Human-Computer Interaction & Emerging Technologies, Vol. 233, 2026, 90–97
https://doi.org/10.54941/ahfe1007507

User Perception of Response Inconsistency and Trust in AI-Assisted Learning

Ashwini Srinivasaprasad, Omer Bugra Kanbur, and Vincent G. Duffy

Edwardson School of Industrial Engineering, Purdue University, USA

ABSTRACT

Generative AI chatbots are increasingly deployed in educational settings, yet their inherent response variability may undermine user trust and long-term adoption. This study examined how perceived response inconsistency and response structure influence user trust, perceived learning, and task performance in AI-assisted learning. Thirty-one graduate students completed GRE-style verbal reasoning tasks with AI assistance delivered via a Wizard-of-Oz chatbot that systematically varied response style (Standard, Lengthy, Unstructured, Ambiguous) across trials, creating response inconsistency by design. Post-task surveys assessed perceived inconsistency, trust impact, and learning experience, while task accuracy served as the objective performance measure. Spearman correlations were employed given the ordinal nature of the survey data. Results revealed that participants who noticed greater response inconsistency reported significantly higher trust damage ($\rho = .617$, $p < .001$) and more negative perceived learning impact ($\rho = .499$, $p < .01$). Critically, however, neither trust nor perceived learning impact correlated with actual task accuracy. This perception-performance disconnect indicates that users felt their learning was impaired when they noticed inconsistency and lost trust, yet their objective performance remained unaffected. Additionally, 86% of participants identified unstructured responses as detrimental to comprehension, while emoji use was rated highly effective for understanding (M = 4.61/5). These findings suggest that response inconsistency poses greater risks for user trust and long-term engagement than for immediate task performance. Users may abandon effective AI tools they do not trust, highlighting the importance of evaluating subjective user experience alongside objective performance metrics in educational AI systems. Design implications include prioritizing structural consistency in AI responses and incorporating visual cues to aid comprehension.

Keywords: Generative AI, Trust, Response inconsistency, Human-AI interaction, User perception, Educational technology

INTRODUCTION

The rapid adoption of generative AI chatbots in educational contexts presents new challenges for understanding how users perceive and interact with AI-generated content. Unlike traditional automated systems that produce deterministic outputs, large language models exhibit inherent response variability where identical queries can yield substantially different outputs in terms of length, structure, and clarity. This variability manifests

Received February 7, 2026; Revised April 2, 2026; Accepted April 19, 2026; Available online July 20, 2026

as inconsistency from the user's perspective and may have significant implications for trust, a factor recognized as critical for sustained technology adoption.

Trust in automation has been conceptualized as a multidimensional attitude comprising beliefs about system reliability, competence, and predictability that influences user willingness to rely on automated systems under conditions of uncertainty (Lee & See, 2004). In the foundational trust literature, Mayer, Davis, and Schoorman (1995) identified ability, benevolence, and integrity as key antecedents of trust, while later work extended these concepts to human-machine interaction. Predictability is particularly salient in human-AI interaction because users develop expectations for consistent system behavior, and violations of these expectations can erode trust through both cognitive and affective processes (Hoff & Bashir, 2015). Research on trust calibration has demonstrated that both over-trust and under-trust lead to suboptimal human-automation collaboration, with under-trust resulting in disuse of beneficial systems (Parasuraman & Riley, 1997).

In educational contexts, cognitive load theory provides a framework for understanding how instructional design affects learning by emphasizing the limited capacity of working memory (Sweller, 1988). Response characteristics such as organization, clarity, and conciseness directly influence the extraneous cognitive load imposed on learners, potentially affecting both comprehension and performance. Recent research on conversational AI has examined factors influencing user satisfaction and engagement, finding that response quality dimensions including conversational contingency and perceived humanness significantly predict user trust and continued use intentions (Cheng et al., 2022). However, existing work has largely examined response quality and conversational attributes in isolation, leaving unclear how users interpret variability across responses and whether perceived inconsistency affects trust and learning independently of objective performance.

The present study addresses this gap by investigating three research questions: (1) Does perceived response inconsistency correlate with reduced user trust in AI-generated information? (2) Does perceived inconsistency affect users' subjective learning experience? (3) Do subjective perceptions of trust and learning impact relate to objective task performance? This work extends our prior research, which examined the relationship between AI response style and cognitive load using physiological measures (Srinivasaprasad et al., 2025). The present analysis focuses on survey-based measures of trust and user perception not previously reported.

METHODOLOGY

Participants

Thirty-one graduate students (16 male, 15 female) were recruited from Purdue University engineering courses. Participants ranged in age from 21 to 36 years. Sixteen participants were native English speakers and fifteen were non-native speakers. All participants had normal or corrected-to-normal vision and provided informed consent prior to participation.

Experimental Setup

The experiment was conducted in a controlled laboratory environment at the Human Integration Lab. Participants were seated at a workstation equipped with a 24-inch monitor displaying the AI chatbot interface. Eye movements and pupil diameter were recorded using Tobii Pro Glasses 3, a head-mounted eye-tracking system with a sampling rate of 100 Hz. The eye-tracker captured gaze position, fixation patterns, and pupil dilation as physiological indicators of cognitive load. Participants also wore inertial measurement unit (IMU) sensors to capture postural data during the task. The physiological data collection and analysis are reported in detail in our prior publication (Srinivasaprasad et al., 2025); the present paper focuses exclusively on survey-based measures of trust and user perception.

The AI chatbot interface was implemented using a Wizard-of-Oz methodology, wherein researchers delivered pre-scripted responses to participants' queries. This approach allowed controlled manipulation of response characteristics while maintaining the ecological validity of a conversational AI interaction. Participants were not informed that responses were human-controlled until debriefing.

Task and Response Style Manipulation

Participants completed GRE-style verbal reasoning tasks, including reading comprehension and text completion questions. The experimental session consisted of two phases: a baseline phase with 4 questions completed without AI assistance, followed by an AI-assisted phase with 16 questions. During the AI-assisted phase, participants could query the chatbot for help understanding passages and answering questions.

A within-subjects design was employed with response style as the independent variable. Participants experienced four distinct response styles across the 16 AI-assisted trials, with four trials allocated to each style. Standard responses consisted of concise bullet points with emoji symbols (checkmarks and crosses) to provide visual clarity regarding correct and incorrect answer options. Lengthy responses comprised detailed explanations spanning four to five paragraphs. Unstructured responses presented information in a bottom-up list ordered from least to most relevant. Ambiguous responses provided vague, indirect guidance lacking definitive answers. This design ensured that each participant experienced all four response styles within a single session, thereby creating response inconsistency by design.

It is important to note that response inconsistency in this study emerged from systematic variation in response style rather than stochastic AI behavior. As such, perceived inconsistency reflects users' subjective experience of variability in response structure, clarity, and guidance across interactions, rather than inconsistency in factual correctness. This approach allows examination of how users interpret stylistic variability as inconsistency, which is particularly relevant for real-world interactions with generative AI systems.

Survey Measures

Following the experimental task, participants completed a post-task survey assessing their perceptions of the AI assistant. Perceived inconsistency was measured with the item "Did you notice any inconsistencies in responses provided by the AI assistant?" on a five-point scale from Never (1) to Always (5). Trust impact was assessed with "Did inconsistencies in the AI's responses affect your trust in the information provided?" ranging from None at all (1) to A great deal (5). Learning impact was measured with "How did inconsistencies in the AI's responses affect your learning experience?" ranging from No impact (1) to Severe negative impact (5). Emoji effectiveness was assessed with "Did the use of emoji help better understand the answers provided by AI?" on a scale from Strongly disagree (1) to Strongly agree (5). A multi-select item asked participants to identify factors affecting their comprehension: lengthy responses, unstructured responses, ambiguous responses, and inconsistent responses. Task accuracy, operationalized as the proportion of correct responses during the AI-assisted phase, served as the objective performance measure.

Statistical Analysis

Given the ordinal nature of the survey data and non-normal distributions confirmed through preliminary analysis, Spearman rank correlations were employed to assess relationships between variables rather than parametric alternatives. Descriptive statistics were computed for all survey measures and task performance. Percentages were calculated for the multi-select comprehension factors item. Given the exploratory nature of this analysis and the sample size, the study focused on bivariate relationships to establish foundational patterns prior to more complex multivariate modeling in future work.

RESULTS AND DISCUSSION

Descriptive Statistics

Table 1 presents descriptive statistics for all survey measures and task performance. Participants reported relatively low levels of noticed inconsistency ($M = 1.97$, $SD = 1.05$), indicating that most participants did not frequently detect response variability. However, those who did notice inconsistency reported moderate levels of trust impact ($M = 3.35$, $SD = 1.20$) and learning impact ($M = 3.26$, $SD = 0.93$). Emoji effectiveness was rated highly ($M = 4.61$, $SD = 1.02$), with 93% of participants agreeing or strongly agreeing that emoji helped their comprehension. Overall accuracy during the AI-assisted phase was 74%.

Table 1: Descriptive statistics for survey measures and task performance.

Measure	M	SD	Range
Noticed Inconsistency	1.97	1.05	1-5
Trust Impact	3.35	1.20	1-5
Learning Impact	3.26	0.93	1-5
Emoji Effectiveness	4.61	1.02	1-5
Accuracy with AI	0.74	0.17	0-1

Relationship Between Inconsistency and Trust

Table 2 presents the correlation matrix for key variables. Participants who noticed greater response inconsistency reported significantly higher trust damage ($\rho = .617$, $p < .001$), representing a large effect size according to conventional interpretations (Cohen, 1988). This finding supports the theoretical proposition that predictability is a core dimension of trust in human-automation interaction (Lee & See, 2004). When AI responses varied unpredictably in format and style, users experienced this as a violation of expectations that undermined their confidence in the system.

Table 2: Spearman correlations between key variables.

Relationship	ρ	p
Noticed Inconsistency → Trust Impact	0.617	< 0.001
Noticed Inconsistency → Learning Impact	0.499	0.004
Trust Impact → Learning Impact	0.529	0.002
Noticed Inconsistency → Accuracy	0.068	0.716
Trust Impact → Accuracy	0.125	0.503
Learning Impact → Accuracy	–0.213	0.250

Note. N = 31.

Inconsistency and Perceived Learning

Perceived inconsistency was also significantly correlated with negative learning impact ($\rho = .499$, $p = .004$). Furthermore, trust impact and learning impact were themselves positively correlated ($\rho = .529$, $p = .002$), indicating a coherent pattern of negative user experience when inconsistency was detected. Participants who noticed more response variability tended to report both greater trust damage and greater perceived harm to their learning experience. This pattern suggests that trust and perceived learning effectiveness may share common antecedents in the consistency of AI behavior.

The Perception-Performance Disconnect

The central finding of this study is a clear dissociation between subjective user perceptions and objective task performance. Neither trust impact ($\rho = .125$, $p = .503$) nor perceived learning impact ($\rho = –.213$, $p = .250$) correlated significantly with actual task accuracy. Similarly, noticed inconsistency showed no relationship with accuracy ($\rho = .068$, $p = .716$). This pattern reveals a perception-performance disconnect: users reported feeling that their learning was impaired and their trust diminished, yet their objective performance remained unaffected by these perceptions.

This disconnect has important implications for AI system evaluation and adoption. Traditional performance metrics may fail to capture user experience dimensions that influence long-term engagement. Users who distrust an AI system may eventually abandon it regardless of its objective effectiveness, a phenomenon consistent with the disuse problem identified in the automation trust literature (Parasuraman & Riley, 1997). The finding suggests that designers and evaluators should monitor trust and subjective experience

alongside accuracy metrics. This pattern suggests that users' trust judgments may be driven more by experiential cues such as response predictability and organization than by the instrumental effectiveness of the AI system.

Factors Affecting Comprehension

When asked to identify factors that affected their comprehension of AI responses, participants most frequently selected unstructured responses (86%), followed by ambiguous responses (72%), lengthy responses (55%), and inconsistent responses (48%). Table 3 summarizes these results. The finding that unstructured responses were cited as problematic by substantially more participants than inconsistency itself indicates that organizational clarity is a primary driver of user comprehension and trust judgments, with perceived inconsistency likely serving as a proxy for structural unpredictability. Users appear to tolerate verbosity better than disorganization.

Table 3: Self-reported factors affecting comprehension.

Factor	n	% Selected
Unstructured responses	25	86%
Ambiguous responses	21	72%
Lengthy responses	16	55%
Inconsistent responses	14	48%

Note. N = 29 participants responded to this multi-select item.

Emoji Effectiveness

The use of emoji symbols in the Standard response condition was rated highly effective for aiding comprehension (M = 4.61, SD = 1.02). This finding aligns with research on visual cues in instructional design, which suggests that clear visual markers can reduce extraneous cognitive load by helping users quickly identify relevant information (Sweller, 1988). Simple emoji such as checkmarks and crosses appeared to provide effective signaling of correct and incorrect answer options, potentially reducing the cognitive effort required to parse text-heavy responses.

Limitations

Several limitations should be acknowledged. First, the single-session design precludes examination of trust dynamics over extended use; trust development and erosion likely unfold over multiple interactions, and users may calibrate their expectations after repeated exposure to an AI system. Additionally, trust was assessed using single-item self-report measures capturing perceived trust impact rather than validated multidimensional trust scales, and thus reflects subjective trust damage rather than calibrated trust as defined in the automation literature. Second, the Wizard-of-Oz methodology, while providing experimental control, may not fully represent the variability of actual generative AI systems in deployed settings. Third, the graduate

student sample, drawn from engineering courses at a single university, limits generalizability to other populations such as younger students or non-technical users. Finally, the correlational nature of the analysis prevents causal inference; experimental manipulation of inconsistency as an independent variable, rather than measuring perceived inconsistency post-hoc, would strengthen claims about its effects on trust.

CONCLUSION

This study demonstrates that users' trust and perceived learning in AI-assisted educational tasks can be undermined by stylistic variability even when objective performance remains unaffected. The results reveal a systematic misalignment between user trust judgments and actual task effectiveness, highlighting a critical challenge for the design and evaluation of educational AI systems. Specifically, the findings demonstrate that trust erosion can occur even when AI systems remain objectively effective, revealing a systematic miscalibration between user trust and system performance. Users may abandon objectively effective AI tools if those tools behave unpredictably and erode trust, highlighting the importance of trust calibration alongside performance optimization.

For the human factors community, these findings carry several practical implications. First, evaluation frameworks for educational AI systems should incorporate trust and user perception metrics alongside traditional performance measures. Second, the finding that unstructured responses were cited as problematic by 86% of participants suggests that designers should prioritize structural organization over mere consistency or brevity. Third, the strong endorsement of emoji effectiveness indicates that simple visual cues can meaningfully enhance information processing in text-based AI interfaces.

Future research should employ validated trust instruments, investigate longitudinal patterns of trust development and erosion across multiple sessions, and explore adaptive response strategies that maintain consistency while preserving the benefits of generative AI flexibility. As AI chatbots become increasingly prevalent in educational and professional contexts, understanding and designing for appropriate user trust will be essential for realizing their full potential.

REFERENCES

Cheng, X., Zhang, X., Cohen, J., & Mou, J. (2022). Human vs. AI: Understanding the impact of anthropomorphism on consumer response to chatbots from the perspective of trust and relationship norms. Information Processing & Management, 59(3), 102940. https://doi.org/10.1016/j.ipm.2022.102940

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.2307/258792

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Srinivasaprasad, A., Sepanloo, K., Jahromi, S. N., Yu, D., & Duffy, V. G. (2025). How AI chatbot response style affects cognitive load and performance in educational tasks. In C. Stephanidis et al. (Eds.), HCI International 2025 – Late breaking papers. Springer.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4