning may be most useful when paired with explicit behavioral cues, structured tasks, or reflective prompts that help learners recognize meaningful differences in patient communication styles.

5.1 Limitations

This study focused on isolating language-based cues in the evaluation of AI-generated patient dialogue. To support this goal, participants were drawn from the general public rather than from dental trainees or practicing clinicians. This approach enabled examination of baseline perceptual judgments without professional training or domain-specific expectations. However, clinicians may evaluate conversational realism, plausibility, and appropriateness differently. Future work is needed to determine whether clinical training influences sensitivity to personality cues and communicative realism. Five of the 15 participants reported prior training in communication or interpersonal skills; the subgroups were too small to test whether such training influenced their judgments.

The small sample size also limited statistical power, particularly for detecting small tracking slopes between assigned and perceived personality traits. Although the within-subjects design provided multiple transcript-level observations per participant, the number of participants was limited. Therefore, nonsignificant effects should be interpreted cautiously and should not be taken as evidence that personality conditioning has no effect. Larger studies are needed to estimate the magnitude and reliability of these associations.

The use of text-only consultation transcripts enabled controlled comparison of wording, content specificity, and interactional structure across conditions. This design choice reduced confounds associated with speech timing, voice quality, and embodiment, allowing focused analysis of linguistic and structural features. However, dialogue presented in text form does not capture the full range of cues present in embodied or interactive simulations. Subsequent studies should examine whether similar evaluative patterns emerge when speech, timing, and nonverbal behaviors are incorporated.

Personality alignment was examined using single dialogue instances paired with brief Big Five trait descriptors. While this supported systematic comparison across personality conditions, personality expression may become more apparent across longer or repeated interactions. At the same time, the consultation length used in this study is comparable to the brief encounters commonly used in communication-focused simulation, and clinicians often form impressions of patient communication styles during relatively short interactions. Consequently, the limited personality recognition observed here may reflect both constraints of the interaction format and limitations in how personality-conditioned dialogue was expressed. Future work should examine whether personality-related communication differences become more perceptible during extended or interactive encounters and whether such differences influence educational outcomes. In addition, the present study focused on a dental consultation scenario. Whether similar patterns would be observed in other clinical contexts, particularly those involving greater emotional intensity or uncertainty, remains unknown and warrants future investigation. Finally, the assigned profile was disclosed after the authorship item but before the alignment and realism ratings, so those two ratings may have been influenced by knowledge of the assigned profile.

A human-authored comparison condition was not included in this exploratory study. Consequently, it remains unclear whether the limited personality inference observed in this sample reflects limitations of LLM-based personality conditioning, the difficulty of conveying personality through brief consultation transcripts, or broader challenges associated with personality perception in this context. Future studies should compare AI-generated and expert-authored patient dialogues to determine whether personality-conditioned LLM outputs differ from human-designed communication scenarios in their ability to convey distinct behavioral characteristics. Without human-authored stimuli, the authorship task also cannot separate discrimination ability from a general tendency to attribute dialogue to human authors, because all five transcripts were AI-generated.

Participants' judgments may also have been influenced by preexisting expectations regarding AI-generated communication. The observed association between perceived realism and human authorship attribution may partly reflect assumptions that AI-generated dialogue is more robotic, formulaic, or less natural than human-written dialogue. Because these expectations were not directly measured, the extent to which authorship judgments reflected dialogue features versus prior beliefs about AI remains unclear. In addition, all participants evaluated the transcripts in the same fixed order (A–E), so transcript-specific effects cannot be separated from order effects such as fatigue or learning.

6 Conclusions

Participants relied primarily on content specificity, language style, and behavioral coherence when evaluating personality alignment, perceived realism, and authorship of AI-generated patient dialogue, whereas emotional cues played a comparatively limited role. In this exploratory study, personality-conditioned dialogue appeared to create different overall conversational impressions, but these impressions did not reliably map

onto the intended Big Five profiles. This suggests that personality conditioning may be more useful as a mechanism for generating varied patient communication styles than as a precise method for conveying specific trait configurations. Perceived realism was closely related to authorship judgments, indicating that participants tended to view more realistic dialogues as more human-like. These findings suggest that effective educational use of AI-driven virtual patients may require empirical evaluation with intended end users, explicit instructional framing, and design strategies that prioritize interactional coherence and contextual realism over latent personality parameterization alone.

Statements and Declarations

Funding. The authors did not receive support from any organization for the submitted work.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. The study protocol was reviewed by the University of South Florida Institutional Review Board and determined to be exempt.

Consent to participate. Participants reviewed an informed consent form and provided electronic consent prior to participation.

Consent for publication. Not applicable.

Data availability. The de-identified participant-level dataset (survey responses and derived variables), the five consultation transcripts, the survey instruments, and the R scripts used for the analyses are available from the corresponding author on reasonable request.

Author contributions. BK: Conceptualization; Methodology; Formal Analysis; Investigation; Project Administration; Supervision; Visualization; Writing – Original Draft; Writing – Review & Editing.
PN: Writing – Review & Editing.
AH: Software; Writing – Original Draft.

References

Adachi T, Tsunekawa Y, Tanimura D (2022) Association between the Big Five personality traits and medication adherence in patients with cardiovascular disease: A cross-sectional study. PLOS ONE 17:e0278534. https://doi.org/10.1371/journal.pone.0278534

Aggarwal R, Mytton OT, Derbrew M, et al (2010) Training and simulation for patient safety. BMJ Quality & Safety 19:i34–i43. https://doi.org/10.1136/qshc.2009.038562

Al-Elq A (2010) Simulation-based medical teaching and learning. Journal of Family and Community Medicine 17:35. https://doi.org/10.4103/1319-1683.68787

Amiri M, Farhoodi F, Abdolvand N, et al (2011) A study of the relationship between Big-five personality traits and communication styles with marital satisfaction of married students majoring in public universities of Tehran. Procedia - Social and Behavioral Sciences 30:685–689. https://doi.org/10.1016/j.sbspro.2011.10.132

Armstrong MJ, Weisbrod NJ, Bylund CL (2021) Strategies to Improve Clinician-Patient Communication Experiences for Patients With Neurologic Conditions. Neurology: Clinical Practice 11:e896–e900. https://doi.org/10.1212/CPJ.0000000000001091

Asan O, Yu Z, Crotty BH (2021) How clinician-patient communication affects trust in health information sources: Temporal trends from a national cross-sectional survey. PLoS ONE 16:e0247583. https://doi.org/10.1371/journal.pone.0247583

Birkhäuer J, Gaab J, Kossowsky J, et al (2017) Trust in the health care professional and health outcome: A meta-analysis. PLOS ONE 12:e0170988. https://doi.org/10.1371/journal.pone.0170988

Blackmore A, Kasfiki EV, Purva M (2018) Simulation-based education to improve communication skills: a systematic review and identification of current best practice. BMJ Simulation & Technology Enhanced Learning 4:159–164. https://doi.org/10.1136/bmjstel-2017-000220

Bosse HM, Nickel M, Huwendiek S, et al (2015) Cost-effectiveness of peer role play and standardized patients in undergraduate communication training. BMC Medical Education 15:183. https://doi.org/10.1186/s12909-015-0468-1

Bowers P, Graydon K, Ryan T, et al (2024) Artificial intelligence-driven virtual patients for communication skill development in healthcare students: A scoping review. Australasian Journal of Educational Technology 40:39–57. https://doi.org/10.14742/ajet.9307

Bucher MA, Suzuki T, Samuel DB (2019) A meta-analytic review of personality traits and their associations with mental health treatment outcomes. Clinical Psychology Review 70:51–63. https://doi.org/10.1016/j.cpr.2019.04.002

Datta R, Upadhyay K, Jaideep C (2012) Simulation and its role in medical education. Medical Journal Armed Forces India 68:167–172. https://doi.org/10.1016/S0377-1237(12)60040-9

Epstein RM, Fiscella K, Lesser CS, et al (2010) Why the nation needs a policy push on patient-centered health care. Health Affairs 29:1489–1495. https://doi.org/10.1377/hlthaff.2009.0888

Furlan R, Gatti M, Menè R, et al (2021) A Natural Language Processing–Based Virtual Patient Simulator and Intelligent Tutoring System for the Clinical Diagnostic Process: Simulator Development and Case Study. JMIR Medical Informatics

9:e24073. https://doi.org/10.2196/24073

Gilligan C, Powell M, Lynagh MC, et al (2021) Interventions for improving medical students' interpersonal communication in medical consultations. The Cochrane Database of Systematic Reviews 2021:CD012418. https://doi.org/10.1002/14651858.CD012418.pub2

Hamdy H, Al-Moslih A, Tavarnesi G, et al (2017) Virtual patients in problem-based learning. Medical Education 51:557–558. https://doi.org/10.1111/medu.13293

Holderried F, Stegemann-Philipps C, Herschbach L, et al (2024) A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study. JMIR Medical Education 10:e53961. https://doi.org/10.2196/53961

Jakesch M, Hancock JT, Naaman M (2023) Human heuristics for AI-generated language are flawed. Proceedings of the National Academy of Sciences 120:e2208839120. https://doi.org/10.1073/pnas.2208839120

Kobayashi M, Iwamoto M, Une S, et al (2022) Simulated communication skills training program effects using augmented reality with real-time feedback: A randomized control study. Alzheimer's & Dementia 18:e062055. https://doi.org/10.1002/alz.062055

Kyung D, Chung H, Bae S, et al (2025) PatientSim: A Persona-Driven Simulator for Realistic Doctor-Patient Interactions. In: Advances in Neural Information Processing Systems, URL https://proceedings.nips.cc/paper_files/paper/2025/hash/24945e3bdc7b3f4b2e64b9979a16f38e-Abstract-Datasets_and_Benchmarks_Track.html

Lee J, Kim KH, Jung D, et al (2020) Effective virtual patient simulators for medical communication training: A systematic review. Medical Education 54:786–795. https://doi.org/10.1111/medu.14152

Letzring TD, Murphy NA, Allik J, et al (2021) The Judgment of Personality: An Overview of Current Empirical Research Findings. Personality Science 2:e6043. https://doi.org/10.5964/ps.6043

Leu HB, Yin WH, Tseng WK, et al (2019) Impact of type D personality on clinical outcomes in Asian patients with stable coronary artery disease. Journal of the Formosan Medical Association 118:721–729. https://doi.org/10.1016/j.jfma.2018.08.021

Leung SK, Bond MH (2001) Interpersonal communication and personality: Self and other perspectives. Asian Journal of Social Psychology 4:69–86. https://doi.org/10.1111/1467-839X.00076

Lipps J, Bhandary S, Meyers L (2017) The expanding use of simulation for undergraduate preclinical medical education. International Journal of Academic Medicine 3:59. https://doi.org/10.4103/IJAM.IJAM_40_17

Meerdink M, Khan J (2021) Comparison of the use of manikins and simulated patients in a multidisciplinary in situ medical simulation program for healthcare professionals in the United Kingdom. Journal of Educational Evaluation for Health Professions 18:8. https://doi.org/10.3352/jeehp.2021.18.8

Monroe BL, Colaresi MP, Quinn KM (2008) Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Political Analysis 16:372–403. https://doi.org/10.1093/pan/mpn018

Paget L, Han P, Nedza S, et al (2011) Patient-Clinician Communication: Basic Principles and Expectations. Tech. rep., National Academy of Medicine, https://doi.org/10.31478/201106a

Pennebaker JW, King LA (1999) Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology 77:1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Redelmeier DA, Najeeb U, Etchells EE (2021) Understanding Patient Personality in Medical Care: Five-Factor Model. Journal of General Internal Medicine 36:2111–2114. https://doi.org/10.1007/s11606-021-06598-8

Scales D (2018) Communicating Against the Current. Journal of Health Care for the Poor and Underserved 29:1046–1053. https://doi.org/10.1353/hpu.2018.0078

Schiffer AA, Denollet J, Widdershoven JW, et al (2007) Failure to consult for symptoms of heart failure in patients with a type-D personality. Heart 93:814–818. https://doi.org/10.1136/hrt.2006.102822

Sharkiya SH (2023) Quality communication can improve patient-centred health outcomes among older patients: a rapid review. BMC Health Services Research 23:886. https://doi.org/10.1186/s12913-023-09869-8

Shneiderman B, Plaisant C, Cohen M, et al (2016) Designing the User Interface: Strategies for Effective Human-Computer Interaction, 6th edn. Pearson

Shorey S, Ang E, Yap J, et al (2019) A Virtual Counseling Application Using Artificial Intelligence for Communication Skills Training in Nursing Education: Development Study. Journal of Medical Internet Research 21:e14658. https://doi.org/10.2196/14658

Shorey S, Ang E, Ng ED, et al (2020) Communication skills training using virtual reality: A descriptive qualitative study. Nurse Education Today 94:104592. https://doi.org/10.1016/j.nedt.2020.104592

Silge J, Robinson D (2016) tidytext: Text Mining and Analysis Using Tidy Data Principles in R. The Journal of Open Source Software 1:37. https://doi.org/10.21105/joss.00037

Soto CJ, John OP (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 113:117–143. https://doi.org/10.1037/pspp0000096

Stamer T, Steinhäuser J, Flägel K (2023) Artificial Intelligence Supporting the Training of Communication Skills in the Education of Health Care Professions: Scoping Review. Journal of Medical Internet Research 25:e43311. https://doi.org/10.2196/43311

Suciu N, Meliţ LE, Mărginean CO (2021) A Holistic Approach of Personality Traits in Medical Students: An Integrative Review. International Journal of Environmental Research and Public Health 18:12822. https://doi.org/10.3390/ijerph182312822

Świątoniowska-Lonc N, Polański J, Tański W, et al (2020) Impact of satisfaction with physician-patient communication on self-care and adherence in patients with hypertension: cross-sectional study. BMC Health Services Research 20:1046. https://doi.org/10.1186/s12913-020-05912-0

Tausczik YR, Pennebaker JW (2010) The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology 29:24–54. https://doi.org/10.1177/0261927X09351676

Themes UFO (2021) 1 New Patient Visit. Pocket Dentistry: Fastest Clinical Dentistry Insight Engine, URL https://pocketdentistry.com/1-new-patient-visit/, accessed 10 September 2025

Zeng J, Qi W, Shen S, et al (2025) Embracing the Future of Medical Education With Large Language Model–Based Virtual Patients: Scoping Review. Journal of Medical Internet Research 27:e79091. https://doi.org/10.2196/79091

Ziv A, Wolpe PR, Small SD, et al (2006) Simulation-Based Medical Education: An Ethical Imperative. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare 1:252–256. https://doi.org/10.1097/01.SIH.0000242724.08501.63