Currents of Inquiry: Insights From Two Years of Real-World AI-Learner Water Conversations

Shao, A., Carradini, S., Harikrishnan, R. S., Kuriakose, J., Manylal, S. , & Ravichandran, S.

Conversational AI | Learning Analytics | STEM education

Applying a Learning Engineering framework to informal STEM education, this study investigates the alignment between user inquiry patterns and conversational AI responses. We present an analysis of two years of interactions between our Waterbot and community users asking about water issues. Treating questions as signals of learners' cognitive focus, we code question types and quantify readability, causal connectives, and lexical markers of certainty for both users' questions and the AI's responses. Overall, learning sessions remained short, with learning depth tied more to chatbot verbosity than to linguistic complexity or certainty. These findings offer an initial design baseline for conversational AI that supports users' learning regarding complex issues. The observed asymmetries in length, complexity, and certainty point to concrete levers for the next design cycle, such as briefer default answers and clearer pathways for follow-up questions. As a study grounded in real learner traces, this work establishes an empirical baseline for the next engineering design cycle of conversational

systems aimed at civic and scientific informal learning regarding water issues.

Introduction

Conversational AI has become a common gateway for public information seeking (IPSOS, 2025). Large language models (LLMs) embedded in chat interfaces now function as informal tutors across a wide range of STEM topics. These conversational systems provide real-time interaction as a responsive partner, which may help with engagement and exploration of complex scientific concepts (Kerry et al., 2008). A growing body of research has focused on speculating learning experiences with AI in experimental settings (Carradini, 2024), yet relatively little is known about how real learners actually phrase their questions and how AI systems respond during unstructured, real-world interactions.

Research on AI tutors shows that conversational agents can assist with explanation, feedback, and step-by-step guidance. Recent studies on LLM-based systems indicate that conversational AI can offer coherent explanations, generate examples, and help learners express scientific ideas (Kasneci et al., 2023; Terzimehić, et al., 2025). However, several challenges persist. LLMs often produce lengthy or overly detailed responses (Su et al., 2025) and avoid taking clear positions on uncertain topics (Yona, Aharoni & Geva, 2024). Researchers have also warned about the problem of over-explanation, where educators may present too much information that overwhelm learners or discourage further inquiry (Fleisher et al., 2025). While many controlled experiments test AI tutors on structured tasks, far less is known about how people outside formal education settings use conversational AI when exploring scientific topics. Informal interactions may involve brief exchanges, loosely formed questions, and variable levels of engagement. Therefore, a challenge remains concerning misalignment between the efficiency-oriented nature of informal information seeking and the sustained cognitive effort required for deep scientific learning. This creates a potential pedagogical mismatch in which system outputs exceed the learner's capacity for productive engagement. Addressing this issue requires close analysis of how such interactions unfold in actual use contexts.

This study addresses these gaps by analyzing about 3,000 real exchanges between users and Waterbot, a water-related conversational system (Rajan, Carradini & Lauer, 2024). The Arizona Waterbot is a community-engaged artificial intelligence project that applies learning engineering principles to environmental communication. Developed in 2023, Waterbot provides clear, context-specific answers about Arizona's water management, conservation, and drought conditions. In addition to its technical structure, Waterbot serves as a pedagogical tool that supports civic learning by helping residents understand scientific and policy information, as well as identify practical ways to promote local sustainability. By analyzing user-chatbot interactions up to July 2025, the project reflects on connecting data, design, and educational science to create equitable, transparent, and human-centered learning solutions. Drawing on principles of learning engineering, the analysis examines linguistic and conversational features and how these elements correspond to AI-mediated learning about environmental issues. By studying authentic interactions, this pilot research offers an empirical basis for improving conversational design and promoting more effective inquiry in future learning systems.

This study frames Waterbot as a learning engineering system, developed to improve learner outcomes through data-driven iteration (Baker et al., 2022; Goodell, Kessler, & Schatz, 2023). While the full learning engineering cycle involves continuous refinement from problem definition, design, data collection, analysis, and redesign (Goodell, Kolodner & Kessler, 2022), this paper specifically reports on one analytic cycle embedded within the longer design trajectory. By examining authentic interactions from the system's initial deployment, we aim to generate evidence that informs the subsequent redesign phase. This approach ensures the AI evolves through human-centered engineering design to better support public engagement with complex environmental topics (Thai et al., 2022).

Methods

To systematically isolate and improve interaction dynamics, this analysis functions as a nested learning engineering cycle within the broader Waterbot design trajectory (Craig et al., 2025). This specific cycle targets the learning problem wherein informal civic learners often engage with complex environmental topics through brief, potentially fragmented interactions. We investigate the design assumption that specific system features may affect the learner's ability to sustain inquiry. To test this, we examine analytic evidence derived from linguistic traces of real-world use, allowing us to track how design decisions influence learner engagement (Totino & Kessler, 2024).

Data source

The dataset consisted of 20,779 raw user-AI interaction logs collected from Waterbot. Initial preprocessing involved removing duplicated entries and unusable inputs such as garbled machine-generated text, resulting in 4,595 cleaned records. To ensure analytic consistency, only messages automatically identified as English-language were retained. This yielded a final analytic dataset of 2,978 user queries paired with the chatbot's responses. Each interaction was tagged with a system-generated conversation ID, allowing individual messages to be grouped into conversational sessions.

The primary units of analysis were individual message pairs, consisting of (a) a single user query and (b) the chatbot's immediate response. All computations of linguistic features (e.g., readability, certainty, causal expressions) were performed separately for users and the chatbot. Conversation-level analyses used the conversation ID to aggregate message-level features. All analyses were performed in Python 3.11 using standard NLP and statistical libraries.

Key Variables and Operationalization

Question Type.

User inputs were categorized according to their interrogative form including what, how, why, who, where, when, yes/no, and non-question prompts. Classification was rule-based, using the first interrogative token or syntactic structure of the message. This variable informed analyses of inquiry depth and the distribution of surface- versus mechanism-oriented questions. Theoretically, question phrasing serves as the learner's entry signal (Chin & Osborne, 2006), distinguishing between users seeking static facts (surface learning) and those attempting to construct mental models (deep learning) (Marton & Säljö, 1976).

Causal Inference Expressions.

To identify causal reasoning, each message was coded for the presence of causal markers such as "because", "therefore", "so that", and "since". This approach aligns with work in computational linguistics and science education that treats causal connectives as indicators of mechanism-focused reasoning. Messages were labeled True or False depending on whether at least one causal marker appeared. In this educational context, causal markers serve as a proxy for mechanistic reasoning (Bachtiar, Meulenbroeks, & van Joolingen, 2022). Their presence reflects a shift from descriptive questions toward explanatory inquiry, which characterizes progression in informal science learning.

Readability (Flesch–Kincaid Score).

Readability was computed using the Flesch-Kincaid Grade Level that estimates the U.S. school grade required to comprehend a text (Flesch, 1948). The score is based on sentence length and syllable count, with higher values indicating more complex language. This measure allowed comparisons between the linguistic complexity of user queries and AI responses. This metric

is leveraged to estimate the cognitive load imposed by the AI. A large disparity between user and AI readability scores may indicate that the system's expert register limits the learner's ability to process and extend the information.

Certainty Level.

Certainty was measured using a lexicon-based method adapted from Pei and Jurgens' (2021) framework for identifying expressions of confidence and uncertainty in natural language. Their approach models epistemic stance through lexical cues associated with strong commitment (e.g., definitely, certainly) and hedging or uncertainty (e.g., possibly, likely). Guided by this framework, each message received a certainty score reflecting the density of certainty-related expressions, with higher values indicating greater epistemic confidence. Certainty was computed separately for user queries and chatbot responses to examine differences in communicative stance between the two parties. Certainty in this study was treated as an indicator of epistemic stance (Rubin, 2010, Schiefer et al., 2022). High user certainty may reflect confirmation-oriented inquiry rather than exploratory reasoning, whereas AI certainty indexes the level of authority expressed in system responses. Disparities between the two can identify cases in which the system does not exhibit the degree of hedging expected in scientific discussion of complex environmental topics.

Results

What are in the user questions?

Users submitted brief questions (M = 10.90 words, SD = 23.95) that were low in Flesch-Kincaid scores (i.e., linguistic complexity) (M = 6.35, SD = 4.81) and expressed high levels of certainty (M = 4.94, SD = 0.14). In contrast, chatbot responses were substantially longer (M = 102.53 words, SD = 76.77), displayed higher linguistic complexity (M = 13.91, SD = 5.73), and conveyed lower certainty (M = 4.63, SD = 0.53). Question types in user-bot messages are presented in Figure 1. The majority of user questions were lower-order forms such as what, yes/no, or topic fragments. Approximately 3% of questions were why questions. There were also a few cases where users asked which-type questions that appeared to require more cognitive effort, indicating that they may have already formed a specific idea before seeking an answer. Expressions of causal reasoning in user messages were relatively uncommon, compared to the more frequent causal indications in the chatbot's language.

Across the dataset of 347 conversations, most interactions were short, typically containing fewer than five conversational turns. These indicate that users primarily engaged in surface-level information seeking rather than extended inquiry or multistep exploration.

How does the AI respond to user questions?

Figure 2 presents a view of the flow from types of questions users asked towards the reading level of Waterbot. A Kruskal-Wallis H test was conducted to examine differences in chatbot response complexity (Flesch-Kincaid Grade Level) across question types. The test indicated a significant effect of question type on response complexity: $H(9) = 213.89$, $p < .001$. Median Flesch-Kincaid scores differed by question type, with the highest medians for why (16.12), yes/no questions (15.47), and when (15.24). The lowest reading level from chatbot answers were for who questions (10.70). These results suggest that the chatbot generated more linguistically complex responses when addressing causal or explanatory prompts compared to factual or identification questions.

To examine alignment between chatbot and user language patterns, Wilcoxon signed-rank tests were conducted for three paired measures: readability, message length, and certainty. The test revealed significant differences for all three comparisons. In each round of conversation, chatbot messages were significantly more complex in readability measures than user messages ($W = 192{,}545.00$, $p < .001$). Chatbot responses were also longer ($W = 92{,}861.00$, $p < .001$) and expressed lower

certainty (W = 764,598.00, p < .001). These results indicate that, while the chatbot tended to produce longer and more syntactically complex messages than users, its language conveyed comparatively less confidence.

How do message characteristics relate to conversation depth?

To examine how session-level characteristics related to conversation length, Spearman's rank-order correlations were conducted between the total number of conversational turns and key linguistic or behavioral variables averaged across each session. Conversation length showed a strong positive correlation with chatbot word count (ρ = .45, p < .001), indicating that longer sessions tended to include more verbose chatbot responses. A smaller but significant positive association was also found with chatbot causal language (ρ = .20, p < .001) and user causal language (ρ = .23, p < .001). All other relationships, including those with readability (Flesch-Kincaid scores) and certainty measures, were nonsignificant (ps > .05). Overall, these results suggest that sessions with more turns were characterized by higher chatbot output and slightly greater use of causal expressions by both the user and the chatbot, but not by increased linguistic complexity or confidence.

Discussion

This pilot study examined how individuals used a text-based AI system to ask questions about water-related issues and how the system responded. By analyzing message-level features across conversational turns, the results provide an initial view of how users engage with an AI information tool in water as a scientific domain and how linguistic patterns shape the structure of these interactions. By analyzing message-level features across conversational turns, the results provide an empirical basis for refining the system to better support public engagement with complex science and public issues.

Across conversations, users submitted short, low-complexity questions that aligned with surface-level information seeking. Most questions were what or yes/no forms, and only a small proportion were why questions. These patterns are consistent with prior research showing that learners often begin with basic inquiries when entering complex scientific domains (Tawfik et al., 2020, Vattam et al., 2011). Causal language was uncommon in user messages, suggesting that most users approached the system with topic interest compared to more advanced conceptual frames. In contrast, the chatbot produced responses that were substantially longer and more linguistically complex. Response complexity varied by question type, with higher complexity for causal or explanatory prompts. Paired comparisons showed that the chatbot consistently used longer sentences, higher readability scores, and lower expressed certainty than users. This echoes prior studies that LLM explanations often mismatch intended education level (Joshi et al., 2025). Although these features reflect the system's tendency to provide thorough explanations, they also indicate a clear asymmetry between user inputs and system outputs. Conversation depth was generally limited, with most sessions ending after only a few turns. The only strong predictor of longer sessions was the chatbot's word count, and weaker associations were observed for causal language use. Taken together, these findings suggest that most users engaged with the system for quick informational purposes rather than extended inquiry.

From a learning-engineering perspective, these results highlight several alignment considerations for future Waterbot system design.

First, the finding that lengthy, complex responses correlate with shorter sessions suggested to us a modification of the system prompt. The redesigned Waterbot needs to prioritize brevity and linguistic simplicity in initial turns to match the user's input style, rather than providing exhaustive explanations by default. Second, the scarcity of causal (why) questions indicates a need for explicit scaffolding. To support deeper exploration, we could replace the passive interface with a suggestion engine that offers optional follow-up questions rooted in causal mechanisms. Finally, the analysis of certainty markers highlighted a need for better calibration of the AI's epistemic stance. Users may phrase questions as assertions or confirmations, or seek for

validation instead of instruction (Gall, 1985, Martin-Arbos, Castarlenas & Duenas 2021), but the AI hedging could be due to safety alignment or policy constraints (OpenAI, 2025). Therefore, the next design phase could incorporate a retrieval-augmented generation (RAG) framework that explicitly cites sources, allowing the system to express higher certainty on established facts while reserving hedging only for truly ambiguous policy areas.

As a pilot project, this study has several limitations. The dataset includes only text interactions without behavioral or contextual data. Certain technical factors, such as the method used to assign session identifiers, may also have influenced the count of unique conversations. Even so, the patterns identified here provide a useful starting point for future research. Later studies could introduce experimental variations in response style or assess user perceptions through real-time feedback collected in an updated system. In light of these, post-July 2025 versions of Waterbot include mechanisms to record real-time user feedback, improved logs of the information sources retrieved during responses, and an interface that supports easier follow-up questions. These updates make it possible to examine how users respond to explanations, what sources they find helpful, and whether lighter or more adaptive responses lead to deeper engagement. Overall, this pilot provides an initial baseline for understanding how people use conversational AI to learn about water-related topics. The findings contribute to ongoing efforts in learning engineering to refine AI systems through iterative analysis of learner traces.

Figure. 1

Distribution of User Inquiry Types for Waterbot (N = 2.978). The most popular questions were instructional prompts, followed by “what” questions

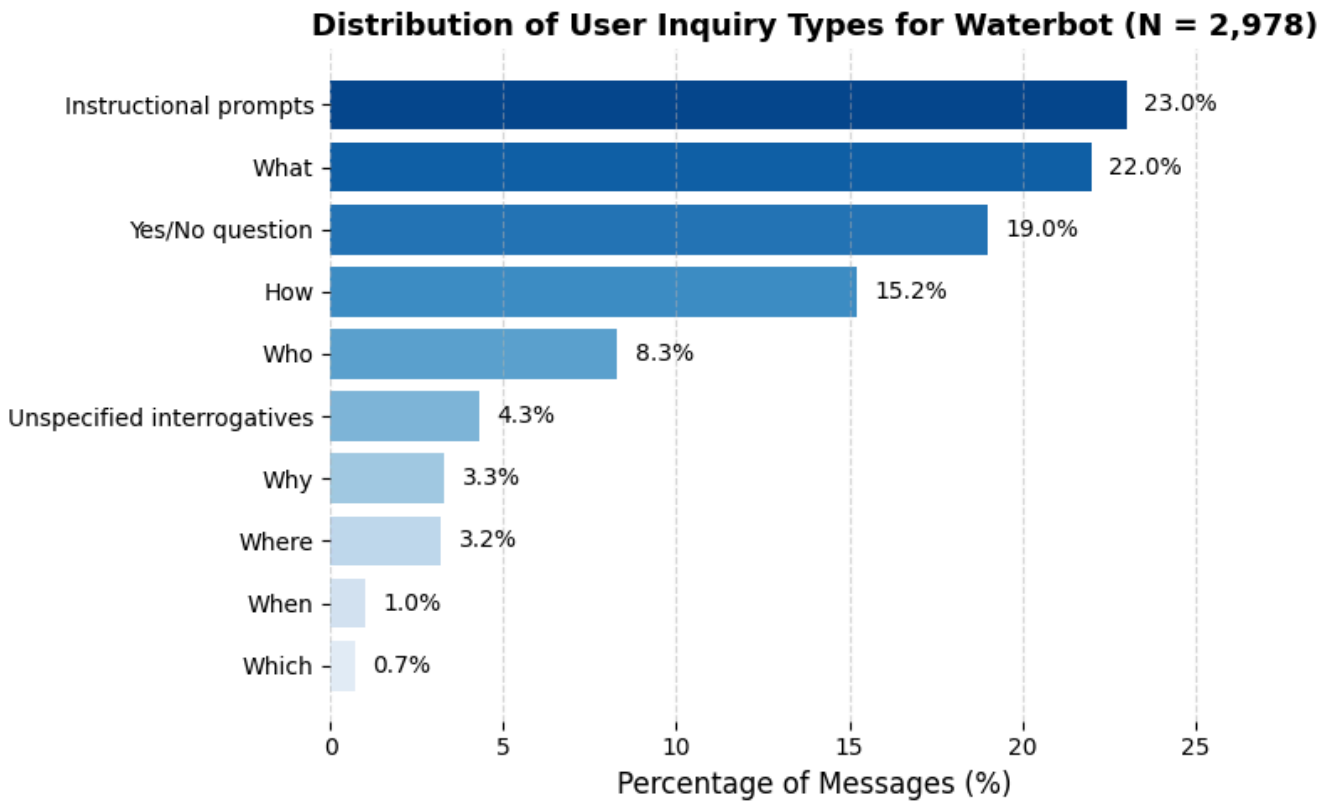
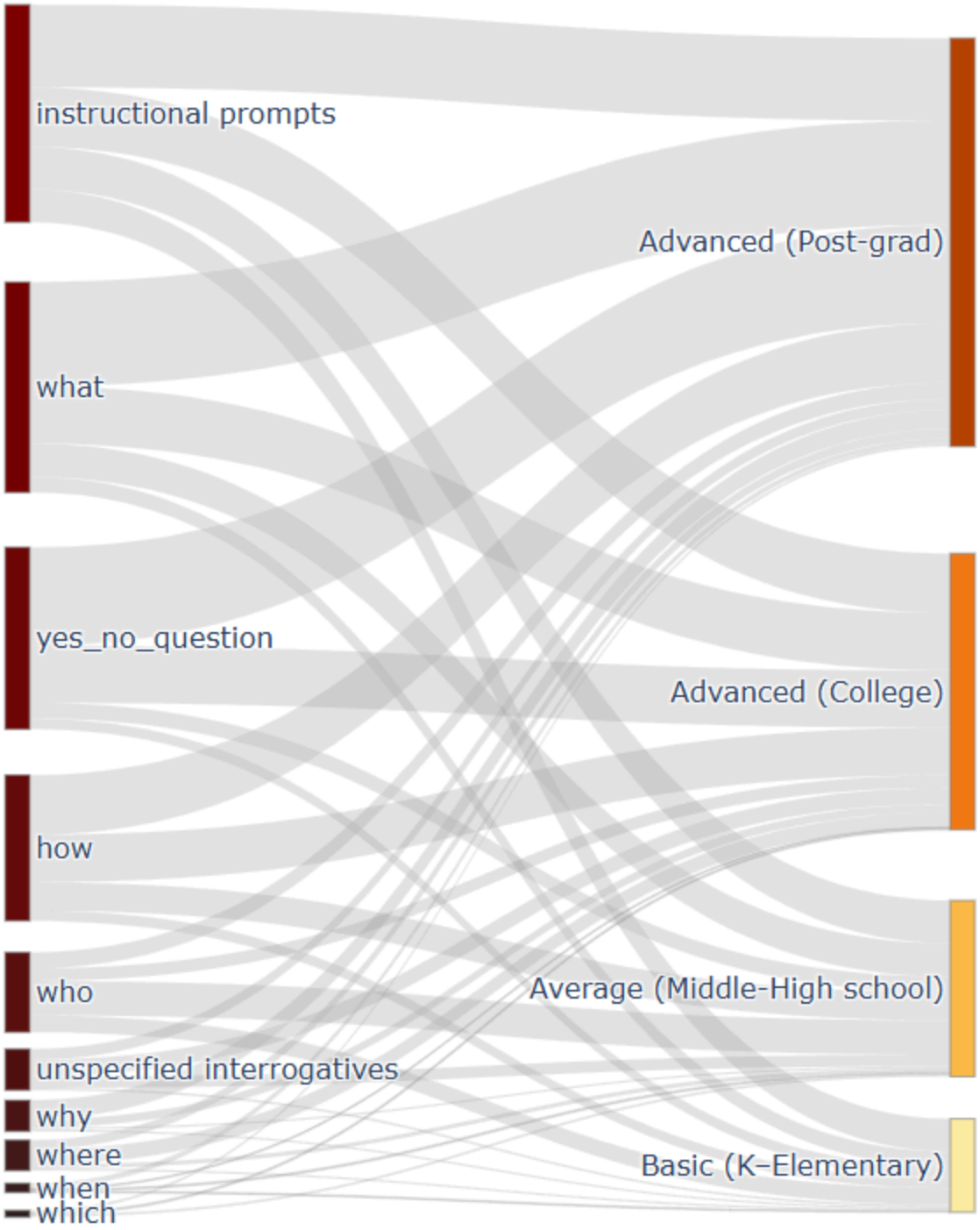


Figure. 2

Flow from User Question Types to Chatbot Readability Levels. Most chatbot responses are of advanced levels, and the answers are most complex for “why” questions.



Acknowledgments

The authors would like to acknowledge the generous support for this work provided by the Arizona Water Innovation Initiative (https://azwaterinnovation.asu.edu/).

References

Bachtiar, R. W., Meulenbroeks, R. F., & van Joolingen, W. R. (2022). Mechanistic reasoning in science education: A literature review. Eurasia Journal of Mathematics, Science and Technology Education, 18(11), em2178.

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.

Carradini, S. (2024). On the current moment in AI: Introduction to special issue on effects of artificial intelligence tools in technical communication pedagogy, practice, and research, part 1. Journal of Business and Technical Communication, 38(3), 187-198.

Chin, C., & Osborne, J. (2008). Students' questions: a potential resource for teaching and learning science. Studies in science education, 44(1), 1-39.

Craig, S. D., Avancha, K., Malhotra, P., Gorman, J. C., Verma, V., Likamwa, R., ... & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale.

Fleischer, H., Noglik, A., Borchers, C., & Schanze, S. (2025). Does Student Learning Rate Depend on Feedback Type and Prior Knowledge?.

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233.

Gall, S. N. L. (1985). Chapter 2: Help-seeking behavior in learning. Review of research in education, 12(1), 55-90.

Goodell, J., Kolodner, J., & Kessler, A. (2022). Learning engineering applies the learning sciences. In Learning engineering toolkit (pp. 47-82). Routledge.

Goodell, J., Kessler, A., & Schatz, S. (2023). Learning engineering at a glance. Journal of Military Learning, 7(1).

IPSOS. (2025). Did you know? As more people engage with AI tools, concerns persist despite technology's recognized role in enabling progress. Ipsos. https://www.ipsos.com/sites/default/files/ct/news/documents/2025-01/ipsos-essentials-infographic-january-2025.pdf

Joshi, B., He, K., Ramnath, S., Sabouri, S., Zhou, K., Chattopadhyay, S.,Chattopadhya, S., & Ren, X. (2025). ELI-Why: Evaluating the Pedagogical Utility of Language Model Explanations. arXiv preprint arXiv:2506.14200.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 103, 102274.

Kerry, A., Ellis, R., & Bull, S. (2008, December). Conversational agents in E-Learning. In International conference on innovative techniques and applications of artificial intelligence (pp. 169-182). London: Springer London.

Martin-Arbos, S., Castarlenas, E., & Duenas, J. M. (2021). Help-seeking in an academic context: A systematic review. Sustainability, 13(8), 4460.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. British journal of educational psychology, 46(1), 4-11.

OpenAI. (2025, January 23). Operator system card. https://openai.com/index/operator-system-card

Pei, J., & Jurgens, D. (2021, November). Measuring Sentence-Level and Aspect-Level (Un) certainty in Science Communications. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 9959-10011).

Rajan, B., Carradini, S., & Lauer, C. (2024, May). The Arizona water chatbot: Helping residents navigate a water uncertain future one response at a time. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-10).

Rubin, V. L. (2010). Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. Information Processing & Management, 46(5), 533-540.

Schiefer, J., Edelsbrunner, P. A., Bernholt, A., Kampa, N., & Nehring, A. (2022). Epistemic beliefs in science—a systematic integration of evidence from multiple studies. Educational Psychology Review, 34(3), 1541-1575.

Su, J., Healey, J., Nakov, P., & Cardie, C. (2025). Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. arXiv preprint arXiv:2505.00127.

Tawfik, A. A., Graesser, A., Gatewood, J., & Gishbaugher, J. (2020). Role of questions in inquiry-based instruction: Towards a design taxonomy for question-asking and implications for design. Educational Technology Research and Development, 68(2), 653-678.

Terzimehić, N., Bühler, B., & Kasneci, E. (2025). Conversational AI as a Catalyst for Informal Learning: An Empirical Large-Scale Study on LLM Use in Everyday Learning. arXiv preprint arXiv:2506.11789.

Thai, K. P., Craig, S. D., Goodell, J., Lis, J., Schoenherr, J. R., & Kolodner, J. (2022). Learning engineering is human-centered. In Learning engineering toolkit (pp. 83-123). Routledge.

Vattam, S. S., Goel, A. K., Rugaber, S., Hmelo-Silver, C. E., Jordan, R., Gray, S., & Sinha, S. (2011). Understanding complex natural systems by articulating structure-behavior-function models. Journal of Educational Technology & Society, 14(1), 66-81.

Yona, G., Aharoni, R., & Geva, M. (2024). Can large language models faithfully express their intrinsic uncertainty in words?. arXiv preprint arXiv:2405.16908.