Exploring the Impact of Avatar Representations in AI Chatbot Tutors on Learning Experiences

Chek Tien Tan
Centre for Immersification
Singapore Institute of Technology
Singapore, Singapore
chektien.tan@singaporetech.edu.sg

Indriyati Atmosukarto
Infocomm Technology
Singapore Institute of Technology
Singapore, Singapore
indriyati@singaporetech.edu.sg

Budianto Tandianus
Centre for Digital Enablement
Singapore Institute of Technology
Singapore, Singapore
budianto.tandianus@singaporetech.edu.sg

Songjia Shen
Centre for Immersification
Singapore Institute of Technology
Singapore, Singapore
songjia.shen@singaporetech.edu.sg

Steven Wong
Centre for Digital Enablement
Singapore Institute of Technology
Singapore, Singapore
steven.wong@singaporetech.edu.sg

Abstract

Despite the growing prominence of Artificial Intelligence (AI) chatbots used in education, there remains a significant gap in our understanding of how interface design elements, particularly avatar representations, influence learning experiences. This paper explores the impact of different AI chatbot avatar representations on students' learning experiences through a mixed-methods within-subjects study, where participants interacted with three distinct types of AI chatbot interfaces with a common large language model (LLM) over a 14-week university course. Our findings reveal that preferences vary according to factors such as learning habits and learning activities. Avatar design also exhibits affordances for specific prompting behaviors, while the perceived human touch influenced learning experiences in nuanced ways. Additionally, real-world relationships with the individuals behind deepfakes influence these experiences. These insights suggest that the thoughtful integration of diverse avatar representations in AI chatbot systems for different learners and settings can greatly enhance learning experiences.

CCS Concepts

• **Human-centered computing** → **Interaction paradigms**; **Empirical studies in interaction design**.

Keywords

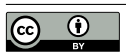
Chatbots, conversational agents, large language models, avatars

ACM Reference Format:
Chek Tien Tan, Indriyati Atmosukarto, Budianto Tandianus, Songjia Shen, and Steven Wong. 2025. Exploring the Impact of Avatar Representations in AI Chatbot Tutors on Learning Experiences. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3706598.3713456

1 Introduction

From early conversational programs like ELIZA [50] to modern Artificial Intelligence (AI) chatbots powered by large language models (LLMs) like ChatGPT [8], such systems have made significant strides in providing personalized learning experiences and on-demand assistance to students. The integration of AI chatbots in educational settings has shown promise in enhancing student engagement, motivation, and learning outcomes [39].

Despite these advances, there remains a gap in understanding how different interface design elements, specifically visual and auditory representations of chatbot avatars, impact learning experiences (Sec. 2). In research, prior work on text-based chatbots showed mimicking human-like interactions may foster believability and engagement [12], for example, through realistic human representations [45] and synthetic non-human characters [35]. The importance of agent design in enhancing learning is also emphasized in pedagogical agent (PA) literature [20, 25]. Meanwhile in the industry, OpenAI has moved from text-only interfaces to the recent GPT-4o model that focused strongly on interfacing with users through natural human-like audio and visual cues on mobile devices [33]. Conversely, research in conversational agents has also highlighted that users often desire human-agent conversations to be transactional and utilitarian [13]. Specific to learning, it is also established that the "decision to include a PA in a computer-supported learning environment is a non-trivial one" [25, p.308]. These varied perspectives highlight the need for deeper investigation into how avatar design — whether more human-like, fictional, or purely functional — affects users' learning experiences, particularly in educational contexts. As the use of AI tools become more prevalent in education, understanding how interface elements like avatars influence experiences in learning can inform more effective design strategies.



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713456

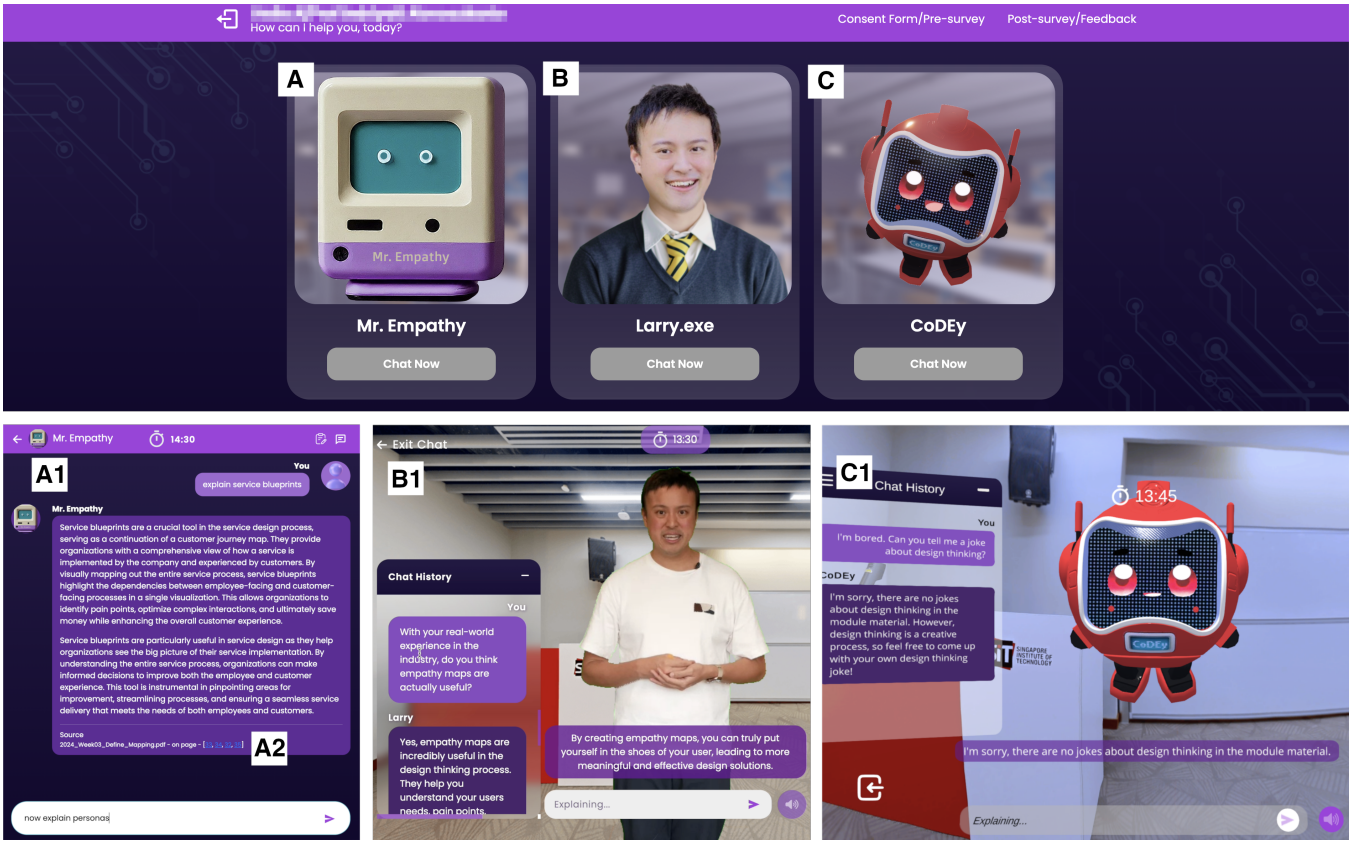


Figure 1: The AI Tutor platform built for the study, with three distinct avatar representations shown in the selection screen on top image: (A) regular text-based interface [NOAVATAR], (B) video-based real-time deepfake avatar [DEEPFAKE], and (C) neutral non-human 3D character [MASCOT]. The respective interactive chat interfaces for (A1) NOAVATAR, (B1) DEEPFAKE, and (C1) MASCOT are shown in the bottom image row. Referenced course documents can be directly accessed from any mode (A2), opening to the specific locations within the document.

This paper hence aims to investigate how various visual and auditory representations of avatars in chatbot AI tutors influence students' learning experiences. Specifically, we ask the following research question:

> How do different avatar representations in AI chatbot tutors impact student learning experiences, and thereby influence AI chatbot interface design in a typical university course?

We developed an AI Tutor platform with three distinct *avatar representations* as shown in Fig.1. By studying how these different interfaces affect learning in a real-world educational environment, we aim to provide actionable insights into the design of more effective AI chatbot systems in education. The AI Tutor platform is an actual tool used by all students in a design thinking course, and serves as an instrument in our user study to explore the effects of different chatbot interfaces on learning. By *typical university course*, we refer to those that encompass both large lecture settings and smaller, tutorial-style sessions, akin to the design thinking course used in this study.

In summary, this paper contributes to (1) enhancing our understanding of the role of avatar representations in AI chatbots, (2) providing insights into how these representations affect student learning experiences, and (3) recommending opportunities for interface design of educational AI chatbot systems.

2 Related Work

Research on pedagogical agents (PAs) has long highlighted the benefits of virtual tutor agents in facilitating learning, though these benefits are contingent upon effective design [20]. Early intelligent tutoring systems (ITS) primarily focused on problem-solving support through computerized communication, while later systems increasingly incorporated more human-like interactions [25]. Meanwhile, increasingly advanced AI chatbots are now widely used in education to provide personalized learning and student support, underscoring their transformative potential [39].

Prior work has highlighted how various technical aspects of AI chatbots affect students' experiences; however, there remains a gap in understanding how interface design elements, particularly visual and auditory representations, influence learning experiences. For

instance, several systematic reviews [15, 24, 49] evaluated the role of AI chatbots in education, identifying their effectiveness in time-saving and improving pedagogy through a range of approaches such as adaptive personalized learning and the use of Retrieval-Augmented Generation (RAG) methods to alleviate hallucinations. Although these studies provide valuable insights into the benefits and technical needs of AI chatbots in various educational settings, they do not unpack the specific interface design elements, such as visual and auditory representations, that could enhance learning experiences in general.

Recent studies have stressed the importance of user interaction and interface design in maximizing the potential of AI chatbots. Clark et al. [13] provided insights on how principles from human-human conversations can inform human-agent interaction design, but also emphasized the need to view chatbot interfaces as a distinct interaction genre requiring novel design considerations. Lo and Hew [29] reviewed the integration of AI-based chatbots in flipped learning, highlighting enhanced engagement and motivation but also noting significant shortcomings in current designs to meet the specific needs of learning environments. El Azhari et al.'s systematic review [16] highlighted the lack of breadth in educational chatbots' knowledge bases and lack of smart interfaces to understand users, such as speech understanding. Rapp et al. [36] discussed how the concept of human-chatbot collaboration concepts can guide strategies for effective chatbot interfaces via a study conducted in a non-educational setting. While these studies consolidate key motivations for interface design in AI chatbots, they do not thoroughly explore how different interface elements, such as avatar representations, specifically influence learning experiences in educational settings.

Visual representations, such as avatars, play a critical role in creating a sense of presence and personal connection in conversational systems. A systematic review on PA personas highlighted the importance of facial expressions and gestures [38] and it is understood that these elements can convey complex meanings and emotions [41]. An embodied agent provides visual affordances to focus user attention and receive cues for interaction [10]. Properly designing avatars is a nuanced process that varies by context, which is crucial for improving user experiences and mitigating negative effects such as the uncanny valley effect [43] and cultural biases [28, 51].

While extensive research has been conducted on PA design, there remains an opportunity to further understand the impact of visual and auditory elements in modern AI chatbots, particularly those driven by LLMs, on learning experiences. Previous studies on general PA design have explored various factors, including the agent's role [22, 27, 30], emotions [4, 38], and instructional strategies [23, 52]. The importance of visual and auditory elements has been established in PA design [19, 46], with early work examining demographic attributes [3, 53] and anthropomorphism [30, 42]. Although these studies provide valuable insights into the general design of PA systems, LLM-driven AI chatbots possess distinct conversational capabilities and affordances beyond prior conversational agent technologies. For example, LLM-driven AI chatbots can generate more contextually relevant responses and provide advanced adaptive personalization capabilities, potentially influencing learner expectations of the agent's visual and auditory representations in nuanced ways.

Existing research on modern AI chatbot avatars has primarily focused on technical innovations and their specific effects on users, rather than exploring the nuanced effects of different avatar representations [1, 35, 48]. For example, Aneja et al. [1] developed a synthetic human-like avatar with user-aware facial expressions and lip-syncing, enhancing perceptions of empathy and believability. Similarly, Qin et al. [35] built CharacterMeet, an LLM-based chatbot with a customizable 3D avatar system to aid writers in character creation, benefiting their creative process significantly. While these studies inform the use of various chatbot technologies and their effects on user experiences, they do not provide detailed knowledge on the impact of different avatar representations on experiences, especially for learning.

In conclusion, while modern AI chatbots hold great promise for transforming educational practices, further research is essential to understand the nuanced effects of avatar interface design on learning experiences.

3 Method

We employed a mixed-methods within-subjects study design on an AI Tutor platform with participants from a university course. We anchored on reflexive thematic analysis (RTA) [6, 7], examining individual interviews and open-ended post-interaction responses. To address our exploratory research question (Sec. 1), RTA allows us to identify experiential patterns through a nuanced, inductively-oriented analysis of the data, guided by the first author's positionality as the course coordinator. RTA allows us to emphasize qualitative depth and contextual understanding, ensuring that our findings were grounded in the participants' rich experiences.

To investigate established experiential constructs, our approach is supplemented by a quantitative analysis on self-reported questionnaires based on the Technology Acceptance Model (TAM) [14], Situational Motivation Scale (SIMS) [18], and Basic Psychological Needs Scale (BPNS) [40]. TAM assesses the perceived usefulness and ease of use afforded by different avatar representations. SIMS and BPNS illuminate the motivational factors, as prior work has shown motivational effects linked to avatars in early PAs [2, 30]. SIMS explores the situational motivational dynamics across different avatar representations, while BPNS evaluates how these avatars satisfy core psychological needs.

3.1 The AI Tutor Platform

The AI Tutor was implemented as a web-based platform accessible through major web browsers. Upon user login, there are three distinct tabs representing the three interface modes (Fig. 1) studied: (1) a pure text-based interface named Mr Empathy (NOAVATAR); (2) a chat interface driven by a video-based real-time human replica of the course lecturer named Larry.exe (DEEPFAKE), and (3) a chat interface driven by a non-human 3D character named CoDEy (MASCOT). We studied these three modes as it enables us to begin our exploration of LLM-driven chatbots at the "global design level" as outlined in the Pedagogical Agents - Levels of Design (PALD) framework [20, p.47]. We chose DEEPFAKE to replicate the course

lecturer rather than a generic virtual human, as it provides opportunities to elicit novel responses in two areas: (1) an agent resembling a known human, generated by modern deepfake technology, and (2) pre-existing relationships with the actual lecturer. This choice also addresses a gap identified in prior research, which highlighted human-like PA designs as the least evaluated [30].

NOAVATAR mimics modern LLM-driven chatbots like OpenAI's ChatGPT [32] or Google's Gemini [17]. The interface is clean and minimalistic, focusing on text-based interactions without additional visual or auditory cues.

DEEPFAKE provides an audio-visual, human-like interaction using an avatar voice and video replica of the human course lecturer (who is not in the authorship team). The dynamic replica was generated using Heygen's API [21], based on the lecturer's recorded video footage. The mouth, facial and gesture animation were driven by the LLM-generated textual responses in real-time. Additionally, the background image depicts a familiar lecture hall at the actual university.

MASCOT provides an audio-visual interaction via a neutral, custom-made non-human 3D character with a cute-sounding voice. Its animations are similarly driven by the real-time LLM-generated responses. The MASCOT mode features a 360 photograph of the lecture hall as its background, enabling users to pan around the synthetic 3D character, highlighting its artificial nature to differentiate users' experience to DEEPFAKE's mimicry of a real-life conversation. Both DEEPFAKE and MASCOT provides text captions alongside their audio responses to ensure clarity in communication.

Across all three modes, a chat history is maintained that includes direct links to specific in-document locations of sourced course materials, allowing students to review and explore the content further. All modes utilized the same state-of-the-art LLM API, OpenAI's developer API [34] with GPT-4o model [33], and a custom implementation of Retrieval-Augmented Generation (RAG) [26] to enable the integration of course materials as contextual vector data.

3.2 Participants

Participants were students enrolled in a design thinking course as part of their computer science-related degree at the Singapore Institute of Technology (the host institution). The AI Tutor platform was incorporated as part of the learning activities for all students in the course. However, only students who provided informed consent and completed all components of the study were included in the data analysis for this paper. Initially, 45 participants volunteered at the start of the study, but only 23 completed all research procedures by the end of the study.

Of the 23 participants (17 males, 3 females, 3 preferred not to say; age range 19 to 29, M = 25.29, SD = 2.14), 7 considered themselves expert users, 13 were comfortable using AI chatbots for specific tasks, and 3 mentioned they were still experimenting. Regarding their primary use of AI chatbots, 10 participants used them mostly for general information, 4 for brainstorming in school projects, 3 for generating solutions for assignments, and the rest for other tasks. Table 1 provides a breakdown of each participant's demographics.

3.3 Procedure

As part of the course, students were allocated 15-30 minutes each week after lectures to individually reflect on the week's content using the AI Tutor platform. In the first three weeks, each student could only access the mode they were assigned to (NOAVATAR, DEEPFAKE, or MASCOT). For the remainder of the course, students had the freedom to use any of the three modes at any time. In the last two weeks, they were also allocated time in tutorial classes to consult the AI Tutor to critic their project deliverables.

For those who provided informed consent, a pre-survey was administered, which included demographic questions and inquiries about their prior experiences with chatbots and AI-driven tutoring applications, as summarized in Sec. 3.2.

To avoid order effects, the order of the three modes was counterbalanced for the first three weeks. After each session with an assigned mode, participants completed a post-session survey that included items from TAM, SIMS, and BPNS, with the item phrasing contextualized for our study.

During the remaining weeks where participants could interact with any mode, they were encouraged to provide experiential open-ended feedback through a form link in the AI Tutor platform. After the course ended, individual interviews (45-60 minutes) were conducted to gather qualitative data on participants' overall experiences with the different modes of the AI Tutor platform.

To address adherence issues caused by the length of the study, we contacted participants who had not completed any post-session surveys to encourage them to do so. However, a significant number did not respond to these follow-up requests, resulting in a final sample size of 23 participants. These participants were compensated with a gift card.

Institutional Review Board approval was obtained for this study at the host institution.

3.4 Qualitative Data Analysis

In alignment with RTA, the first author conducted the primary analysis, allowing for a nuanced and in-depth interpretation of the data based on the author's intimate involvement in the course as the coordinator and primary interviewer. This is consistent with the notion that RTA "works especially well with a single researcher" [7]. RTA was performed on a combined dataset of interview transcriptions and open-ended post-interaction responses for each participant.

The first author, a computer science faculty at the host institution who has coordinated this course for the past four years, conducted the majority of the interviews (20 out of 23). The remaining interviews were conducted by the research team, with the first author subsequently reviewing the recordings and transcripts. This process greatly facilitated the first author's initial immersion in the data. All interviews were conducted after the completion of the course, ensuring that information regarding participation during the course was not disclosed to the first author. As the course coordinator, the first author observed all learning activities but did not participate in any assessment activities. Note that the first author was not the lecturer behind DEEPFAKE, and the lecturer was not involved in the interviews.

An interview guide was employed to ensure consistency across interviews, focusing on participants' overall experiences with the

ID	Gender	Age	Familiarity with LLM chatbots	Most used task with LLM chatbots	Ranking of AI Tutor modes after course ended (rank 1 on left)
P1	Male	25	Comfortable	General info	MASCOT, DEEPFAKE, NOAVATAR
P2	Male	25	Experimenting	General info	NOAVATAR, DEEPFAKE, MASCOT
P3	Male	24	Expert	Assignment solutions	NOAVATAR, MASCOT, DEEPFAKE
P4	Male	26	Comfortable	Proofreading	DEEPFAKE, MASCOT, NOAVATAR
P5	Prefer not to say	24	Experimenting	Assignment solutions	DEEPFAKE, NOAVATAR, MASCOT
P10	Male	28	Comfortable	General info	NOAVATAR, DEEPFAKE, MASCOT
P15	Male	26	Comfortable	Others	MASCOT, NOAVATAR, DEEPFAKE
P21	Female	24	Comfortable	Entertainment	MASCOT, NOAVATAR, DEEPFAKE
P22	Male	24	Expert	Brainstorming	NOAVATAR, DEEPFAKE, MASCOT
P24	Male	25	Comfortable	General info	NOAVATAR, MASCOT, DEEPFAKE
P28	Male	29	Comfortable	Assignment solutions	DEEPFAKE, NOAVATAR, MASCOT
P29	Male	24	Comfortable	General info	DEEPFAKE, MASCOT, NOAVATAR
P30	Male	27	Comfortable	General info	NOAVATAR, MASCOT, DEEPFAKE
P31	Female	23	Comfortable	Brainstorming	MASCOT, DEEPFAKE, NOAVATAR
P33	Prefer not to say	19	Expert	General info	NOAVATAR, DEEPFAKE, MASCOT
P35	Male	25	Comfortable	Others	NOAVATAR, MASCOT, DEEPFAKE
P37	Male	24	Experimenting	Proofreading	NOAVATAR, MASCOT, DEEPFAKE
P38	Male	26	Expert	General info	DEEPFAKE, MASCOT, NOAVATAR
P40	Prefer not to say	25	Expert	Brainstorming	DEEPFAKE, MASCOT, NOAVATAR
P41	Male	27	Comfortable	General info	NOAVATAR, DEEPFAKE, MASCOT
P42	Male	27	Comfortable	General info	NOAVATAR, DEEPFAKE, MASCOT
P44	Female	24	Expert	Brainstorming	DEEPFAKE, NOAVATAR, MASCOT
P45	Male	29	Expert	General info	DEEPFAKE, NOAVATAR, MASCOT

Table 1: Summary of participant demographics and ranking of AI Tutor mode preferences after the 14-week course.

AI Tutor, their comparative perceptions of the avatar representations, and suggestions for improvement. This guide facilitated systematic data collection while allowing for the development of rich, contextual insights.

To enhance transparency and rigor, the first author maintained a codebook to document the codes and themes. While the analysis was primarily reflexive, having a codebook served as a structured reference for ongoing theme development, and was used for review by the second author, also a computer science faculty member at the host institution, to ensure consistency and reliability. A reflexive journal was also maintained to document the first author's reflections and interpretations of the data, grounding the codes in the data and the researcher's reflexivity.

The codebook was updated and revised as additional interview transcript became available, incorporating both the participants' narratives and the researcher's perspectives and experiences from the reflexive journal. Individual codes were iteratively organized into themes, and the names and descriptions of these themes were continuously refined to ensure they accurately represented the data. This approach allowed for a comprehensive and nuanced understanding of the participants' interactions with the AI Tutor, while also accounting for the researcher's positionality.

After the first author coded the complete data corpus guided by the reflexive journal, the second author reviewed the initial codes and themes. Subsequently, the research team engaged in discussions to refine and validate the codes and themes in the codebook, thereby enhancing the reliability of the analysis.

3.5 Quantitative Data Analysis

Despite the eventual sample size of 23 participants being smaller than anticipated, we proceeded with the quantitative analysis to provide a comprehensive picture of the study's findings and to offer preliminary quantitative insights.

The constructs from TAM, SIMS, and BPNS were aggregated for analysis (Fig. 3). The constructs exhibited mixed normality as assessed by the Shapiro-Wilk test but demonstrated homogeneity of variance according to Levene's test. Given the small sample size and the mixed normality of the data, we utilized the non-parametric Friedman Test to determine statistical significance in comparing the construct scores across the three avatar representations.

In the interviews, participants were also asked to rank the three avatar representations based on their preferences (Table 1). Given that ranking data is ordinal, we utilized the Friedman Test to determine statistical significance.

4 Results

The results of our study are primarily anchored in rich qualitative data, augmented by quantitative measures. Through the eventual themes, we gain detailed insights into each participant's experiences and preferences, providing a comprehensive understanding of how different avatar representations influence student experiences.

4.1 Avatar preferences vary among students

When asked to rank the three avatar representations based on their preferences, participants' responses varied widely (summarized in Fig. 2 with detailed rankings in Table 1). There were no statistically significant differences in rankings among the three AI Tutor modes

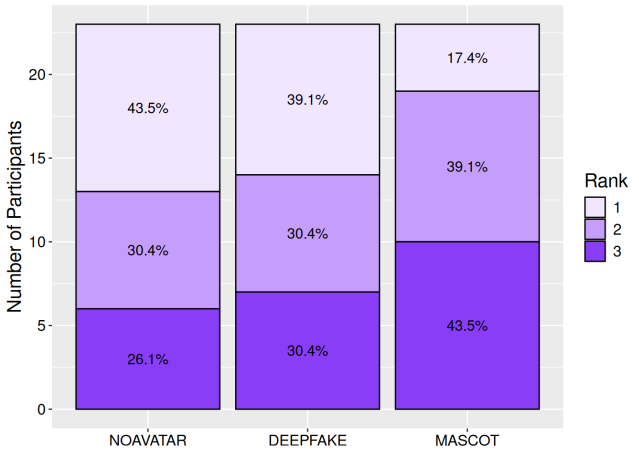


Figure 2: Rankings of participant preferences for the three AI Tutor modes.

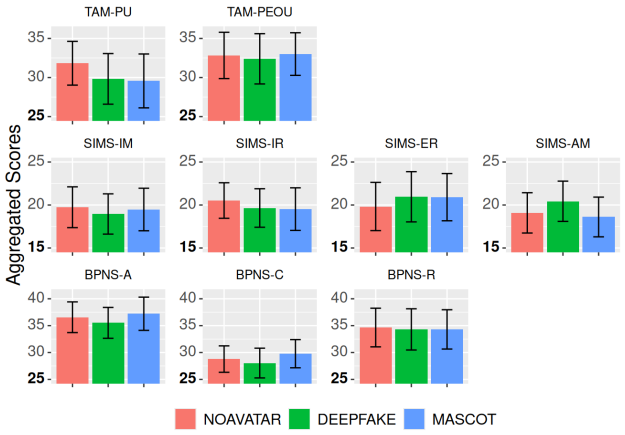


Figure 3: Mean aggregated construct scores from TAM (Perceived Usefulness (PU), Perceived Ease of Use (PEOU)), SIMS (Intrinsic Motivation (IM), Identified Regulation (IR), External Regulation (ER), Amotivation (AM)), and BPNS (Autonomy (A), Competence (C), Relatedness (R)) questionnaires. The y-axis in each plot is truncated to begin at higher values to enhance the visibility of comparisons.

($\chi^2(2) = 2.43, p > 0.05$). This suggests that there is no clear single preference for any of the three avatar representations, and further studies may be needed to confirm this.

Usability and motivational scores, as measured by the TAM, BPNS, and SIMS questionnaires, were generally high across all three avatar representations (Fig. 3). The results indicated no statistically significant differences across the three avatar representations for any of the evaluated constructs: TAM-PU ($\chi^2(2) = 1.58, p > 0.05$), TAM-PEOU ($\chi^2(2) = 0.34, p > 0.05$), SIMS-IM ($\chi^2(2) = 0.11, p > 0.05$), SIMS-IR ($\chi^2(2) = 2.03, p > 0.05$), SIMS-ER ($\chi^2(2) = 0.74, p > 0.05$), SIMS-AM ($\chi^2(2) = 4.08, p > 0.05$), BPNS-A ($\chi^2(2) = 2.84, p > 0.05$), BPNS-C ($\chi^2(2) = 2.16, p > 0.05$), and BPNS-R ($\chi^2(2) = 0.24, p > 0.05$). This suggests that the different representations did not significantly influence perceived usefulness, ease of use, or the various motivational constructs, though further investigation is warranted. Consequently, individual participant preferences for the avatars may not have been directly related to these specific usability and motivational constructs.

However, thematic analysis of the qualitative data provided rich insights into user preferences related to learning habits and activities.

When considering different learning activities, participants felt that the activities influenced their preferences: "Mr Empathy [NOAVATAR], if I want to ask a quick question ... like in lectures or in tutorial. I just like need an immediate answer to double check something ...But I feel like for CoDEy [MASCOT] and Larry [DEEPFAKE] it's more like when I'm self studying like, I would prefer that UI when I'm self studying because it feels more inviting." (P15)

Conversely, there were others who did not find that different activities mattered: "Different learning activities... really don't see any strategies for choosing one (mode) over the other." (P41)

When considering learning habits however, there was common consensus that they influenced avatar preferences: "Yes, so actually, Mr. Empathy [NOAVATAR] suits more to my learning style, because I want to see all the text." (P45)

Generally, participants who preferred NOAVATAR were more task-oriented (Sec. 4.1.1), while those who preferred DEEPFAKE and MASCOT were more engagement-oriented (Sec. 4.1.2).

4.1.1 Task-oriented learners prefer text-based interfaces. Participants who preferred NOAVATAR valued the simplicity and directness over the other modes: "Like you're multitasking ... I think the 1st one [NOAVATAR] is way easier, like you get an answer straight, you find keywords." (P3)

Compared to the other modes, participants also appreciated NOAVATAR's faster, distraction-free responses: "Mr Empathy is more straight-forward and less distracting to user and the latency feels lower" (P35)

There were also specific efficiency concerns related to computer resource usage: "They provide similar results but the other two require more cpu/mem usage." (P3)

These participants were also more inclined to provide feedback on core features such as the text box sizes and keyboard shortcuts, even when the conversation primarily focused on avatars: "Definitely one is the large text box for the input prompts ... then also large response ... large box for text output." (P41)

It also appears that participants whose learning habits heavily relied on popular text-based chatbots tended to prefer NOAVATAR: "Maybe I got used to ChatGPT already, so the text-based one is more comfortable to me also." (P33)

4.1.2 Engagement-oriented learners prefer avatar-based interfaces. Participants who favored DEEPFAKE and MASCOT generally sought features that provided them with a sense of engagement. Those who preferred DEEPFAKE highlighted its realistic and relatable human-like interactions: "It feels more personable because I am talking to the AI version of the professor that is teaching me." (P29)

Participants who favored MASCOT were enticed by its high interactivity: "If we see something cute, we are actually more encouraged to use it." (P1)

Those who preferred having avatars also frequently highlighted the lack of attraction to the NOAVATAR mode: "Like I would feel less inclined to use it since there's nothing to look at while interacting with it." (P15)

The avatars (DEEPFAKE and MASCOT) were often seen as more engaging, including participants who did not rank it as their top preference: "I think engagement-wise. I would say the text-to-speech [DEEPFAKE/MASCOT] is better. Yeah, like text one [NOAVATAR] is quite dry, like you have to read yourself." (P24)

In addition, some participants who preferred NOAVATAR did not totally object to the presence of avatars but thought they could take a more secondary role: "It would be good if the chat logs and history were in the center instead of the corner... As users will be mainly interacting with and looking at it often instead of CoDEy [MASCOT]." (P30)

4.2 Avatar design affords prompting behavior and perceived response quality

Participants noted that the look-and-feel of the AI Tutor shaped the questions they asked: "If I talk to the Larry.exe [DEEPFAKE], maybe I won't really go out of scope that much, because it feels like they are talking to the Prof. So you might ask question more related to the course itself. But for CoDEy [MASCOT], because of the avatar, you feel like you can go more outside of the scope." (P40)

The first author witnessed firsthand instances where participants asked MASCOT to generate code snippets, likely influenced by its name ("CoDEy") and appearance. Interviews supported this, as participants expressed how avatar attributes influenced their prompts: "Mr. Empathy [NOAVATAR] is more for the social things. Then Larry [DEEPFAKE] is more just for the module. And CoDEy [MASCOT], I thought, he's gonna write me some some computer code." (P45)

Interestingly, the avatar design also influenced the perceived quality of the responses, even though the underlying LLM-based system was the same. For instance, some felt that DEEPFAKE provided more accurate and reliable information for questions related to course content: "Larry [DEEPFAKE] is able to understand questions better ... Only responses matter to me. Larry [DEEPFAKE] gave the best response compared to the other two." (P4)

While others instead thought higher of the NOAVATAR: "From what I see, the Empathy [NOAVATAR] AI is able to provide more structured answers compared to Larry.exe [DEEPFAKE]." (P2)

Comparing the avatar-based interfaces, DEEPFAKE was seen as more capable of understanding complex questions, which led to more in-depth discussions: "For this course, It's Larry.exe [DEEPFAKE]. Maybe some other course. These two [NOAVATAR/MASCOT] will be better, because I don't know how it's trained more for each model. I assume they are trained differently. (Participant enquires about implementation details like model version, temperature, etc.)" (P4)

On the other hand, the MASCOT was seen as fun and engaging, but less reliable in providing accurate information: "When interacting with Mr Empathy [NOAVATAR], I thought of asking it to help me do trivial and more manual tasks such as generating flash cards and asking it to come up with possible quiz questions. I think with Larry.exe [DEEPFAKE], I didn't think of it (those things)." (P42)

4.3 Influence of existing real-world relationships on trust and engagement with deepfakes

Unpacking the preferences for DEEPFAKE revealed a rich interplay of factors that relate to the participants' real-world relationships with the course lecturer. An important observation was that familiarity with the person behind the deepfake could influence its perceived trustworthiness: "Larry.exe [DEEPFAKE] feels more reliable because of the voice sounding like Prof. Larry ... but the other one is like it doesn't feel as trustworthy to me." (P5)

The knowledge of the lecturer behind the DEEPFAKE also seemed to affect the perceived capability of the AI Tutor: "Larry [DEEPFAKE] gave me the subconscious idea that I am talking to someone capable." (P44)

Participants mentioned they only chose the DEEPFAKE because they knew the professor personally and liked his/her teaching style. They indicated that if it was a professor they disliked, they would not have chosen the DEEPFAKE: "If I would not like the Prof., I don't think I would use it to be honest ... Yeah, it's quite, very personal." (P38)

Participants also noted the familiar teaching style provided by the DEEPFAKE, which they found beneficial due to their relationship with the course lecturer: "The way he [DEEPFAKE] voiced out the answer also seemed very easy to digest because, when I listen to Larry's lecture, I have the idea that it's easy to digest, easy to understand. Then, when the bot delivers it the same way, it gives me the same feeling." (P44)

The consistency of the familiar teaching style was also highlighted as desirable, with participants finding it more effective when consuming newly generated information: "We went on YouTube to look for tutorial videos (that offer alternative explanations of lecture content) and it turned out to be a bit hard to understand... that's where I feel that if it's actually a Prof., our own [DEEPFAKE] Prof. talking to us, it'll be much better." (P1)

At the same time the entertaining aspect to the DEEPFAKE also seemed to draw users, i.e., the humor of having the lecturer made into a chatbot. "They might find it interesting. Oh, I have my professor on my tablet." (P10)

4.4 Impact of human touch on interaction expectations and learning experiences

Participants generally felt that interactions with human-like avatars provided a better learning experience. Some mentioned that NOAVATAR lacked this human touch, describing interactions as feeling like "talking to a wall" (P1).

Even those who preferred NOAVATAR (Sec. 4.1.1) found that having some form of typing text animation improved the perceived human touch compared to the immediate display of responses: "It (the text response) just pops up, there isn't like a visual effect ... it just feels like a little bit out of nowhere." (P37)

Others who mentioned similar sentiments about the text animation suggested that the name "Mr Empathy" affords certain

expectations that were not met by NOAVATAR: "It feels very sudden (the text appearing) like one shot everything. The name (Mr Empathy [NOAVATAR]) doesn't fit what it's doing." (P40)

Generally, when asked why human qualities like emotions are important to learning, participants were unable to provide a clear answer: "I'm sorry I can't really explain this that well." (P4)

However, they consistently mentioned that the human touch made the learning experience more engaging and enjoyable: "Mr. Empathy [NOAVATAR] definitely affected my engagement because I feel a bit bored when interacting with it. But everytime with Larry [DEEPFAKE] and CoDEy [MASCOT] it's like a different kind of engagement. Cause it's interactive, they actually responds to me. And I feel like I'm talking to someone instead of I'm talking to (a bot)." (P1)

An interesting perception is how the NOAVATAR lacked human touch, but it was actually what was preferred: "I feel like the emotional aspect of acknowledging that this is a robot. It's not a human feels more natural to me." (P41)

Related to affordance (Sec. 4.2), the recognition of human touch also influenced the prompting behaviors. "Maybe it is because of the human element, like you wouldn't really ask your professor to make flash cards for you or give you quiz questions. Maybe the human element can prevent some types of questions being asked from the user." (P42)

4.5 Perceived benefits of tighter integration with learning content and activities

Participants highly valued features allowing the AI Tutor to integrate seamlessly with course materials, particularly through the RAG system for accessing specific pages in module content (Fig. 1, A2). This integration provided a means to verify generative responses, addressing issues such as chatbot hallucinations: "And it also provided the link to ensure that at least I can fact check it myself." (P30)

Participants appreciated the ability to search through course materials efficiently, which facilitated faster knowledge retrieval: "Ten plus PDFs on [LMS], then might not be easy to find what you want. So in these situations, the LLM really helps a lot. You can get the answer faster." (P10)

There was also a strong desire to enhance the AI Tutor's integration with the broader learning environment, including interactions with related courses within the degree program and connections to additional learning tools such as note-taking platforms: "you know, if this is connected to the [university LMS] and the modules right ... yeah, it'll be like immediate feedback." (P15)

Further suggestions included involving course instructors in the AI Tutor platform, allowing the AI to mediate student-instructor interactions and thereby enhance contextual information to generate more relevant responses: "It would help for AI Tutors to provide constant feedback to the professors over the course of the modules, as sometimes students may not feel comfortable giving feedback, especially negative, to the Prof.s (directly)." (P28)

Participants also proposed adding "sensing" technology for real-time feedback during learning activities. "So as I'm writing the assignment, you know using C code, ..., then on the side or in the comments of the C code ... the chatbot or text saying that, hey, actually I noticed that you wrote this section of code, there might be a possible bug." (P42)

Avatar interactions appeared to prompt participants to even suggest gamification as a means to improve engagement: "Maybe to have more connection to the bot (AI Tutor). We can have a leveling system like, grow your bot, grow your character kind of thing." (P44)

While these findings may initially appear unrelated to avatar representations, they surfaced from the data—even when participants were not specifically prompted—and significantly enhance our understanding of the interplay between avatar design and the broader learning environment (Sec. 5.3).

5 Discussion and Recommendations

Based the results, we discuss the key insights and position it in the context of existing literature in each subsection below. We also provide actionable recommendations for AI chatbot design in educational settings.

5.1 Diverse Preferences and Their Implications

The wide variation in participants' rankings of the three avatar representations indicates diverse preferences, with nuances related to learning habits and activities revealed through the rich qualitative data (Sec. 4.1). Participants who preferred NOAVATAR were primarily task-oriented, appreciating the distraction-free and efficient nature of the text-based interface (Sec. 4.1.1). In contrast, those who favored DEEPFAKE and MASCOT were more engagement-oriented, valuing the interactive and human-like features of these avatars (Sec. 4.1.2). These differing preferences may underline the importance of offering both minimalist and interactive options to support various learning styles effectively, aligning with work on general PAs that highlights the benefits of providing different avatar choices [2] and customization [52].

When considering different learning activities, some participants felt that the choice of avatar representation mattered, whereas others did not (Sec. 4.1). This further highlights the complexity of designing AI Tutor interfaces that can cater to a wide range of user needs and preferences.

Previous studies have demonstrated value in both text-based [13, 42] and avatar-based interfaces [30] for educational chatbots in different contexts. The differently perceived values of the three avatar representations in our study align with these findings, with new insights in the context of LLM-based AI Tutors. Additionally, we enhance this understanding by connecting these varied preferences to specific user types—task-based versus engagement-based learners.

Recommendations: Provide options for varied avatar representations, such as text-based, human-like, and non-human animated avatars, and allow students to personalize their learning environment. An advanced approach would be to develop customizable avatars that can adapt to individual learning habits and preferences.

5.2 The Role of Human-like Interaction in Learning

The perceived human touch in AI Tutor interactions was seen to impact learning experiences. For participants who value this,

these avatars' characteristics (realistic, cute, humorous, etc.) helped provide participants a sense of relatability and connection, which they found beneficial for their learning (Sec. 4.4).

A key finding in our study is the significant impact that real-world relationships have on the perceived credibility of deepfake AI Tutors, highlighting the crucial role of personal connections in educational interactions. While prior research in pedagogical agents (PAs) indicates that human-like qualities can enhance trust [37], it remains unclear if this applies to deepfakes with whom users have existing relationships. Our findings suggest that familiarity with the lecturer behind the deepfake increases the perceived trustworthiness and capability of the AI Tutor, suggesting valuable opportunities for leveraging deepfake technology in educational settings where students know the real-world instructors. Some participants even desired having other tutors represented by deepfake avatars.

Further supporting this perspective, recent research emphasizes the importance of personalization and communication approaches in enhancing user engagement and trust. Sun et al.'s [44] investigation into personalizing LLMs for more engaging experiences highlights the positive impact of deepfake avatars, aligning with our findings about enhanced learner experiences with realistic human replicas. Additionally, Metzger et al. [31] found that authoritative communicative styles in conversational agents can elicit greater trust, which complements our observation that deepfake avatars can produce similar effects but through the perception of authority rather than actual communication style.

Interestingly, some participants preferred NOAVATAR's non-human nature, finding it more suitable for straightforward, task-oriented interactions (Sec. 4.4). However, even these participants appreciated typing text animations to enhance the interface's human touch. This diversity in preferences suggests that while human-like interactions can enhance engagement, there is also a place for more utilitarian, text-based interfaces for those who prioritize efficiency.

The avatar design also influenced prompting behavior, with participants adjusting their questions based on the perceived capabilities and characteristics of the avatars (Sec. 4.2). For instance, DEEPFAKE led to more course-related questions, while MASCOT encouraged creative inquiries. An aspect that surprised us, the designers of the current AI Tutor, was how even the name (e.g., "Mr Empathy") —something we did not initially consider to have an effect—influenced prompting behavior. This suggests that the design and appearance of avatars significantly shape the scope of interactions, even in the absence of more advanced personalization features in the LLM, such as providing stylized generated responses for each mode. It extends prior research on personality in text-based chatbots [47] and suggests avatar design could interact with personality-driven systems. Beyond confirming previous research on stereotypical expectations arising from static PA avatars [46], our findings extend these insights to modern LLM-based chatbots with dynamic avatars, highlighting their further influence on prompting behaviors.

Relationships with the real-world lecturer behind DEEPFAKE also appeared to mediate prompting behavior, influencing the perceived credibility and trustworthiness of the AI Tutor (Sec. 4.3), and has the potential to enhance learning outcomes. Prior research on traditionally animated PAs has shown enhanced learning outcomes when avatars feature realistic and non-traditional characteristics in expert roles [3, 22]. Our results may present a contemporary manifestation of a "realistic non-traditional expert" achieved via an LLM-driven deepfake. The added positive associations with the real-world expert (lecturer) could further enhance learning through the prompting affordances. This relationship between the deepfake avatar and its real-world counterpart presents unique opportunities for future research into learning outcomes facilitated by modern AI Tutors.

It is interesting that DEEPFAKE was not generally perceived negatively despite the growing ethical and credibility concerns associated with deepfake technology [5, 9]. The rankings of DEEPFAKE were not significantly different from NOAVATAR and MASCOT, and there were no reports of negative feelings such as being deceived or misled by DEEPFAKE. This suggests that perceived concerns may not be as severe as anticipated, possibly due to the course context and familiarity with the lecturer. However, broader ethical implications warrant careful consideration, such as the risk of trusting misinformation from deepfake avatars associated with familiar individuals. Although incorporating RAG in our system alleviates misinformation to an extent, it is important to remain cautious in educational settings. Furthermore, deepfakes can create unrealistic expectations of instructors. These nuances highlight the need for responsible use and governance of such technologies to maintain trust and integrity in educational environments.

Our findings also suggest that deepfake avatars representing familiar individuals can evoke humorous thoughts, which contribute positively to student experiences. This loosely associates with Ceha et al.'s [11] research, which highlights that affiliative humor in conversational agents can significantly enhance motivation and effort. However, our findings primarily involve perceptive humor relating to familiar associations rather than direct conversational humor. Humor, whether conversational or perceptive, can act as a catalyst for positive attitudes, fostering a more welcoming and less intimidating learning environment. Moreover, in our study, while the humorous feelings towards deepfake avatars did not negatively affect the learning experience, it underscores the importance of understanding the appropriate use of humor in educational tools to maximize its benefits.

Recommendations: Include the design of avatars that provide relatable human-like visual and auditory cues that enhance the interaction without overwhelming the user. Avatars should complement textual responses with appropriate gestures, expressions, and contextual information to facilitate better comprehension. Additionally, explore integrating humor in avatar interactions to enhance engagement and motivation.

5.3 Enhancing Engagement through Integration and Interaction

Participants consistently highlighted the value of integrating the AI Tutor with the learning environment, even when not explicitly prompted (Sec. 4.5). The Pedagogical Agents-Conditions of Use Model (PACU) illustrates how a PA's functions and its learning environment are closely linked with avatar design [20]. This framework allows us to contextualize our findings in relation to the research question (Sec. 1).

Features such as direct links to course documents were deemed crucial for knowledge retrieval and information verification, particularly in addressing concerns about chatbot hallucinations (Sec. 4.5). This supports the "information processing" function within PACU [20, p.46], which can moderate the higher trust placed in deepfake avatars, thereby potentially alleviating ethical concerns regarding misinformation and its impact on learning outcomes.

There was a strong desire for enhanced integration of the AI Tutor into the broader educational ecosystem, including real-time feedback on assignments and quizzes, connections to related courses, and compatibility with additional learning tools such as note-taking platforms (Sec. 4.5).

This integration supports the "monitoring and directing" and "transfer of information" functions of PACU [20, p.46]. This offers PA designers insights into tailoring different avatar representations, such as using an authoritative-looking DEEPFAKE for monitoring assessments and a neutral-looking MASCOT for guiding students to related courses.

Prominent avatar interactions led many participants to suggest adding "sensing" technology and gamification, highlighting the perceived value for interactive elements to increase motivation and engagement in learning (Sec. 4.5). This aligns with the "motivation" function of PACU [20, p.46], indicating how designers might use fun-looking MASCOT avatars to create a gameful experience through visual cues and fun interactions.

Overall, these findings emphasize the importance of designing AI Tutor avatar representations with consideration for their integration and interaction within the broader learning environment. This underscores the pressing need for AI chatbots in modern education to incorporate robust interfaces and comprehensive knowledge bases [16].

Recommendations: Enhance relevant avatar representations with abilities to provide real-time assessment feedback, integrate with course content, and connect to broader learning management systems. Additionally explore incorporating gamified elements to enhance engagement in sustained chatbot interactions for learning.

6 Conclusion

This study explored the impact of different avatar representations in AI chatbot tutors on students' learning experiences within a university design thinking course. Our findings reveal the nuances of how diverse avatar representations influence student engagement and learning experiences. While some participants favored a distraction-free text-based interface, others favored more engaging interactions facilitated by human-like or animated avatars. The affordances of different avatar designs also influenced prompting behavior and the perceived quality of responses, with existing real-world relationships with the deepfake tutor notably influencing perceptions. As a learning tool, it was also found that students valued the AI Tutor's tight integration with the learning environment, suggesting that interface features such as direct access to course materials and assignment integration are essential for enhancing the learning experience.

6.1 Limitations

The study was limited to students from a single course, which may affect the generalizability of the findings. However, this may enhance the internal validity of the study for similar contexts. Additionally, the smaller-than-anticipated sample size may have impacted the statistical power of the quantitative analysis.

Via the open-ended responses, participants also indicated that they experienced fatigue due to the lengthy surveys that had to be completed three times, which could have contributed to the high attrition rate.

Although this study was eventually able to provide rich qualitative insights into the impact of avatar representations on student learning experiences, the fatigue issues and small sample size nevertheless limit the ability of the quantitative analysis to substantially augment the findings. Our findings should hence be interpreted with the limitations in mind.

6.2 Future Work

A straightforward extension of this study would be expand the sample size and scope of participants to include a more diverse range of students across different courses and institutions. This will enhance the quantitative analysis and provide a more comprehensive understanding of the impact of avatar representations on student learning experiences.

Longitudinal studies investigating longer-term engagement and learning outcomes with AI tutor avatars will also provide deeper insights into their effectiveness over time. Moreover, future studies should explore the cultural sensitivity and inclusivity of avatar designs to develop more universally effective educational tools. Investigating the integration of advanced features like real-time feedback, deeper learning environment integration, and gamified elements will also contribute to creating more engaging and interactive learning experiences.

Overall, this study contributes to the growing body of knowledge on AI-enabled conversational tools and highlights the importance of thoughtful interface design in enhancing learning outcomes with chatbots. By addressing the diverse needs of students through customizable avatars and integrated learning experiences, AI tutors have the potential to significantly improve educational practices and outcomes.

Acknowledgments

This work was supported by the Centre for Digital Enablement (CoDE) at the Singapore Institute of Technology (SIT). The study would not have been possible without the dedicated efforts of Ms Jacinta Ong and Mr Pek Isaac from CoDE, who provided expertise in LLM development and logistical support for the user study. We extend our heartfelt appreciation to Mr Larry Yeung, the course lecturer and design thinking practitioner, for his invaluable support in facilitating this study, despite the demands of his teaching commitments. Our gratitude also goes to the reviewers for their thoughtful and detailed feedback throughout the submission iterations, which greatly contributed to the significantly improved final version presented here.

References

[1] Deepali Aneja, Rens Hoegen, Daniel McDuff, and Mary Czerwinski. 2021. Understanding Conversational and Expressive Style in a Multimodal Embodied Conversational Agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 102, 10 pages. https://doi.org/10.1145/3411764.3445708

[2] Amy L. Baylor. 2005. The Impact of Pedagogical Agent Image on Affective Outcomes. In *Proceedings of the Workshop on Affective Interactions: Computers in the Affective Loop, International Conference on Intelligent User Interfaces*. ACM, San Diego, CA, USA.

[3] A. L. Baylor and Y. Kim. 2004. Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role. In *Intelligent Tutoring Systems. ITS 2004*, J. C. Lester, R. M. Vicari, and F. Paraguaçu (Eds.). Springer, Berlin, Heidelberg.

[4] M. Beege and S. Schneider. 2023. Emotional design of pedagogical agents: the influence of enthusiasm and model-observer similarity. *Educational Technology Research and Development* 71, 3 (2023), 859–880. https://doi.org/10.1007/s11423-023-10213-4

[5] BlackBerry. 2024. *Deepfakes Unmasked: The Ethics and Threats of Deepfake Technology*. Technical Report. BlackBerry. https://www.blackberry.com/us/en/solutions/threat-intelligence

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[7] Virginia Braun and Victoria Clarke. 2022. *Thematic Analysis: A Practical Guide*. SAGE Publications, London.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Red Hook, NY, USA, 1877–1901. Virtual Conference.

[9] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, H. Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna J. Bryson, Roman V. Yampolskiy, and Dario Amodei. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *ArXiv* abs/1802.07228 (2018). https://api.semanticscholar.org/CorpusID:3385567

[10] Justine Cassell. 2000. Embodied conversational interface agents. *Commun. ACM* 43, 4 (apr 2000), 70–78. https://doi.org/10.1145/332051.332075

[11] Jessy Ceha, Ken Jen Lee, Elizabeth Nilsen, Joslin Goh, and Edith Law. 2021. Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 685, 14 pages. https://doi.org/10.1145/3411764.3445068

[12] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758. https://doi.org/10.1080/10447318.2020.1841438

[13] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[14] Fred D. Davis. 1985. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results*. PhD Thesis. Massachusetts Institute of Technology, Cambridge, MA, USA.

[15] Jinming Du and Ben Kei Daniel. 2024. Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence* 6 (2024), 100230. https://doi.org/10.1016/j.caeai.2024.100230

[16] Khadija El Azhari, Imane Hilal, Najima Daoudi, and Rachida Ajhoun. 2023. SMART Chatbots in the E-learning Domain: A Systematic Literature Review. *International Journal of Interactive Mobile Technologies (iJIM)* 17, 15 (Aug. 2023), pp. 4–37. https://doi.org/10.3991/ijim.v17i15.40315

[17] Google. 2024. Google Gemini - Next Generation AI Model. https://www.google.com/. Accessed: 2024-09-09.

[18] Frédéric Guay, Robert J. Vallerand, and Céline Blanchard. 2000. On the Assessment of Situational Intrinsic and Extrinsic Motivation: The Situational Motivation Scale (SIMS). *Motivation and Emotion* 24, 3 (2000), 175–213. https://doi.org/10.1023/A:1005614228250

[19] Agneta Gulz and Magnus Haake. 2006. Design of animated pedagogical agents-A look at their look. *Int. J. Hum.-Comput. Stud.* 64, 4 (2006), 322–339. https://doi.org/10.1016/j.ijhcs.2005.08.006

[20] Steffi Heidig and Geraldine Clarebout. 2011. Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review* 6, 1 (2011), 27–54. https://doi.org/10.1016/j.edurev.2010.07.004

[21] HeyGen. 2024. HeyGen - AI Video Generator. https://www.heygen.com/. Accessed: 2024-09-09.

[22] Y. Kim and A. L. Baylor. 2016. Research-Based Design of Pedagogical Agent Roles: a Review, Progress, and Recommendations. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 160–169. https://doi.org/10.1007/s40593-015-0055

[23] Y. Kim, A. L. Baylor, and PALS Group. 2006. Pedagogical Agents as Learning Companions: The Role of Agent Competency and Type of Interaction. *Educational Technology Research and Development* 54, 3 (2006), 223–243.

[24] Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2023. Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education* 20 (10 2023). https://doi.org/10.1186/s41239-023-00426-1

[25] H. Chad Lane and Noah Schroeder. 2022. Pedagogical Agents. In *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application* (1 ed.), Birgit Lugrin, Catherine Pelachaud, and David Traum (Eds.). Vol. 2. Association for Computing Machinery, New York, NY, USA, Chapter 21, 307–329.

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[27] M. Liao, X. Luo, H. Yang, and K. Zhu. 2024. The interactive effects of pedagogical agent role and voice emotion design on children's learning. *Current Psychology* 43, 36 (2024), 29170–29188. https://doi.org/10.1007/s12144-024-06559-4

[28] Zihan Liu, Han Li, Anfan Chen, Renwen Zhang, and Yi-Chieh Lee. 2024. Understanding Public Perceptions of AI Conversational Agents: A Cross-Cultural Analysis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 155, 17 pages. https://doi.org/10.1145/3613904.3642840

[29] Chung Kwan Lo and Khe Foon Hew. 2023. A review of integrating AI-based chatbots into flipped learning: new possibilities and challenges. *Frontiers in Education* 8 (2023). https://doi.org/10.3389/feduc.2023.1175715

[30] A. S. D. Martha and H. Santoso. 2019. The Design and Impact of the Pedagogical Agent: A Systematic Literature Review. *The Journal of Educators Online* 16, 1 (2019). https://doi.org/10.9743/jeo.2019.16.1.8

[31] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 481, 19 pages. https://doi.org/10.1145/3613904.3642122

[32] OpenAI. 2023. ChatGPT: Optimizing Language Models for Dialogue. (2023). https://openai.com/chatgpt Accessed: 2024-09-09.

[33] OpenAI. 2024. GPT-4o: Omni-modal AI Model by OpenAI. https://www.openai.com/blog/gpt-4o Accessed: 2024-08-29.

[34] OpenAI. 2024. OpenAI Developer API. https://platform.openai.com/. Accessed: 2024-09-09.

[35] Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1051, 19 pages. https://doi.org/10.1145/3613904.3642105

[36] Amon Rapp, Arianna Boldi, Lorenzo Curti, Alessandro Perrucci, and Rossana Simeoni. 2023. Collaborating with a Text-Based Chatbot: An Exploration of Real-World Collaboration Strategies Enacted during Human-Chatbot Interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 115, 17 pages. https://doi.org/10.1145/3544548.3580995

[37] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. *International Journal of Human-Computer Interaction* 37, 1 (2021), 81 – 96.

[38] Taejung Park Robert O. Davis and Joseph Vincent. 2021. A systematic narrative review of agent persona on learning outcomes and design variables to enhance

personification. *Journal of Research on Technology in Education* 53, 1 (2021), 89–106. https://doi.org/10.1080/15391523.2020.1830894

[39] Jürgen Rudolph, Shannon Tan, and Samson Tan. 2023. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *1* 6, 1 (2023). https://doi.org/10.37074/jalt.2023.6.1.23

[40] Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* 55, 1 (2000), 68.

[41] Magdalena Rychlowska, Antony S. R. Manstead, and Job van der Schalk. 2019. *The Many Faces of Smiles.* Springer International Publishing, Cham, 227–245. https://doi.org/10.1007/978-3-030-32968-6_13

[42] Noah L. Schroeder, Olusola O. Adesope, and Rachel Barouch Gilbert. 2013. How Effective are Pedagogical Agents for Learning? A Meta-Analytic Review. *Journal of Educational Computing Research* 49, 1 (2013), 1–39. https://doi.org/10.2190/EC.49.1.a

[43] Jun'ichiro Seyama and Ruth S. Nagayama. 2007. The Uncanny Valley: Effect of Realism on the Impression of Artificial Human Faces. *Presence: Teleoperators and Virtual Environments* 16, 4 (08 2007), 337–351. https://doi.org/10.1162/pres.16.4.337

[44] Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 35, 6 pages. https://doi.org/10.1145/3640794.3665887

[45] Carmen Vallis, Stephanie Wilson, Daniel Gozman, and John Buchanan. 2024. Student Perceptions of AI-Generated Avatars in Teaching Business Ethics: We Might not be Impressed. *Postdigital Science and Education* 6, 2 (2024), 537–555. https://doi.org/10.1007/s42438-023-00407-7

[46] George Veletsianos. 2010. Contextually relevant pedagogical agents: Visual appearance, stereotypes, and first impressions and their impact on learning. *Computers & Education* 55, 2 (2010), 576–585. https://doi.org/10.1016/j.compedu.2010.02.019

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

[48] Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. 2024. Building LLM-based AI Agents in Social Virtual Reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 65, 7 pages. https://doi.org/10.1145/3613905.3651026

[49] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large Language Models for Education: A Survey and Outlook. arXiv:2403.18105 https://arxiv.org/abs/2403.18105

[50] Joseph Weizenbaum. 1966. ELIZA—A Computer Program For The Study of Natural Language Communication Between Man and Machine. *Commun. ACM* 9, 1 (1966), 36–45.

[51] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating Systems: Gender, Race, and Power in AI. https://ainowinstitute.org/discriminatingsystems.html.

[52] Shan Zhang, Chris Davis Jaldi, Noah L. Schroeder, Alexis A. López, Jessica R. Gladstone, and Steffi Heidig. 2024. Pedagogical agent design for K-12 education: A systematic review. *Computers & Education* 223 (2024), 105165. https://doi.org/10.1016/j.compedu.2024.105165

[53] Sarah A. Zipp, Tyler Krause, and Scotty D. Craig. 2017. The Impact of User Biases Toward a Virtual Human's Skin Tone on Triage Errors Within a Virtual World for Emergency Management Training. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, 1 (2017), 2057–2061. https://doi.org/10.1177/1541931213601998