

Talking to Avatars: How Human-Like Should a Virtual Health Assistant Be in Medical History Collection?

Ying Chen

y.chen11@student.tue.nl

Eindhoven University of Technology

Eindhoven, the Netherlands

Abstract

High-quality medical history is essential for accurate clinical diagnosis. However, traditional consultations often face barriers such as time limits, low patient expressiveness, and weak emotional connection. In recent years, the combination of large language models (LLMs) and Virtual Avatar technologies has provided new possibilities to enhance the efficiency and user experience of medical history taking. This study investigates the impact of virtual characters with different anthropomorphic levels on user behaviour and data quality in AI-driven medical history taking.

In this study, a virtual consultation system was developed, integrating GPT-based dialogue capabilities, and three virtual human characters with different anthropomorphism levels were designed: low anthropomorphism (static icons), medium anthropomorphism (stylised characters), and high anthropomorphism (realistic characters). Through a within-group experimental design, a total of 25 participants completed the medical history taking task with each of the three personas, and quantitative data such as user response quality, usability, and experience of use, as well as qualitative feedback such as semi-structured interviews, were collected.

The results showed that the medium-high anthropomorphic characters significantly increased users' emotional comfort, trust, and motivation to express information. However, the overly realistic appearance may cause discomfort due to the 'Uncanny Valley effect', which may inhibit natural expression. Behavioural synchronisation, emotional responsiveness and social role setting are the key design dimensions that affect user engagement. Based on this, this study proposes a multidimensional, context-sensitive anthropomorphic design framework that emphasises the balance between visual realism and interaction quality.

This study provides empirical insights into anthropomorphic design strategies for virtual health assistants and offers guidance for building more empathetic digital agents..

Keywords: Digital people; virtual characters; anthropomorphic design; medical history taking; user experience; large language modelling; healthcare human-computer interaction

1 Introduction

1.1 Research Background

Even in today's highly developed era of medical imaging and genetic testing, detailed and accurate medical history information remains an indispensable key component of clinical diagnosis. Research indicates that accurate diagnosis relies on comprehensive and precise medical history information; incomplete or inadequately recorded medical histories may lead to inappropriate treatment decisions[5], which could even jeopardize patient safety. However, in actual clinical practice, the collection of high-quality medical history information faces numerous challenges: unclear patient expression, communication barriers, time constraints, and heavy workloads for doctors can all lead to missing or low-quality information. These issues prompt us to consider whether new technological tools can be used to assist in this process. The Prospects of Artificial Intelligence Chatbots and Virtual Avatars in Clinical Consultations In recent years, the development of large language models (LLMs) has enabled chatbots to demonstrate significant capabilities in understanding natural language, probing for details, and generating structured information [12] [1]. When these systems are combined with virtual avatars—digital humans capable of expressing natural speech, facial expressions, and body language—they have the potential to create more approachable and immersive interactive experiences, thereby potentially alleviating patients' psychological stress and facilitating the disclosure of sensitive information[15][25]. The integration of LLMs with virtual avatars holds promise for faster, more comprehensive, and patient-centered solutions for medical history collection.

1.2 Research Motivation

As artificial intelligence and Virtual Avatar technology continue to advance in medical settings, their "human-like degree" has been shown to influence users' trust, intimacy, and willingness to disclose information[25]. Highly human-like characters may enhance interactivity but may also trigger the "uncanny valley" effect[27]; while simplified characters may be safer, they may reduce expression willingness and interaction efficiency[21]. Therefore, balancing the human-like degree is a key issue in the design of digital medical systems. However, existing research has primarily focused on single character styles, lacking systematic comparisons

of characters with different levels of anthropomorphism under unified tasks.[6]. Additionally, there have been few attempts to evaluate chat engines and virtual avatars as an integrated system[10]. This study builds upon this foundation, systematically examining the impact of different levels of anthropomorphism on user behavior and data quality.

1.3 Research Objectives and Contributions

Building on prior exploratory findings, this study designed and developed a medical history collection system integrating large language models with virtual Virtual Avatars, aiming to enhance the efficiency and quality of information collection while providing users with a natural and seamless experience. To systematically evaluate the effects of different levels of anthropomorphism, this study designed three distinct Virtual Avatar interfaces:

- a High-fidelity, expressive, and animated anthropomorphic virtual avatars;
- b Stylized 3D characters with basic body movements;
- c Simplified static avatars with voice feedback.

Therefore, I propose two research questions:

- RQ1: In AI chatbot-driven conversations for collecting medical history information, how does the quality of user responses differ between characters with different levels of anthropomorphism?
- RQ2: In AI chatbot-driven conversations for collecting medical history information, how does the user experience differ between characters with different levels of anthropomorphism?

The main contributions of this study include:

- Designing a functional LLM-driven virtual medical history collection system and applying it to real-user testing;
- Comparing three different levels of anthropomorphism through experiments to reveal their specific impacts on user information disclosure and experience;
- Summarizing users' expectations and feedback regarding AI-based consultations to provide suggestions for future optimizations
- Providing a reference model and data support for designing smarter, more user-friendly medical communication tools in the future.

2 Related Work

In recent years, the integration of medical informatics and human-computer interaction technology has been driving a new revolution: the capabilities of intelligent systems are no longer limited to “judging right from wrong,” but have begun to “generate content”; interaction interfaces are no longer cold tables or text, but have begun to become concrete and humanized[10]. Against this backdrop, medical history collection, a fundamental clinical process, has also ushered

in a new definition and research perspective. This section reviews three areas of work closely related to this study: (1) the development of medical chatbots and voice assistants; (2) intelligent dialogue systems aimed at information collection; and (3) the exploration of Virtual Avatar technology applications in medical dialogue. It further identifies gaps and shortcomings in current research, outlines the preliminary explorations conducted under the PFMP program, and explains the research focus and significance of this paper.

2.1 The Evolution of Medical Chatbots and Voice Assistants

The earliest medical chatbots are more like “voice-enabled questionnaire forms”—the system could only ask questions along a predefined path, and patients could only respond selectively[6]. Once the conversation deviated from the script, the process was likely to be interrupted. While this highly structured approach could control risks, it also limited users' freedom of expression and the authenticity of the conversation.

With the development of natural language processing technology, especially breakthroughs in semantic understanding through deep learning, medical AI systems have begun to develop stronger contextual understanding capabilities. For example, tools like Ada Health and Buoy Health can extract keywords from user expressions and dynamically adjust follow-up questions based on semantic content, enhancing the flexibility and specificity of conversations.

The true turning point came with the rise of large language models (LLMs). Researchers such as Fadhil and Schiavo [10] demonstrated the high adaptability of LLMs to medical contexts; Athotaathota2020chatbotintegrated generative dialogue into triage systems, enabling automatic summarization and integration with electronic health records, thereby reducing the burden of manual record-keeping.

The addition of voice assistants has further expanded interaction methods. Platforms like Google Assistant and Amazon Alexa have launched voice features for medical scenarios, allowing users to record health data such as blood sugar and blood pressure using spoken language. This method is particularly suitable for users with visual impairments or those unfamiliar with text. Research shows that in chronic disease management, voice input increased follow-up completion rates from less than 50 percent to nearly 70 percent [12].

However, most voice assistants currently rely on pre-set scripts, offering conservative diagnostic suggestions and often responding with a generic “I recommend you seek medical attention.” Additionally, they lack the ability to understand and respond to emotions: when users express anxiety, fear, or confusion, the system typically cannot provide comforting responses, which remains a significant barrier to their deeper integration into clinical settings.

2.2 Intelligent Dialogue Systems Focused on Information Collection

Unlike systems aimed at providing treatment recommendations, information collection-oriented chatbots focus on enabling users to “speak more, speak accurately, and speak truthfully.” Traditional form-filling methods are relatively mechanical, often limiting users to fixed options and making it difficult to express their true circumstances.

Semi-structured dialogue[12], however, offers a more flexible approach—users can start by discussing the issues they are most concerned about, and the system then supplements the information through follow-up questions. For example, Chen[20] found in their study on adverse drug reaction reporting that, compared to traditional methods, conversational interfaces can enhance the effectiveness of reports, particularly in capturing more details related to daily life in free-text fields.

The advantage of such systems lies in their ability to understand users’ language in real time and map free-form narratives to standard medical coding systems (e.g., SNOMED CT[16]), thereby reducing manual data entry time and workload.

However, most such systems currently use anonymous text or voice formats, lack concrete character imagery, and do not have a clear “questioner” presence. In scenarios involving sensitive topics (such as mental health or sexual health), whether the questioner is trustworthy and has social attributes may directly affect whether users are willing to speak up.

This leads to the next direction worth exploring: the introduction of Virtual Avatars.

2.3 Virtual Avatars and the Role of Anthropomorphic Design

Virtual Avatars—animated avatars capable of speech, gaze, and facial expression—introduce a visual and social dimension to health dialogues. Kurniawan[11][9] found that even 2D avatars with emotional voice increased user trust in health advice. Stock[25] reported that animated facial expressions helped reduce embarrassment thresholds in sexual health consultations, promoting richer self-disclosure. Users often unconsciously apply social norms when interacting with anthropomorphic agents, as explained by the Media Equation theory [22] and the CASA paradigm[19]. So behavioral cues such as nodding or brow furrowing have been shown to signal active listening, increasing perceived empathy and prolonging user engagement in psychological counseling contexts.

However, more realism does not always lead to better outcomes. Robertson[23] highlighted the “uncanny valley” effect[18][13]: when avatars appear realistic but lack synchronized expressions or natural behavior, users may experience discomfort or even aversion. Similar effects occur with

desynchronized lip movements—delays beyond 240ms can significantly lower perceived professionalism.

In the context of this study, anthropomorphism refers to the extent to which a digital character exhibits human-like characteristics. This includes visual realism, behavioral traits, and emotional cues. Prior work[7, 11] has shown that these dimensions collectively influence how users perceive and engage with digital agents. While prior research has tested avatars with different levels of realism, studies often focus on single styles, lacking direct comparison across a spectrum of anthropomorphism within the same task and user group.

This study addresses that gap by systematically comparing three levels of avatar anthropomorphism—abstract, stylized, and realistic—within a unified system for medical history collection. By holding task structure and user population constant, it investigates how the degree of anthropomorphic design affects user experience, and the completeness of medical disclosure.

3 Overall Research Approach and Process

This project adopts a Research through Design (RtD) methodology[29], combining iterative prototyping, user testing, and reflection to generate insights into human-avatar interaction in medical contexts. The process also results in a final design outcome, which embodies the validated interaction principles and user-centered features identified during the study. As shown in Figure 1.

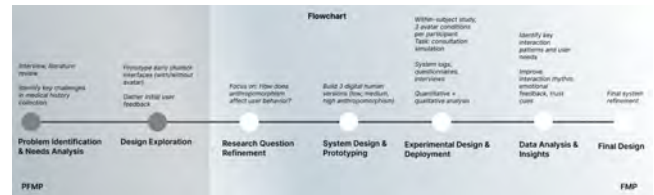


Figure 1. Overall Research Process

3.1 Problem Identification Needs Analysis

I began by identifying challenges with the current patient history taking interface, such as lack of engagement, poor completeness of responses and limited emotional connection. A literature review and expert interviews guided this stage.

3.2 Design Exploration

I conducted a small-scale design exploration by prototyping early chatbot interfaces with different interaction modalities (with avatar and without avatar). This stage helped reveal user preferences and guided the initial design space.

3.3 Research Question Refinement

Based on insights from exploration, I refined the core research question: How does the degree of anthropomorphism in Virtual Avatars affect user experience and the quality of user responses?

3.4 System Design Prototyping

A virtual avatar interface was developed in Unity, integrating GPT dialogue, voice interaction, expression, gestures, and structured data collection. Three levels of anthropomorphism are implemented.

3.5 Experimental Design Deployment

I analyzed quantitative (e.g., data completeness, user ratings) and qualitative (e.g., interview themes) results to evaluate the effects of anthropomorphism. Insights are used to refine both system and design recommendations.

3.6 Data Analysis Insight

I analyzed quantitative (e.g., data completeness, user ratings) and qualitative (e.g., interview themes) results to evaluate the effects of anthropomorphism. Insights are used to refine both system and design recommendations.

3.7 Final Design

Based on the study results, design optimization suggestions are proposed, leading to a round of prototype iteration. Revisions are made to areas identified by users as needing improvement, culminating in the final version of the design.

4 Exploration: PFMP work summary

4.1 Problem Identification Needs Analysis

During the discovery phase, interviews with healthcare professionals and a review of relevant literature highlighted key issues in traditional history taking: limited patient expressiveness, inefficient data collection, and a lack of emotional connection during communication. Healthcare professionals emphasized that they wanted a system that would encourage patients to share information while organizing the collected data in a structured format they are familiar with. Based on these insights, three primary design goals are defined:

- a Improve the Efficiency of History Taking Chatbot should use patient-centered communication methods (Deveugele et al., 2005) to promote positive patient response. And A structured history report is then generated to inform the physician's diagnosis (Grice et al., 2017).
- b Providing a Good Interactive Experience Face-to-face counselling through the use of verbal and non-verbal behaviors such as empathy and immediacy can increase trust and satisfaction between patients, thus promoting better health communication and understanding (Berman Chutka, 2016). As well as creating a more immersive experience through the introduction of avatars, increasing user willingness to use (Stock et al. 2023).
- c Promoting Precision Medicine Ensure the completeness and accuracy of medical history data to provide physicians with reliable decision support (Peterson et

al., 1992). Encourage patients to give more comprehensive answers through multiple rounds of questioning to detect potential symptoms and hidden problems and avoid missing key information (Burt et al., 2017).

4.2 Exploratory Design Study 1

To achieve the design objectives, this phase adopted the "rapid prototyping-initial testing" method to complete the design of two prototype versions. Both Version 1 and Version 2 systems integrate large language models, supporting the collection of medical history information through dialogue and the generation of structured medical history reports. The difference lies in the presentation form of the dialogue interface:

- Version 1: A traditional chat box input interface with a 2D virtual avatar, featuring a clear interface structure that aligns with current user habits for chatbot usage; As shown in Figure 2.
- Version 2: A voice input interface with a full-body virtual avatar, providing a stronger sense of immersion through natural language, facial expressions, and voice intonation, which better aligns with users' natural conversation. As shown in Figure 3.

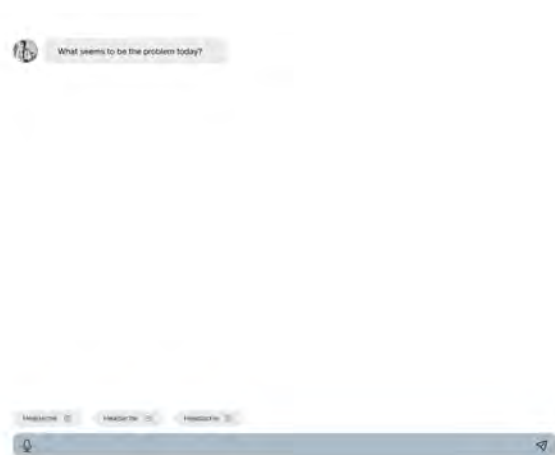


Figure 2. Chat box based GUI

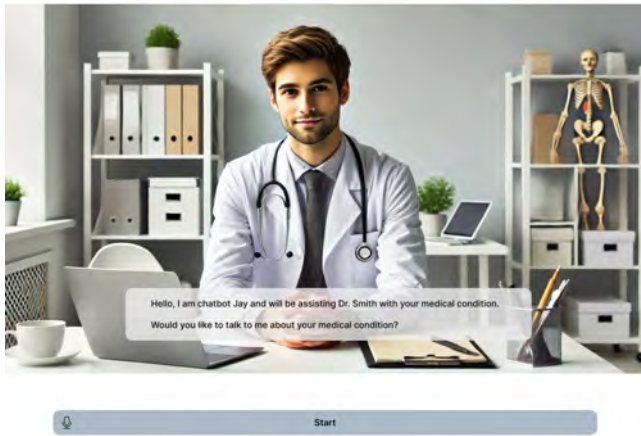


Figure 3. meeting-style GUI

This not only enables the collection of complete and structured symptom information but also offers a more empathetic and engaging user experience. This combination aims to simulate realworld medical consultations, helping users feel more connected and natural during interactions. Additionally, avatar design plays a crucial role in user engagement. A young male doctor was selected, who appears calm and professional and can synchronize voice through text-to-speech (TTS) technology.

4.3 Preliminary User study

4.3.1 Procedure. To evaluate the initial effectiveness of the two prototypes, a small-scale user study was conducted with six participants aged 20–30, all of whom had prior medical consultation experience. Each participant completed two interaction sessions—one with each version—within a simulated consultation scenario involving persistent headaches. Sessions lasted approximately five minutes, during which participants described their symptoms and responded to the system’s questions. Post-session data are collected through semi-structured interviews and open-ended feedback.

4.3.2 Data Collection. The following data sources are used:

- Observational logs: including response length, turn-taking behavior, and interaction time;
- Interviews: covering user experience, emotional responses, and perceptions of the avatar;
- Open-ended feedback: focused on system design, voice interaction, and visual preferences.

4.3.3 Key Findings and Research Focus. The results revealed clear differences between the two interaction modes. All participants showed greater initial engagement with the voice-based interface, expressing curiosity about the speaking avatar. Version 2 elicited longer, more emotionally expressive responses, with users often speaking in full sentences

and using descriptive language. In contrast, Version 1 interactions tended to be shorter and more concise, though occasionally more efficient. Participants appreciated the system’s structured question flow and polite tone in both versions. The dynamic follow-up questions driven by the LLM improved the completeness of symptom reporting. Feedback on the avatar was generally positive—users described it as professional and reassuring—but some noted its lack of emotional variability and expressed interest in more customizable features. These findings highlight the potential of more human-like interfaces to stimulate user expression and foster trust. While previous studies have shown that anthropomorphic design elements can enhance communication, the specific effects of varying degrees of anthropomorphism remain underexplored. With this as a basis, the research questions for this project are further defined.

5 Design and Implementation

5.1 Design Objectives and Constraints Under Research-Driven Design

The system was developed not as a final product but as an experimental platform to test how avatar anthropomorphism influences user behavior in medical consultations. Accordingly, the design followed three key principles:

- **Structural consistency:** All versions of the system share identical task flow, dialogue logic, and data handling, ensuring comparability across conditions.
- **Variable control:** Anthropomorphism is the only manipulated factor. Interaction rhythm, question content, and voice input mechanisms remain constant.
- **Data traceability:** All user responses are automatically converted into structured fields, supporting quantitative analysis of response completeness and clarity.

This setup enables a controlled environment for evaluating behavioral and perceptual differences between avatar styles.

5.2 virtual avatar Construction and Unified Settings

To ensure that virtual avatars under different anthropomorphization conditions have a unified basic setting while maintaining operational style tiers, this study designed and constructed three visually distinct virtual avatars (as shown in Figure 6-1) centered on the role identity of a “female medical assistant.”

5.2.1 Unified Character Settings. All three characters are set as female medical assistants in their 30s, with a unified attire style (white coat + light blue shirt) to create a professional, approachable, and culturally neutral image. This setting aims to create a professional yet trustworthy, non-authoritarian medical interaction atmosphere, facilitating users’ trust and willingness to express themselves during virtual consultations. The selection of female characters is based on common findings in health communication and

virtual assistant research: users tend to trust female characters more in health-related interactions, especially when it comes to symptom description and emotional expression, as female characters are more likely to elicit openness from users [26][2]. The character’s age is set in the middle-to-young adult range to strike a balance between “professional credibility” and “communication friendliness,” avoiding the sense of distance that comes with an older age and the lack of authority that comes with excessive youthfulness [?]. The facial design style avoids excessive cultural characteristics and personalized elements, with simple, neutral hairstyles and makeup to reduce cognitive interference caused by physical features and ensure the universality of the image. .



Figure 4. Visual comparison of three avatar anthropomorphism levels.

5.2.2 Three expressions of anthropomorphic level.

The three virtual characters in this study are developed based on a unified design framework, forming a hierarchical structure in terms of anthropomorphism—from abstract icons to stylized cartoons to realistic human-like models.

- a :Low-anthropomorphic characters (left) are represented by two-dimensional static icon-style abstract images. These avatars lack facial, limb, or gender features and adopt a blue-green iconographic design with prominent medical symbols (e.g., nurse caps, cross marks). Their purpose is to convey a basic “medical identity” while eliminating emotional or social cues, thereby emphasizing the tool-like nature of the interaction.
- b : Medium-anthropomorphic characters (center) take the form of stylized 3D female doctor figures, featuring large eyes, simplified facial traits, and smooth animated contours. With exaggerated hairstyles and soft expressions, these avatars offer a friendly and approachable visual impression. Although the human form is preserved, the overall presentation is cartoonish and deliberately low in realism.
- c : High-anthropomorphic characters (right) are highly realistic 3D female models created using Character Creator 4. These avatars feature natural facial proportions, detailed skin texture, and lifelike eye movement, closely resembling real medical professionals in appearance. The design emphasizes visual realism and emotional naturalness to maximize user immersion.

To systematically manipulate the core variable—degree of anthropomorphism—this stratification was not limited to visual appearance alone. Following the operational model proposed[11], the design further incorporated layered differences in speech expression and non-verbal behavior, such as emotional voice variation, facial motion, and gesture responsiveness. Rather than simply categorizing the interface as “with or without a virtual avatar,” this layered strategy finely adjusts multiple factors related to social presence and interactive performance.[8] This enables more precise control of the experimental variable and enhances the internal validity of the comparative study.

Dimension	High-Fidelity Group (Photorealistic Robot)	Medium-Fidelity Group (3D cartoon character)	Low-Fidelity Group (Abstract UI)
Visual Representation	Full-body 3D realistic model, standing naturally with distinct	3D stylized cartoon character, rendered in non-realistic style	Static abstract image No specific form
Dynamic Expressiveness	Facial changes, synchronized lip-sync, + 10 basic gestures controlled via timeline	Simple expressions (such as smiling) Stylized gestures	No expressions, or gestures
Speech Content	Azure TTS synthesis using SSML markup	User speech input System speech output	User speech input System speech output
Interaction Modality	User speech input system speech output	Moderate Some gestures	Low, static image, no language sense
Social Presence Intensity	Stimulate social touch and interaction Simulate a robot’s interactive social scene	Construct a medium-fidelity state Investigate dynamic representations on	Build a low-social touch scenario foundation Emphasizes visual elements over language

Figure 5. Three expressions of anthropomorphic level

5.3 Consultation Process Design and Interaction Strategy

The system draws inspiration from Stewart and Levenstein’s “patient-centered six-step interview method” [17] in consultation process design, and has been localized to accommodate the characteristics of virtual avatar interaction. The overall design adopts a semi-structured dialogue mechanism, balancing openness and structure. The system guides users to freely state their thoughts through open-ended questions; if the system identifies missing key information, it automatically switches to closed-ended follow-up questions to ensure the continuity and completeness of information. Based on the field-filling status, the system evaluates information coverage in real-time and triggers a dialogue closure prompt once the predefined standards are met, actively asking users, “Is there anything else you’d like to add?” to ensure a balanced structure between structured and open-ended elements in the consultation process.

5.4 voice input system and Interface Strategy

To support natural and low-friction voice interaction, the system uses a “click-to-start + silence detection” input mechanism. Users initiate recording by tapping a central button,

and the system automatically stops recording after two seconds of silence, minimizing manual effort and enabling accurate segmentation for backend processing. The user interface adopts a minimalist and status-aware design. A single, centrally located button changes dynamically to reflect interaction states—for example, “Start” before recording and “Listening...” during input. This provides clear feedback across all interaction stages, helping users understand the system’s status and maintain control. This design ensures accessibility for users with diverse backgrounds and supports a seamless, voice-first consultation experience.

5.5 Module Structure and Technology Selection Research Adaptability Description

The system prototype was developed using the Unity engine, chosen for its support of 3D virtual avatar models, real-time rendering, and flexible animation control. All avatars adopt the HDRP pipeline for consistent visual quality, with animation controlled via Unity’s Animator and Timeline systems for gesture and lip-sync coordination. The voice interaction system integrates Unity’s built-in module with a silence detection mechanism, enabling smooth, pause-sensitive speech input. Language understanding and response generation are handled by GPT models, with identical dialogue logic applied across all avatar conditions to ensure fairness. Speech synthesis varies by condition: high-anthropomorphism avatars use Azure Text-to-Speech with emotion styles (e.g., cheerful, empathetic), while medium and low use neutral-tone engines to avoid emotional bias. The architecture follows a modular design, enabling independent deployment and orchestration of core components—including speech recognition, dialogue management, semantic extraction, emotion analysis, and avatar control. This structure ensures adaptability, scalability, and reusability for future research or deployment in varied consultation contexts.

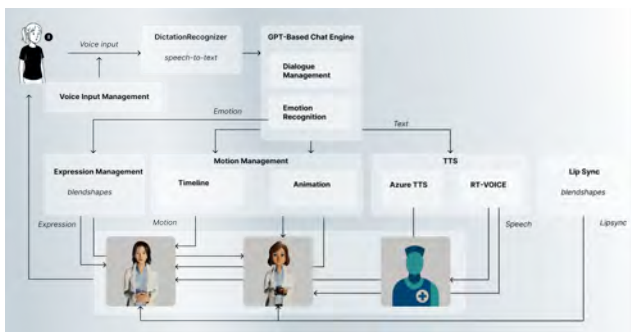


Figure 6. System architecture.

6 Experiment

To assess the impact of the degree of anthropomorphism in virtual avatars on user experience and the quality of medical history information, I designed a user experiment based on

a virtual consultation scenario. The experiment utilized the virtual avatar medical history collection system I developed as its platform, combining a standardized task scenario with system functionality. Participants are invited to engage in a complete online consultation simulation with the virtual avatar. During the experiment, the “degree of anthropomorphism” controlling the virtual avatar’s appearance and behavior was manipulated to observe the effects of different presentation methods on user behavior and attitudes.

6.1 Scenario Setup and Experimental Tasks

The experiment simulated a real online medical consultation scenario. Under system guidance, participants sequentially completed medical history collection tasks with three virtual avatars of different anthropomorphism levels. The tasks included describing symptoms (e.g., headache, cold, etc.) and answering system-generated questions, covering key details such as the onset time, location, severity, accompanying symptoms, and past medical history of the symptoms. The content of the virtual avatar consultation tasks remained consistent across all avatars, with differences only in their appearance and interaction styles to ensure experimental comparability.

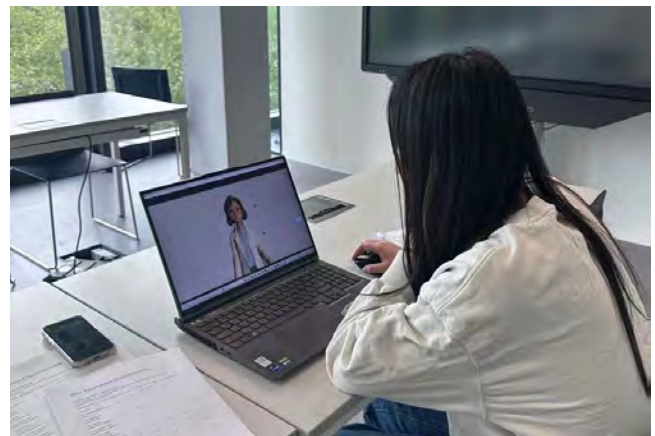


Figure 7. A participant interacting with the virtual consultation system during the experiment.

6.2 Experimental Variables and Grouping Strategy

This experiment employed a single-factor between-subjects design, with the independent variable being the “anthropomorphism level of virtual digital avatars.” Three levels are defined. Dependent variables (Measured Variables) include:

- a Information quality of responses, which include information quantity, specificity, relevance, clarity.
- b User experience

Information quality is assessed through a combination of structured data and human scoring, while user experience is analyzed using standardized questionnaires (BUS-11, Godspeed, UEQ-S) and interviews. This experiment employed

a single-factor within-subject design. Each participant completed the same consultation task across three avatar conditions (low, medium, high anthropomorphism), with brief breaks between rounds to reduce fatigue. Although the presentation order of avatars was not strictly counterbalanced, Kruskal-Wallis H test for order effects should no significant influence on key outcome variables ($p > 0.05$), supporting the validity of the results.

6.3 Experimental Sample Design and Statistical Power Analysis

Sample size estimation was conducted using the G*power 3.1 tool, with an effect size $f = 0.3$, $\alpha = 0.05$, statistical power = 0.80, and a repeated measures one-way ANOVA model. The results indicated that the minimum required sample size was 20 participants. A total of 27 participants are recruited for this study. After screening the data for completeness and validity during the experiment, 25 valid samples are ultimately retained. The participants are primarily college students aged 18–30, all of whom possessed good communication skills and basic digital device operation abilities. All participants signed informed consent forms prior to the experiment and received a standardized experiment task description.

6.4 Experimental Process and Data Collection

The complete experimental process is as follows: 1. Participant Recruitment and Grouping: Participants are recruited through campus channels, completed online registration and eligibility screening, and are randomly assigned interaction sequences. 2. Experiment Task Execution: Each participant independently completed three rounds of virtual doctor consultation tasks. The system automatically recorded all interaction processes, including user input and system responses. After completing each round of tasks, participants are required to fill out a questionnaire. Following the completion of three rounds of simulated consultation tasks, a semi-structured interview was conducted. The entire process took approximately 40 minutes. 3. Data Collection and Recording: Data are collected from multiple sources:

- System logs: All user-system dialogues are saved for later evaluation of response quality (informativeness, specificity, relevance, and clarity).
- Observation notes :Observation notes: Researchers recorded notable user behaviors such as hesitation, repetition, and emotional cues during interaction.
- Questionnaire : After each interaction, participants completed standardized instruments including BUS-11[4], UEQ-S[24], and the Godspeed[14]. BUS-11 assesses chatbot usability, UEQ-S captures both pragmatic and hedonic aspects of user experience, and the Godspeed scale evaluates perceived anthropomorphism and related social traits.
- Interviews : Semi-structured interviews explored participants' subjective impressions, emotional reactions, and suggestions for system improvement.

These data sources enabled both quantitative and qualitative evaluation of user experience and interaction effectiveness.

7 Results

This section combines quantitative questionnaire data with qualitative interview content to analyze the differences in user experience and expression quality among the three types of virtual avatars. By comparing the scoring results with user feedback, key design insights are extracted to provide a basis for optimizing virtual avatar imagery and interaction.

7.1 Quantitative result

7.1.1 Comparison of Performance Differences Among Groups A, B, and C. To investigate differences in social perception among different virtual avatars, a Friedman paired samples test was conducted to statistically analyze the scores of the three groups (A, B, C) across four key dimensions: anthropomorphism, animacy, likeability, and perceived intelligence. The results show:

- Anthropomorphism: There were significant differences among the three avatars, $\chi^2(2) = 32.435$, $p < .001$. Wilcoxon signed-rank tests showed significant pairwise differences between A and B ($p < .001$), A and C ($p < .001$), and B and C ($p = .009$), following the pattern $C > B > A$.
- Animacy :Significant differences were observed, $\chi^2(2) = 34.758$, $p < .001$. Pairwise comparisons revealed significant differences between A and B ($p < .001$), and A and C ($p < .001$), but no significant difference between B and C ($p = .162$).
- Likeability : The Friedman test indicated significant differences, $\chi^2(2) = 18.149$, $p < .001$. Avatar A differed significantly from both B ($p = .006$) and C ($p = .004$), while B and C showed no significant difference ($p = .336$).
- Perceived Intelligence :Significant differences were found, $\chi^2(2) = 7.267$, $p = .026$. A significant difference was found between A and C ($p = .005$), while comparisons between A and B ($p = .077$) and B and C ($p = .249$) were not significant.
- Perceived Safety : The Friedman test did not reach statistical significance, $\chi^2(2) = 5.092$, $p = .078$. However, post-hoc tests showed that A differed significantly from both B ($p = .016$) and C ($p = .019$), while no difference was observed between B and C ($p = .872$).

In summary, Role A scored significantly lower than B and C across all four dimensions, while B and C had similar scores across most dimensions with no statistically significant differences. This result indicates that although the three

digital characters share a unified functional structure, their differences in emotional characteristics and social appearance design are sufficient to significantly influence users' subjective perceptual experiences.

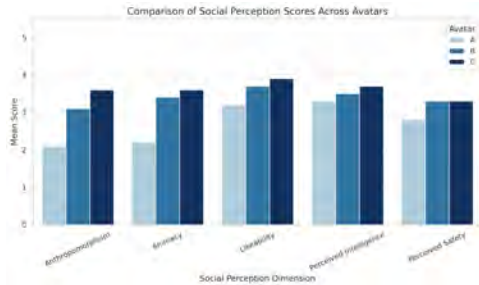


Figure 8. Comparison of social perception scores across three avatars (A, B, C).

7.1.2 Differences in user experience (UX) and usability between different Virtual Avatar characters. To investigate the differences in performance between different Virtual Avatar characters in terms of user experience (UX) and usability (Useability), Friedman paired sample tests and subsequent pairwise comparisons are conducted on the subjective evaluation data of each participant for characters A, B, and C.

- **User Experience Dimension** The results of the Friedman test should significant differences among the three groups of characters in the UX user experience dimension, $\chi^2(2) = 14.990$, $p = .0006$. Further Wilcoxon paired tests indicated that Avatar C achieved significantly higher user experience scores than both Avatar A ($p = 0.0006$) and Avatar B ($p = 0.0032$), while no significant difference was observed between Avatar B and C ($p = 0.47$). The average scores for the three groups are: A ($M = 3.01$, $SD = 0.68$), B ($M = 3.47$, $SD = 0.68$), and C ($M = 3.55$, $SD = 0.63$). These results suggest that Avatar C is perceived more positively overall, though its advantage over Avatar B is not statistically conclusive.
- **satisfaction with the chatbot Dimension** In the satisfaction with the chatbot dimension, the Friedman test did not reveal any significant differences, $\chi^2(2) = 2.571$, $p = 0.213$, indicating that there is no significant differences in users' ratings of the three characters in terms of comprehensibility, operability, and efficiency.

This suggests that Avatar C is superior on the emotional and perceptual dimensions of the user, rather than just being more functional or efficient to interact with.

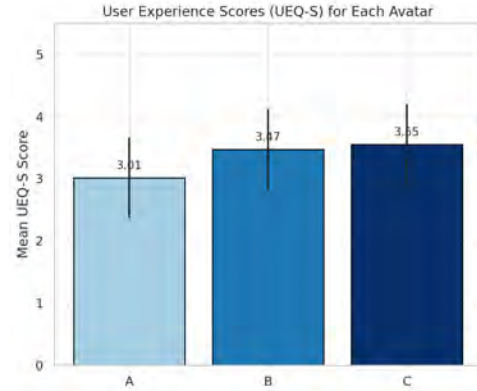


Figure 9. User experience scores (UEQ-S) for each avatar.

7.1.3 The correlation between anthropomorphism and satisfaction with the chatbot and overall user experience. To examine the relationship between the “anthropomorphism” scores of Virtual Avatars and user experience dimensions, this study first conducted a normality test on the distribution of variables. The Shapiro-Wilk test was used to assess the “anthropomorphism scores” and all user experience dimensions (including UX items and satisfaction with the chatbot evaluation items). The results showed that, except for the dependent variable, most independent variables deviated significantly from normal distribution ($p < 0.05$). Given this, to enhance the robustness of the analysis results, the non-parametric Spearman's rank correlation analysis was employed to assess the relationship between anthropomorphism scores and UX/satisfaction items. Personification scores showed a moderately strong positive correlation with users' overall system experience (UX) and are also significantly correlated with satisfaction with the chatbot evaluations, indicating that users who perceive higher system satisfaction are more likely to perceive its personification characteristics. This suggests that the system's “human touch” has an interactive enhancement effect with multiple factors such as “usability,” “pleasure,” and “trust.” More specifically, the results of the Spearman correlation analysis show that “anthropomorphism scores” are significantly positively correlated with multiple user experience dimensions. Among these, attractiveness (Attractive, $r = 0.547$, $p < 0.001$), pleasantness (Pleasant, $r = 0.520$, $p < 0.001$), and innovativeness (Inventive, $r = 0.556$, $p < 0.001$), and willingness to use again (Willing to use again, $r = 0.539$, $p < 0.001$) are the strongest correlated items. Additionally, users' trust in the chatbot, perceived intuitiveness, and perceived efficiency are also significantly positively correlated with anthropomorphism (all $p < 0.01$), indicating that anthropomorphic characteristics of the role may enhance users' overall evaluation of its functionality and emotional experience. A few variables (such as “system response speed” and “ease of understanding chat content”) did not reach significant correlation, suggesting

that these satisfaction sub-dimensions may not be directly influenced by anthropomorphic perception.

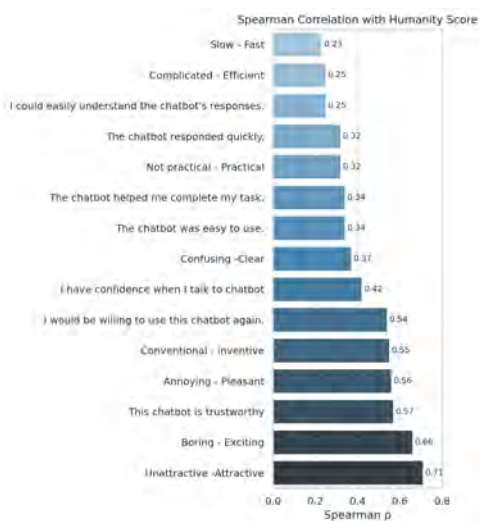


Figure 10. Correlation between perceived anthropomorphism and chatbot evaluation items.

7.1.4 User Response Quality Analysis. This analysis is based on the question-and-answer data from three groups of users, totaling about 800 user responses, and statistically analyzed the performance of responses across four dimensions: information content ($\log_2(\text{Informativeness} + 1)$) processing, subjective specificity, relevance, and clarity.[? ? ? ?] Inter-group nonparametric tests are conducted to assess the impact of different conditions on user expression quality.

To assess the quality of user responses, we adopted a simplified version of the scoring framework from Stock[28]. Each response was evaluated across four dimensions:

- **Informativeness:** Automatically calculated based on the number of structured medical fields successfully filled in each response. We applied a log transformation ($\log(n + 1)$) to reduce the effect of outliers.
- **Relevance, Specificity, Clarity:** Be rated on a 0–2 scale, reflecting how precise and concrete the user’s expression was.

Specificity, relevance, and clarity were manually scored by the author. A total quality score was obtained by summing the four dimensions, resulting in a maximum possible score of 8 per response.

- **Average Response Quality Score (by Group)** The informativeness dimension uses $\log_2(\text{Informativeness} + 1)$ for logarithmic compression to align with the 0–2 interval of the subjective scoring dimension.

Group (x)	Informativeness (log2)	Specificity	Relevance	Clarity
A	5.85	0.92	1.76	1.71
B	6.13	1.06	1.82	1.78
C	6.38	1.21	1.87	1.84

Figure 11. Average Response Quality Score

Group C had the highest scores in all three dimensions (log information content and subjective ratings); Group A had the lowest scores, with significant differences in performance.

- **Normality Test: Shapiro-Wilk Results** The p-values for the Shapiro-Wilk normality test are all far less than 0.05 (almost 0) across all dimensions and groups, indicating:
 - The four-dimensional scoring data does not follow a normal distribution
 - Therefore, parametric methods are abandoned, and more robust non-parametric test methods are adopted
- **Kruskal-Wallis multi-group test results**

Dimension	Kruskal p	significant difference
Informativeness (log2)	0.0036	Yes
Specificity	0.0021	Yes
Relevance	0.0053	Yes
Clarity	0.0037	Yes

Figure 12. Kruskal-Wallis multi-group test results

This indicates that there are statistically significant differences among the three groups across all dimensions.

- **Mann-Whitney two-group comparison test results**

Comparison Groups	Dimension	p	significant difference
A vs B	Informativeness	0.0732	No
A vs B	Specificity	0.1074	No
A vs B	Relevance	0.1054	No
A vs B	Clarity	0.0791	No
A vs C	Informativeness	0.0008	Yes
A vs C	Specificity	0.0011	Yes
A vs C	Relevance	0.0022	Yes
A vs C	Clarity	0.0015	Yes
B vs C	Informativeness	0.0893	No (marginally)
B vs C	Specificity	0.0752	No (marginally)
B vs C	Relevance	0.0985	No (marginally)
B vs C	Clarity	0.0804	No (marginally)

Figure 13. Mann-Whitney two-group comparison test results

The results show that Group C significantly outperforms Group A, demonstrating the best performance; Group A exhibits a clear disadvantage compared to the other two groups. Although there is no significant difference between Groups B and C, marginal differences are observed across all four metrics (p-values ranging from 0.07 to 0.1), suggesting that Group C users may have a slight advantage in terms of expression.

7.2 Qualitative results Thematic Analysis of Interviews

7.2.1 Theme 1: Character Style Shapes Trust and Tone.

Participants consistently linked avatar appearance to how seriously they engaged in the consultation. The highly anthropomorphic avatar was often described as “doctor-like,” leading many to adopt a more formal, structured tone when speaking. One user explained, “The third character looks most like a doctor, so I speak to her most seriously” (P24). This avatar’s professional image appeared to promote cognitive framing of the task as a real medical consultation.

However, over half the participants noted discomfort related to this avatar’s facial expressions or timing. Several described the mouth movements and voice as “out of sync,” evoking terms like “robotic,” “mannequin-like,” or “uncanny.” One user remarked, “It’s too realistic, but something’s off—it made me nervous” (P22).

By contrast, the medium (cartoon) avatar received the most consistent praise. Twenty-one users called it “approachable,” “non-threatening,” or “easy to talk to,” citing its relaxed style and friendly gestures. Static icons were widely considered disengaging, with multiple users comparing them to “speaking to a wall.”

This suggests that users favor avatars that strike a balance between credibility and emotional comfort—supporting a middle-ground approach in avatar realism.

7.2.2 Theme 2: Feedback Behaviors Influence Expression Willingness.

A strong recurring theme was the importance of dynamic, responsive feedback in creating a sense of being heard. Participants mentioned gestures like nodding, blinking, or brief affirmations (e.g., “I see”) as cues that encouraged further disclosure. “The cartoon avatar nodded when I spoke—I felt more comfortable going on,” said one user (P03). Such nonverbal signals appeared to lower inhibition and sustain conversational flow.

Conversely, feedback that was delayed, inconsistent, or mismatched with speech (e.g., emotionless facial expression while saying “I understand”) undermined trust. These issues were most frequently reported for the highly realistic avatar, with participants expecting greater behavioral consistency due to its lifelike appearance. As one user noted, “The face didn’t move at all while talking. That made me feel like it wasn’t really listening” (P17).

Mid- and low-fidelity avatars elicited fewer expectations in this regard, leading to greater tolerance of limited feedback. Still, users consistently rated avatars higher when they exhibited timely, coordinated responses.

This highlights the importance of multimodal behavior alignment as a key driver of perceived responsiveness and expression quality.

7.2.3 Theme 3: Rhythm and Interaction Flow Affect Comfort.

Many users identified interaction rhythm—including

system pace, pause recognition, and transition timing—as central to a natural consultation experience. Several participants expressed frustration with being cut off mid-sentence or not receiving cues when to speak. “It skipped ahead before I was done,” said one user (P20). Another added, “Sometimes I just didn’t know if it was still listening or had already moved on” (P19).

Short but natural pauses for reflection were often misinterpreted by the system as the end of a turn, resulting in premature transitions. These disruptions not only made users feel rushed, but also reduced the completeness of their responses.

Users generally preferred systems that provided explicit conversational cues—such as slight delays, backchanneling gestures, or summary confirmations—to help them gauge timing. Highly anthropomorphic avatars were particularly prone to breaking immersion when their rhythm didn’t match the conversational flow.

This suggests that interaction timing should be explicitly managed at both the system and avatar levels to ensure natural pacing and reduce communication friction.

7.2.4 Theme 4: Emotional Safety Supports Disclosure.

The emotional tone projected by the avatar was found to significantly influence users’ willingness to share sensitive or personal details. Especially when participants reported feeling unwell or anxious, they preferred avatars with a softer demeanor. The cartoon avatar was widely praised for being “gentle,” “friendly,” and “not judgmental.” One user shared, “When I feel sick or low, I don’t want a serious face looking at me—I want something warm” (P08).

Highly realistic avatars occasionally triggered discomfort due to overly intense eye contact or lack of warmth in expression. “It kept staring at me—I didn’t like that,” said one participant (P35). Meanwhile, the static avatar was perceived by some as emotionally neutral, and by others as simply disconnected.

A few participants expressed a desire to choose the avatar style based on their mood or consultation type, suggesting the value of customizable options.

This indicates that emotional comfort—not visual fidelity—may be the most important factor in promoting self-disclosure in healthcare interactions.

7.3 Insights and Design Inspirations

This project focuses on the impact of the degree of anthropomorphism in virtual avatars on user experience and expressive behavior. Through a systematic analysis of user interview data, I have identified four key themes: user preferences, behavioral feedback, task collaboration, and technological evolution. These themes are discussed in the context of real-world usage scenarios, and targeted design recommendations are summarized. These insights not only provide theoretical support for the design of anthropomorphic virtual avatars

but also offer practical guidance for future product iterations and role-based functional divisions.

7.3.1 Insight 1: Anthropomorphism Should Be Treated as a Multi-Dimensional and Context-Sensitive Design Strategy. This study reveals that avatar anthropomorphism is not a linear spectrum from low to high. Rather, it comprises multiple intersecting dimensions—visual realism, behavioral coordination, emotional expression, and perceived social role. Users do not respond to anthropomorphism in isolation, but to how these cues work together within the interaction context. High realism may convey professionalism but can break immersion if behavioral coordination is lacking. Moderate anthropomorphism enhances comfort but may reduce task seriousness. Therefore, anthropomorphism should not be seen as a fixed design choice, but as a flexible, adaptive strategy that evolves with user profiles, task phases, and interaction goals

7.3.2 Insight 2: Behavioral Feedback Consistency Shapes Perceived Responsiveness. Across all avatar types, the strongest predictor of trust and expressive willingness was behavioral synchrony—whether voice, gesture, and facial feedback aligned in timing and tone. Users did not necessarily care about visual fidelity but are highly sensitive to the perceived presence and “listening quality” of the avatar.

7.3.3 Insight 3: Expressiveness is not just about comfort, but characterisation is also key. While friendly avatars can ease users’ nerves, they don’t necessarily lead to more complete or valuable clinical information. Some users admitted that they tend to be more relaxed and casual when confronted with cartoon avatars, while they are more serious and articulate when the avatar looks more like a real doctor.

This illustrates the importance of characterization: users subconsciously adjust their expressions according to the characterization conveyed by the avatar - whether it is a “doctor”, “friend”, “voice assistant”, “friend”, “friend”, “friend”, or “voice assistant”. or ‘voice assistant’.

7.3.4 Insight 4: Emotional temperature is an independent and important design dimension in health scenarios. In situations of discomfort, anxiety or vulnerability, users particularly value the ability of an avatar to convey warmth and empathy. It is worth noting that this ‘emotional security’ does not come from the realism of the appearance, but from the details of the design, such as a soft tone of voice, gentle eyes, and a relaxed and non-oppressive posture.

Many users mentioned that their ideal digital persona would be ‘friendly’, “supportive” or ‘reassuring’ - even if they knew it was an AI. -even if they knew it was AI.

This further suggests that emotional warmth does not depend on whether an image is realistic or not, but is a separate dimension that deserves specialised design.

7.3.5 Design recommendations. Based on the above insights, the following design implications and recommendations are provided:

- Anthropomorphism is not a level to be chosen, but a dynamic composition to be tuned.
- Behavioral synchrony is the foundation of perceived presence and trust.
- Visual framing shapes how users position themselves in the interaction.
- Emotional warmth is not a byproduct of realism, but a deliberate design choice.
- virtual avatars should adapt their form and function to match the phase and emotional tone of the task.

8 Final Design

Based on the experimental results and design insights, I gradually optimized the product during the design process to address these issues, ultimately creating a usable, realistic, and deployable virtual avatar medical history consultation system prototype.

8.1 Design Iteration Focus

In response to the user interviews in the previous chapter, I identified multiple pain points related to interaction and trust, and the iterations will focus on these key dimensions. I focused on three core objectives: enhancing the naturalness of the conversation; strengthening user trust; and optimizing the interface and interaction rhythm. The system design was improved in the following areas:

8.1.1 Interaction Rhythm and Process Prompts. To avoid users feeling abrupt at the beginning of the conversation, the system uses natural small talk as an opening. The voice input method is also optimized, allowing users to start and end conversations without clicking buttons. And an automatic mute detection mechanism is introduced: when users stop speaking for a certain period of time, the system automatically ends voice input, enhancing the natural rhythm of voice interaction. Additionally, the system provides diverse feedback statements when recognition fails, reducing frustration and mechanical feel.

8.1.2 Feedback Mechanisms and Information Confirmation. Users prefer more transparent feedback and clearer confirmation of understanding. Therefore, I designed dialogue information display and natural language repetition confirmation to help users understand which information the system has “understood,” thereby enhancing users’ trust in the system’s judgments. The system also supports displaying a unified field confirmation page before the consultation ends, improving information transparency and a sense of control.

8.1.3 Emotional Perception and virtual avatar Responses. Multiple users expect the system to provide understanding

and response when they express discomfort or negative emotions. Based on this, I have enhanced emotion recognition and emotional feedback mechanisms that include both linguistic and non-linguistic cues, making the virtual avatar more empathetic.

8.1.4 Privacy and Role Identity Prompts. Considering that some users may have sensitive concerns about being monitored and the use of their data, the system clearly states the role identity and scope of data usage at the beginning, and offers two types of Virtual Avatar characters—low-fidelity and high-fidelity—to accommodate different users' acceptance thresholds for anthropomorphism and privacy exposure.

8.2 Final System Design

8.2.1 virtual avatar Performance and Humanized Feedback. The system integrates two styles of virtual avatars to accommodate users' varying preferences for anthropomorphism. The high-anthropomorphism version is based on HDRP high-quality models, featuring voice emotional expression, eye contact, and nodding, among other facial and body movements, to create a more immersive companion-like experience for users; the medium-anthropomorphism version uses stylized characters with a friendly and lively style, suitable for users seeking stress relief or those sensitive to photorealistic Virtual Avatars. The system incorporates a semantic-level emotion recognition mechanism, utilizing emotion recognition APIs to automatically assess users' emotional tendencies based on their spoken content. It then provides appropriate voice style feedback and triggers corresponding facial expressions and movements, such as nodding, frowning, or gazing. This multi-channel feedback approach enhances the virtual avatar's ability to "understand" users.



Figure 14. human-like character motions



Figure 15. Stylised character motions

8.2.2 Interface Design and voice input system. In interface design, the system adheres to minimalist principles, emphasizing clear operations, transparent information, and visual accessibility. The entire consultation interface is highly intuitive, driven by voice commands, eliminating the need for frequent button clicks. Users can respond directly after the virtual avatar poses a question, and the system automatically detects the start and end of voice input, providing rapid feedback upon completion, typically within approximately one second. Additionally, to enhance the perceptibility of voice interaction, the system incorporates a subtitle synchronization display mechanism. During user speech, the system generates real-time text from voice recognition and presents it in bubble subtitle format, helping users confirm whether the system accurately understood their statements. Considering that some users may find bubble subtitles distracting or disruptive to the conversation experience, users can choose whether to enable subtitles. Furthermore, the system incorporates dynamic progress indicators, using a progress bar to display the current status of the consultation process, thereby enhancing predictability of the overall workflow. At the conclusion of the process, the system displays a field confirmation page to guide users through final verification and confirmation, ensuring data completeness and user satisfaction. This mechanism not only enhances user engagement but also provides reliable assurance for subsequent data output and utilization. Different versions all feature stable voice response mechanisms and basic feedback functions in their interaction performance, while maintaining consistent UI presentation visually, ensuring consistency and switchability in the user experience.



Figure 16. User interface sequence of the virtual consultation system

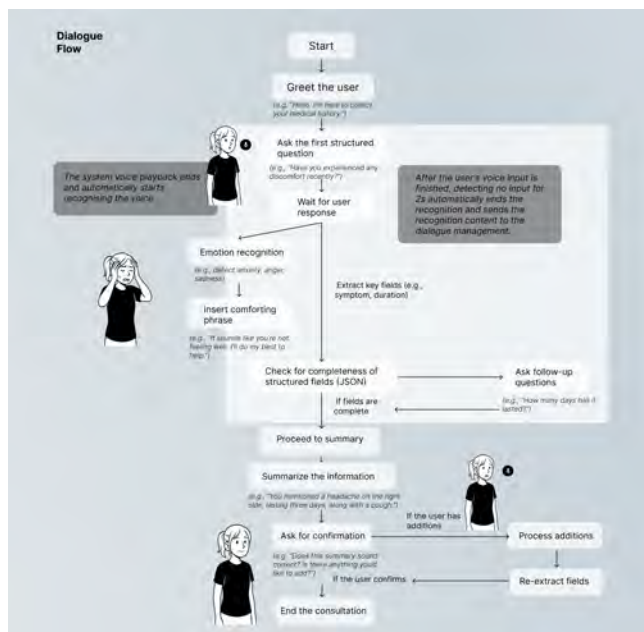


Figure 17. Dialogue flow with structured field extraction and emotional feedback.

8.2.3 System Response and Deployment Feasibility.

Although the core focus of this study is on user interaction and experience design, to ensure consistency in system behavior and experimental controllability, I have constructed a complete technical closed-loop system, with the specific architecture shown in the figure below: The system runs on the Unity platform, leveraging efficient collaboration among multiple functional modules to complete the entire workflow of voice input, language understanding, virtual avatar response, field extraction, and data export. The system architecture is modular and clear, facilitating future customization and expansion based on device performance or usage scenarios (e.g., tablet devices, remote consultations, etc.). The overall response is smooth, with voice recognition and generation latency controlled within 1–1.5 seconds, making it suitable for medical scenario experiments requiring rapid response. The voice synthesis module supports changes in tone and style, while the language understanding module can output structured JSON-formatted fields, facilitating integration with subsequent systems.

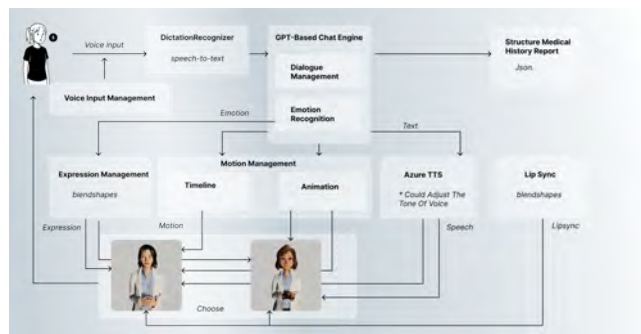


Figure 18. Final architecture of the consultation system with structured output generation.

8.3 Evaluation

I invited 24 users to evaluate the user experience of the final design. The system was evaluated using the simplified version of the UEQ (UEQ-S, User Experience Questionnaire - Simplified), which consists of two dimensions: pragmatic quality and hedonic quality to measure the system’s usability and emotional appeal, respectively. Participants completed the questionnaire immediately after interacting with the system. Ratings are made on a 5-point scale (1 = strongly disagree, 5 = strongly agree). The test results should that the mean quality scores for both dimensions are above 4.0, with a utility quality score of 4.25 and a hedonic quality score of 4.30, indicating that the system was positively evaluated by users in terms of both functional usability and emotional experience. The final design effectively achieves the expected user experience goals and lays a solid foundation for the subsequent application and promotion of the system.

9 Discussion

9.1 Theoretical Contributions: How Anthropomorphic Design Influences User Expressive Behavior

This study employs the “Research through Design” (RtD) methodology and finds that virtual characters with higher levels of anthropomorphism are more likely to stimulate users’ willingness to express themselves, thereby enhancing their level of seriousness and the completeness of information in communication. This finding aligns with the Cognitive-Affective-Social Interaction (CASA) theory and the Media Equation theory, which suggest that users instinctively perceive “human-like” systems as social interaction partners, thereby altering their expression methods, emotional investment, and trust levels. It is worth emphasizing that in the healthcare context, which heavily relies on trust and information authenticity, moderate anthropomorphism not only helps establish rapport but also enhances the system’s professionalism and credibility, laying the foundation for more effective data collection and health assessments.

9.2 Design Reflection: Balancing Professionalism and Social Pressure

User feedback indicates that highly anthropomorphic characters do not always lead to positive experiences. On one hand, their lifelike voices and appearances enhance the immersion and professionalism of communication; on the other hand, overly realistic expressions or gaze behaviors may cause user discomfort or even inhibit expression. This phenomenon suggests that anthropomorphic design should aim for “just right” rather than “as realistic as possible.” An ideal virtual avatar should strike a balance between warmth and pressure, conveying trust and professionalism without overly infringing on users’ psychological comfort zones. virtual avatars with “style adjustment capabilities” will be able to adapt to different user characteristics and task contexts, dynamically adjusting tone, expressions, speech rate, and eye contact to provide a more comfortable and natural interaction experience.

9.3 System Limitations and Improvement directions

Although this study has achieved preliminary results in system implementation and user testing, it also has some limitations: First, the system has not yet fully overcome technical bottlenecks in terms of performance. Currently, the voice, expressions, and movements of virtual Virtual Avatars are primarily controlled through a combination of predefined animations, script logic, and emotion recognition results. Although the system has basic emotion recognition and tone adaptation capabilities, enabling it to partially utilize matching voice tones and facial expressions, its overall performance still relies on limited predefined resources. The system has not yet achieved truly semantically driven dynamic behavior generation and natural coordination, leading to user perceptions such as “natural voice but stiff movements” and “lack of eye contact” in highly anthropomorphic characters. This indicates that the system still struggles to achieve a truly unified “human-like” experience between visual presentation and behavioral performance. Secondly, the system’s interactive adaptability remains insufficient. Although it possesses basic emotion analysis capabilities, it is still unable to continuously perceive users’ tone of voice, language rhythm, or non-verbal feedback, and adjust the Virtual Avatar’s behavioral style or response strategy in real time accordingly. This makes some interactions feel stiff or inflexible, limiting users’ expressive space and the comfort of interaction. More importantly, the sample size in this experiment was limited, with participants primarily concentrated in specific demographics (such as college students and young users), resulting in relatively homogeneous age structures, cultural backgrounds, and actual medical experiences. This sample composition restricts the external applicability of the research conclusions. For example, different age groups, cultural groups, or real patients may have varying degrees of

acceptance and expression strategies when encountering anthropomorphic characters. Therefore, future research should expand in two directions: on the one hand, continue to enhance the system’s capabilities in multimodal coordination, contextual adaptation, and behavioral consistency; on the other hand, expand the sample coverage to include more authentic and diverse user groups to validate the universality and effectiveness of anthropomorphic strategies in complex medical contexts. These observations also suggest that the current system and design strategies are still in their infancy and require further integration of AI capabilities with user feedback to expand the design space.

9.4 Future Possibilities for Virtual Avatars: The Potential Role and Challenges of AI Generation Technology

Although this study did not directly employ generative AI technologies, the technical limitations encountered—such as rigid facial expressions, unsynchronized gestures, and limited adaptive feedback—highlight the current bottlenecks in building truly human-like virtual agents[7]. These are precisely the areas where generative technologies may play a transformative role in the near future. Advancements in multimodal large language models, real-time emotional voice synthesis, and generative facial animation are paving the way for Virtual Avatars that can dynamically align speech content with appropriate tone, facial expressions, and gestures. Such capabilities would enable virtual agents to respond not only with accurate information but also with social-emotional coherence, greatly enhancing realism and user engagement.[? ?] In this view, generative AI may represent the next step in the evolution of avatars - transforming them from ‘by-the-book’ information transmitters to emotionally responsive intelligent companions in healthcare interactions.

However, this transition comes with significant challenges:

- **Risk of trust and miscalculation:** As avatars become more and more realistic, users may overestimate their expertise or even mistake them for real healthcare practitioners. In this context, it becomes particularly important to clarify role definition and communication boundaries to prevent over-reliance in clinical situations.
- **Aligning functional goals with expressive style:** The generative and expressive capabilities of avatars should serve the core task of helping to achieve clear, organized, and empathetic communication, rather than merely pursuing a ‘human-like’ appearance. In healthcare, accuracy and professionalism of information always come first.
- **Ethical and adaptive design considerations:** Freely generated content must be subject to rigorous controls, especially when dealing with sensitive or

emotional dialogue. Future systems should be context-aware, able to flexibly adjust tone and presentation according to different users and scenarios.

In summary, generative AI cannot replace thoughtful design work, but it can allow avatars to behave more naturally, consistently, and with greater emotional responsiveness. If combined with clear ethical boundaries and purpose-driven interaction design, it will lay a solid foundation for the next generation of trusted, intelligent virtual healthcare agents while breaking through current limitations.

10 Conclusion

This study investigated how different levels of avatar anthropomorphism affect user experience and response quality during medical history taking. By comparing three avatar styles—abstract icons, stylized figures, and highly realistic models—the research systematically examined differences in behavioral interaction, emotional engagement, and communication outcomes.

In response to RQ1, results suggest that virtual avatars with a high degree of anthropomorphisation have a tendency to enhance the quality of user responses in AI-driven questioning. Participants were more expressive and better at using complete sentences and detailed descriptions when confronted with more humanised characters. Regarding RQ2, user experience also improved with the level of anthropomorphisation, particularly in terms of emotional engagement, trust [3][7] and empathy perception. However, overly realistic avatars may trigger user discomfort in cases of poor behavioural coordination, instead of affecting natural expression. Therefore, moderate anthropomorphisation is more effective in achieving a good user experience than simply pursuing a high degree of realism.

These findings emphasise that avatar design should go beyond the pursuit of physical realism and be more flexible in integrating social cues to respond to the cognitive and emotional needs of users. [22][19] Anthropomorphism should be viewed as a task-oriented, adaptive and socially meaningful interaction strategy rather than a fixed choice of visual style.

Methodologically, this study combines user testing and system prototyping; conceptually, a three-component structural-behavioural-emotional incarnation design framework is proposed; and practically, functional prototypes for healthcare scenarios are constructed. These results provide theoretical support and practical experience for building more adaptive and human-centred digital health interfaces in the future.

With the continuous progress of AI in behaviour generation, voice interaction and personalised regulation, avatars will increasingly have stronger expressiveness, intelligence and emotional responsiveness. The adaptive anthropomorphisation strategy proposed in this study is not only applicable to medical consultation, but also has the potential for

wide application in digital therapy, mental health support and health education, which indicates that the virtual avatar is shifting from a functional tool to an intelligent partner with cognitive support and social awareness.

References

- [1] Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, and Ajay Rana. 2020. Chatbot for healthcare system using artificial intelligence. In *2020 8th International conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO)*. IEEE, 619–622.
- [2] Jeremy N Bailenson, Andrew C Beall, Jack Loomis, Jim Blascovich, and Matthew Turk. 2005. Transformed social interaction, augmented gaze, and social influence in immersive virtual environments. *Human communication research* 31, 4 (2005), 511–537.
- [3] Timothy Bickmore and Toni Giorgino. 2006. Health dialog systems for patients and consumers. *Journal of biomedical informatics* 39, 5 (2006), 556–571.
- [4] Simone Borsci and Martin Schmettow. 2024. Re-examining the chatBot Usability Scale (BUS-11) to assess user experience with customer relationship management chatbots. *Personal and Ubiquitous Computing* 28, 6 (2024), 1033–1044.
- [5] Jenni Burt, John Campbell, Gary Abel, Ahmed Aboulghate, Faraz Ahmed, Anthea Asprey, Heather Barry, Julia Beckwith, John Benson, Olga Boiko, et al. 2017. Improving patient experience in primary care: a multimethod programme of research on the measurement and improvement of patient experience. *Programme Grants for Applied Research* 5, 9 (2017), 1–452.
- [6] Niezen, G., Van der Vlist, B. J., Hu, J., & Feijs, L. M. (2010). From events to goals: Supporting semantic interaction in smart environments. In conference; IEEE Symposium on Computers and Communications (ISCC), 2010, Riccione, Italy; 2010-06-22; 2010-06-25 (pp. 1029-1034). Institute of Electrical and Electronics Engineers.
- [7] Oscar Hengxuan Chi, Christina G Chi, and Dogan Gursoy. 2025. Seeing Personhood in Machines: Conceptualizing Anthropomorphism of Social Robots. *Journal of Service Research* 28, 1 (2025), 78–92.
- [8] Oscar Hengxuan Chi, Christina G Chi, and Dogan Gursoy. 2025. Seeing Personhood in Machines: Conceptualizing Anthropomorphism of Social Robots. *Journal of Service Research* 28, 1 (2025), 78–92.
- [9] Xu, W., Kreijns, K., & Hu, J. (2006, April). Designing social navigation for a virtual community of practice. In *International Conference on Technologies for E-Learning and Digital Entertainment* (pp. 27-38). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [10] Ahmed Fadhil and Gianluca Schiavo. 2019. Designing for health chatbots. *arXiv preprint arXiv:1902.09022* (2019).
- [11] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.
- [12] Gloria R Grice, Nicole M Gattas, Theresa Prosser, Mychal Voorhees, Clark Kebodeaux, Amy Tiemeier, Tricia M Berry, Alexandria Garavaglia Wilson, Janelle Mann, and Paul Juang. 2017. Design and validation of patient-centered communication tools (PaCT) to measure students' communication skills. *American Journal of Pharmaceutical Education* 81, 8 (2017), 5927.
- [13] Chin-Chang Ho and Karl F MacDorman. 2010. Revisiting the uncanny valley theory: Developing and validating an alternative to the God-speed indices. *Computers in Human Behavior* 26, 6 (2010), 1508–1518.
- [14] Chin-Chang Ho and Karl F MacDorman. 2010. Revisiting the uncanny valley theory: Developing and validating an alternative to the God-speed indices. *Computers in Human Behavior* 26, 6 (2010), 1508–1518.
- [15] Hu, J., Le, D., Funk, M., Wang, F., & Rauterberg, M. (2013, July). Attractiveness of an interactive public art installation. In *International Conference on Distributed, Ambient, and Pervasive Interactions* (pp. 430-438). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [16] Dennis Lee, Nicolette de Keizer, Francis Lau, and Ronald Cornet. 2014. Literature review of SNOMED CT use. *Journal of the American Medical Association* 311, 12 (2014), 1253–1254.

Informatics Association 21, e1 (2014), e11–e19.

- [17] Joseph H Levenstein, Eric C McCracken, Ian R McWhinney, Moira A Stewart, and Judith B Brown. 1986. The patient-centred clinical method. 1. A model for the doctor-patient interaction in family medicine. *Family practice* 3, 1 (1986), 24–30.
- [18] Masahiro Mori. 2012. The uncanny valley. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. Reprint of 1970 original.
- [19] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56, 1 (2000), 81–103.
- [20] Masayuki Nigo, Hong Thoai Nga Tran, Ziqian Xie, Han Feng, Bingyu Mao, Laila Rasmay, Hongyu Miao, and Degui Zhi. 2022. PK-RNN-V E: A deep learning model approach to vancomycin therapeutic drug monitoring using electronic health record data. *Journal of Biomedical Informatics* 133 (2022), 104166. doi:10.1016/j.jbi.2022.104166
- [21] Kristine L Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 12, 5 (2003), 481–494.
- [22] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.
- [23] Scott Robertson, Rob Solomon, Mark Riedl, Theresa Wicklin Gillespie, Toni Chociemski, Viraj Master, and Arun Mohan. 2015. The visual design and implementation of an embodied conversational agent in a shared decision-making context (eCoach). In *Learning and Collaboration Technologies: Second International Conference, LCT 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings 1*. Springer, 427–437.
- [24] Martin Schrepp, Andreas Hinderks, et al. 2017. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). (2017).
- [25] Anna Stock, Stephan Schlögl, and Aleksander Groth. 2023. Tell me, what are you most afraid of? Exploring the effects of agent representation on information disclosure in human-chatbot interaction. In *International Conference on Human-Computer Interaction*. Springer, 179–191.
- [26] Joseph C Ugrin and J Michael Pearson. 2013. The effects of sanctions and stigmas on cyberloafing. *Computers in Human Behavior* 29, 3 (2013), 812–820.
- [27] Jinchun Wu, Xiaoxi Du, Yixuan Liu, Wenzhe Tang, and Chengqi Xue. 2024. How the Degree of Anthropomorphism of Human-like Robots Affects Users' Perceptual and Emotional Processing: Evidence from an EEG Study. *Sensors* 24, 15 (2024). doi:10.3390/s24154809
- [28] Darian Zamanzadeh, Mitchell Wood, Srjana Srivastava, Amy Ogan, Justine Cassell, and Collin Richey. 2023. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-Ended Questions. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 7. ACM, 1–28.
- [29] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 493–502.

Appendix A: Questionnaire Items

.1 BUS-11

- The chatbot was easy to use.
- The chatbot is trustworthy.
- The chatbot responded quickly.
- I have confidence when I talk to chatbot
- I could easily understand the chatbot's responses.
- The chatbot helped me complete my task.

- I would be willing to use this chatbot again.
- The information provided by the chatbot was relevant.
- Interacting with the chatbot was pleasant.
- The chatbot's interface was intuitive.
- The chatbot did not confuse me during the interaction.
- The chatbot offered the functionalities I expected.
- I felt the interaction with the chatbot was efficient.

.2 Godspeed

- Fake-Natural
- Machinelike-Humanlike
- Unconscious-Conscious
- Artificial-Lifelike
- Moving rigidly-Moving elegantly
- Dead-Alive
- Stagnant-Lively
- Mechanical-Organic
- Inert-Interactive
- Apathetic-Responsive
- Dislike-Like
- Unfriendly-Friendly
- Unkind-Kind
- Unpleasant-Pleasant
- Awful-Nice
- Incompetent-Competent
- Ignorant-Knowledgeable
- Irresponsible-Responsible
- Unintelligent-Intelligent
- Foolish-Sensible
- Anxious-Relaxed
- Calm-Agitated
- Still-Surprised

.3 UEQ-S

- Boring-Exciting
- Confusing-Clear
- Conventional-inventive
- Not practical-Practical
- Unattractive-Attractive
- Complicated-Efficient
- Annoying-Pleasant
- Slow-Fast

Appendix B: Interview Guide

We'd like to briefly talk (5–10 minutes) about your experience using the virtual consultation system. There are no right or wrong answers—please feel free to share your thoughts and feelings.

- Can you describe how the entire interaction process felt for you?
- How did the virtual person's appearance, voice, or movements influence your willingness to talk?

- In your opinion, in what ways did the virtual human seem “human-like”? And in what ways did it still feel artificial or machine-like?
- Today, you interacted with one version of the virtual human. Now thinking about the three versions (with different levels of realism and expressiveness), how do you think these differences might affect your willingness to interact or how natural the conversation feels?
- Imagine this system was really implemented in hospitals or health apps—would you feel comfortable using it? Why or why not?
- Is there anything you think could be improved to enhance the experience? This could be anything—how it speaks, the speed, how it asks questions, or even how it looks.
- You experienced one version of the virtual human today. If you had to choose between them, which one would you prefer for a health consultation, and why do you think that version would work better for you?
- Is there anything else you’d like to share about your experience—something we didn’t ask, but you think is important?