Charm-bots: The Impact of A.I.’s Sycophancy Language on User Trust

Molina, M., Lum, H. C., Evans, S., Inderberg, L. H. , Glass, I., & Michael, A.

AI in Education
Artificial Intelligence
chatbots
Large Language Models
student-AI interaction
sycophancy
tone
trust calibration

With the rapid advancement of artificial intelligence (AI) technologies, particularly Large Language Models (LLMs) like ChatGPT, it has become increasingly important for researchers to explore how these innovations influence human-technology interactions, education, and broader societal dynamics. The language used by chatbots can shape how users engage with them. This study aims to examine how the "tone" of AI affects users' perceptions of chatbots and their interactions with them. By comparing interactions in two contexts: objective (e.g., solving chemistry problems with chatbot assistance) and subjective (e.g., seeking advice for medical symptoms), the research also investigates whether context moderates these interactions. Ultimately, this study contributes to the growing body of knowledge on the factors influencing human-AI interaction and seeks to inform the design of educational AI systems that promote appropriately calibrated trust and support productive learning.

Introduction

As society heads towards an increasingly automated future, its implication on humanity becomes a significant area of focus. One sector of particular interest is human-robot interaction (HRI) research, or the study of human-robot communication—the exchange of information (Goodrich & Schultz, 2007) and, increasingly, student–AI interaction within educational technology and learning systems (e.g. intelligent tutoring systems) (Baker et al., 2022). Large language models (LLM), such as ChatGPT and Google's Gemini, have captured public attention, with students and the common person using these chatbots for various tasks (Chatterji et al., 2025). The literature on such advanced systems—which produce human-like responses based on large datasets of information—suggests that people do interact differently with chatbots than they do with other humans.

Amongst learning and decision making, subsequent behavior towards automation such as reliance is a significant topic of discussion in the context of understanding how HRI guides human behavior. Research indicates that trust is a crucial factor in predicting reliance on robotic systems, demonstrating a positive relationship (Hoff & Bashir, 2015; Lee & See, 2004). While trust and reliance research in AI is saturated, there remains a gap in work examining the interaction between students and learning systems. The design question must be reframed—not whether trust is higher, but whether student trust is calibrated and appropriate to the task context (e.g. when students rely on the AI and when they choose to seek a second opinion). However, there are caveats to this relationship, such that trust and reliance in AI can be rapidly restored after early and late incidence of error, although early errors have a much more significant negative impact on reliance compared to later errors made (Kahr et al., 2024). Shahrasbi (2025) proposes a theoretical experimental design in which participants interact with a

neutral tone agent and an empathetic tone agent, resulting in a preference for the empathetic tone agent for future use, which they identify as a behavioral measure of trust. They build on this by suggesting a second experiment, where individuals would complete a multi-step task with the option of switching between the two agents varying in tone after each response, hypothesizing that participants will be more tolerant of an erroneous empathetic agent if they began the task with it, thus providing an established bond.

The current study at hand serves to address the research gap on how chatbot tone and subjective versus objective contextual questions influence subsequent student trust calibration and help-seeking/continued-use intention toward an LLM-based AI, while investigating user attribute interaction effects. As chatbots increasingly appear in learning environments such as tutoring and educational decision-support, understanding how tone interacts with question type is directly relevant to inform learning system design. To our knowledge, the proposed framework is novel within the literature.

Methods

Measures

Participants will complete a post-study survey that includes demographic information and measures of GenAI usage, prior exposure and experience, and fairness-related questions. To capture individual differences relevant to technology adoption and responses to AI, established scales are included, such as the Technology Readiness Index (TRI) Parasuraman, 2000), the Big Five Personality Dimensions (Gosling et al., 2003), the Maranell Religiosity Scale (Maranell, 1974), and the AI Anxiety Scale (Wang & Wang, 2019).

Procedure

A between-subjects design will be used, where a total of 34 participants (students) randomly assigned to one of two groups. One group will interact with the LLM, ChatGPT, which will be personalized to use a “friendly” tone (n = 17), while the other group will interact with the same chatbot prompted to maintain a “neutral” tone (n = 17).

Once participants arrive at the study site, they will be given a consent form to read and fill out in order to proceed with the study. After participants provide their consent, they will interact with ChatGPT as a proxy for a chat-based AI tutor interface that could appear in a learning platform.

Participants will be tasked to ask ChatGPT objective questions; help understanding chemistry concepts (e.g. “What’s benzene’s most distinctive feature?”) and subjective questions; advice on vague medical symptoms (e.g. You lost your appetite since the fall). This will allow us to analyze how interactions differ between interpreting confusing course topics and engaging with other decision-support prompts where the “best” answer is subjective rather than strictly correct. After finishing the chatbot interaction, participants will complete a short series of questionnaires in Qualtrics on generative AI usage and preferences, personality test questions, technology acceptance, and demographics questions.

Results

This study investigates the influence of a chatbot’s language style on user trust and the potential moderating role of individual differences. We hypothesize that the use of sycophancy language by the chatbot will increase students’ self-reported trust and intention to continue using the chatbot, with trust being higher in objective domains than in subjective ones. This objective–subjective contrast is intended to surface whether trust is calibrated to question type. Additionally, we will examine whether individual characteristics affect trust, including familiarity with generative AI (genAI) technologies, attitudes toward new

technologies, AI-related anxiety, and personality traits such as extraversion, agreeableness, conscientiousness, and openness to experience. Specifically, we expect that users with greater familiarity with genAI technologies and more positive attitudes toward new technologies will exhibit higher trust, whereas those with higher AI-related anxiety will exhibit lower trust.

Discussion

If the hypotheses of this study are supported, the findings could have far-reaching implications for the design, deployment, and ethical governance of large language models (LLMs) in educational technology and learning systems (e.g., AI tutors). Demonstrating that a sycophancy tone increases user trust would suggest that developers could strategically use conversational style to enhance engagement and repeated use of AI systems in student–AI interaction. At the same time, such effects raise important concerns regarding transparency, informed user consent, and the potential for overreliance on AI-generated information. Dependence on AI could be particularly problematic in educational decision-making contexts, where reliance on inaccurate or misleading outputs could result in tangible consequences. Moreover, if individual differences such as familiarity with technology, personality traits, or AI-related anxiety are shown to influence trust, these findings could inform the development of adaptive AI systems that calibrate their communication style according to user characteristics, thereby balancing usability with caution and supporting productive learning. From an ethical standpoint, these results highlight the necessity of incorporating safeguards to prevent manipulation or unintended biases from user behavior. They also underscore the importance of promoting AI literacy and public awareness to ensure that users approach AI outputs with an appropriate level of critical evaluation. Ultimately, such insights contribute to a more responsible, user-centered approach to AI deployment, emphasizing both the benefits and ethical responsibilities associated with advanced generative technologies in learning engineering and learning system design.

References

Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. Technology, Mind, and Behavior. Advance online publication. https://doi.org/10.1037/tmb0000058

Goodrich, M. A., & Schultz, A. C. (2008). Human–robot interaction: A survey. Foundations and Trends® in Human–Computer Interaction, 1(3), 203–275.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the big five personality domains. Journal of Research in Personality, 37, 504-528.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434.

Kahr, P. K., Rooks, G., Snijders, C., & Willemsen, M. C. (2024, March). The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice. In Proceedings of the 29th International Conference on Intelligent User Interfaces (pp. 609-622).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

Maranell, G. M. (1974). Responses to religion: Studies in the social psychology of religious belief. Lawrence, KS: University Press of Kansas.

Parasuraman, A. (2000). Technology Readiness Index (TRI): A multiple-item scale to measure readiness to embrace new technologies. Journal of Service Research, 2(4), 307–320.

Shahrasbi, Nasser, "Empathy in AI Agents: Impacts on User Trust and Error Tolerance" (2025). AMCIS 2025 TREOs. 119.

Wang, Y. Y., & Wang, Y. S. (2019). Development and validation of an artificial intelligence anxiety scale: an initial application in predicting motivated learning behavior. Interactive Learning Environments, 30(4), 619–634. https://doi.org/10.1080/10494820.2019.1674887