Can Learners Navigate Imperfect Generative Pedagogical Chatbots? An Analysis of Chatbot Errors on Learning

Tiffany Wenting Li
University of Illinois Urbana-Champaign
Urbana, IL, USA
Stevens Institute of Technology
Hoboken, NJ, USA
wenting7@illinois.edu

Yifan Song
University of Illinois Urbana-Champaign
Urbana, IL, USA
yifan33@illinois.edu

Hari Sundaram
University of Illinois Urbana-Champaign
Urbana, IL, USA
hs1@illinois.edu

Karrie Karahalios
University of Illinois Urbana-Champaign
Urbana, IL, USA
kkarahal@illinois.edu

Abstract

Generative pedagogical chatbots offer a promising solution to transform personalized learning at scale, but their benefits are at risk because of the potential of providing inaccurate information. We have a limited understanding of how effectively learners handle factual chatbot errors and how these errors affect learners with varying backgrounds. This study addresses these questions in an ecologically valid open-ended online STEM learning environment. Using Bayesian causal inference and thematic analysis on survey and interview data from a quasi-experimental setting, we found that most participants struggled to detect factual errors even with access to reading materials and the Internet. Undetected errors harmed learning outcomes and self-efficacy, underscoring the need to help learners evaluate chatbot responses. By analyzing participants' evaluation strategies, we identified challenges during error management and suggested ideas on designing effective supporting resources and learner empowerment. Finally, we revealed differential impacts of chatbot errors across learners and called for personalized support and deployment.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → **Interactive learning environments**; • **Information systems** → *Information systems applications.*

Keywords

Pedagogical Chatbot; Conversational Agent; Error Management; Differential Impact; Fairness; Hallucination; Reliance; STEM Learning; Large Language Model

ACM Reference Format:
Tiffany Wenting Li, Yifan Song, Hari Sundaram, and Karrie Karahalios. 2025. Can Learners Navigate Imperfect Generative Pedagogical Chatbots? An Analysis of Chatbot Errors on Learning. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale (L@S '25), July 21–23, 2025, Palermo, Italy.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3698205.3729550

ACM ISBN 979-8-4007-1291-3/2025/07
https://doi.org/10.1145/3698205.3729550

1 Introduction

How well do learners manage errors when they seek information from generative-model-based pedagogical chatbots, and how do these errors impact learners? A pedagogical chatbot based on generative models promises to transform personalized learning at scale by offering timely, interactive support tailored to learners' individual needs [14, 93]. Yet, concerns persist regarding the reliability of such chatbots in educational settings [70, 95] due to frequent inaccuracies (e.g., hallucinations [6, 31] and logical fallacies [21, 71]) or incomplete information [20, 44]. We have a limited understanding of how effectively learners handle these errors and how such errors ultimately affect learning.

Although we see a recent surge of interest in generative chatbots, chatbots have been used in education for decades for question-answering [19], tutoring [91], and enhancing learner engagement [10]. Earlier systems commonly relied on rule-based or retrieval-based methods, which offered accurate responses but required costly domain modeling and had limited interaction scope [17, 22, 43]. In contrast, **generative pedagogical chatbots** leverage recent advances in natural language generation (NLG), such as large language models (LLMs), to produce flexible responses for a wide range of learner requests [13, 45, 54, 80, 90, 94]. However, since LLMs "learn" domain knowledge by capturing patterns from training corpora, generative pedagogical chatbots often yield inaccurate or incomplete information [82], resulting in incorrect programming feedback [7, 8], tutoring instructions [68], or homework solutions [3, 76].

When learners encounter erroneous information, whether in human-graded assignments [12, 29, 32], automated feedback [12, 51], or on the Internet [59], research has shown that undetected mistakes can harm learning. Conversely, error detection and correction, or the attempt to do so, sometimes bring positive impacts, such as evaluating peers' mistakes [26, 34, 55, 67] and receiving erroneous suggestions from a peer pedagogical agent [1, 41, 50].

While error detection is critical, learners often struggle [46, 51, 53, 89]. Generative chatbots may amplify these concerns. Learners may trust an eloquently phrased chatbot response more than solutions from a peer or a web search result [16, 25, 63, 75]. Furthermore, the convenience of on-demand dialogue interactions can reduce cross-checking behaviors [92]. These factors raise important questions about how effectively learners detect, interpret, and correct chatbot-generated errors in educational contexts and how these errors affect their learning outcomes, self-efficacy, and trust in the system.

In this paper, we systematically measure how well adult learners manage factual chatbot errors when seeking information to understand factual and conceptual STEM knowledge [5], and how these errors affect learning outcomes and self-efficacy. We chose to focus on factual and conceptual STEM knowledge because factual and logical accuracy is more critical for them compared to the subjective, open-ended content in non-STEM subjects. We structure our investigation with three research questions:

RQ1 How well do learners identify and correct factual chatbot errors with and without logical fallacies?

RQ2 How does a factual chatbot error affect learners' learning outcomes and self-efficacy on the topic?

RQ3 How do learners' prior knowledge, prior perceptions and experience with chatbots, learning motivations, and demographic factors influence their ability to manage chatbot errors and how much they get impacted by chatbot errors?

To answer these questions, we developed an ecologically valid learning platform that incorporated a state-of-the-art LLM and conducted a quasi-experimental study. Participants completed a "pre-test - learning session - post-test" learning cycle on two topics in Introductory Statistics with a chatbot introduced as imperfect, in either a survey (N = 180) or an interview setting (N = 26). We randomly assigned each participant a chatbot that is either completely accurate or would make one factual error (either with or without a logical fallacy). We performed a mixed-methods analysis with Bayesian causal inference on survey data and a thematic analysis on qualitative data. Our work contributes to the use of imperfect generative pedagogical chatbots in three ways.

Established the need to support STEM learners in chatbot error management. A majority of our participants struggled to detect and correct factual chatbot errors with alternative resources. Undetected chatbot errors harmed learning outcomes and self-efficacy, while detecting chatbot errors predicted learning gains vs. no error in some cases. This suggests that better detection of chatbot errors will not only protect learners from harm but also open up possibilities for educators to turn the errors into learning opportunities.

Showcased the challenges learners experienced during chatbot error detection. The reasons why participants failed to detect chatbot errors with the help of reading materials and Google search included relying on evaluation strategies dependent on accurate prior knowledge, having low confidence in their ability to evaluate, having unwarranted trust, and more. To overcome these challenges, we proposed ideas on designing for the effective use of supporting resources and learner empowerment.

Revealed potential fairness concerns across learner groups. We identified learner groups that were more vulnerable to the harms of imperfect chatbots, such as beginners on a topic, learners with limited experience with chatbots, and non-native English speakers. These differential impacts stressed the need to use caution during deployment and develop personalized support.

2 Related Work

2.1 Generative Pedagogical Chatbots and Errors

Recent advancements in natural language generation, especially with LLMs, have spiked interest in **generative pedagogical chatbots**, chatbots based on generative models developed and/or used for pedagogical purposes [45, 54, 69, 80]. However, despite their potential to accommodate a broader range of learner requests [13, 90], these systems sometimes provide *inaccurate or incomplete* information [20, 44, 78], including hallucinations [6, 31], logical reasoning failures[21, 71], or missing key details [20, 44]. ChatGPT, for instance, struggles with math problem-solving [6], runtime analysis [3], and retrieving factual knowledge about underrepresented regions [61]. In pedagogical contexts, studies found that GPT-generated feedback for programming contains inaccuracies in at least half of its responses [7, 8], sometimes even replicating the original errors from students [28]. A GPT-4-based tutor has been shown to mislead learners to incorrect math solutions [68]. These reliability issues have raised concerns among educators, researchers, and learners about the use of generative chatbots in education [70, 95].

2.2 The Impact of Chatbot Errors on Learners

Understanding the impacts of chatbot errors in educational contexts is crucial. Extensive studies have demonstrated that erroneous information from instructors, peers, or online resources affects learning outcomes and self-efficacy. Failing to differentiate reliable and unreliable online information can lead to misconceptions [59], and undetected incorrect feedback from human or automated systems impedes learning across tasks, such as match-to-sample exercises [12, 29] and programming code explanations [51]. Conversely, identifying and correcting errors in others' work (e.g., peers, virtual peer agents) can enhance learning and self-efficacy [26, 34, 41, 50, 55, 67], though these gains may vary with prior knowledge [1, 2].

Besides impacts on learning outcomes, errors from AI systems frequently erode trust and reduce users' willingness to rely on these technologies [16]. In educational contexts, framing automated evaluation tools as unreliable decreases learner acceptance [75], and perceptions of error rates in NLP-based autograders negatively impact satisfaction, causing students to overtrust positive and undertrust negative evaluations [30]. Therefore, examining the impacts of pedagogical chatbot errors may help optimize learners' performance and attitudes.

Prior research has examined learners' use of generative chatbots across subjects such as programming [36], STEM [52], creative tasks [79], and language learning [27, 60]. Yet, few have systematically studied how their errors affect learners, particularly in non-programming STEM contexts. For example, Lieb and Goel found that 40% of secondary students distrusted a physics chatbot but did not measure learning outcomes [52]. Therefore, we tackle this open question in an ecologically valid STEM learning context.

2.3 Learners' Management of Chatbot Errors

McBride et al. characterized human management of automation errors into three phases: detection, explanation, and correction [56].

Error detection and correction are particularly relevant in the chatbot error management context, as literature has shown that how learners manage and use erroneous information moderates its impact [12, 12, 29, 32, 51]. In non-educational information-seeking and task-completion contexts, prior studies have found that users frequently struggle to detect chatbot errors [38, 40, 65, 81, 84]. For instance, users perform worse when relying on AI-infused search engines that provide incorrect medical answers [40], overtrust misleading product comparisons from LLM-based tools [84], and often fail to detect inaccuracies from imperfect assistants in productivity tasks [38]. These studies, however, provided limited access to alternative resources for error detection and rarely investigated users' error correction behaviors. In contrast, educational settings usually provide learners with multiple alternative resources, which may affect the error detection and correction outcomes.

However, no work has systematically examined how well learners manage chatbot errors in a non-programming STEM learning setting. For instance, Shoufan observed that students without prior knowledge struggled to evaluate ChatGPT's solutions in a computer science exam setting, but not in a learning setting [83]. Kazemitabaar et al. observed students using verification strategies on chatbot-generated code but sometimes failed, yet code verification is different from evaluating natural language responses [36]. Our work examines how well learners detect and correct errors in natural language chatbot responses in STEM learning.

2.4 Chatbot Error Management Variations Across Learners

Individual learner attributes may shape how people evaluate, interpret, learn from, and respond to AI-generated information. Inspired by prior work, we examine the following attributes:

Prior Knowledge & Motivations. Foundational knowledge or confidence (e.g., high pre-test scores, STEM background, self-efficacy) may reduce learners' reliance on chatbots and enhance error detection [1, 33, 77]. Perceived task value and motivation (intrinsic or extrinsic) may encourage self-regulated learning behaviors such as verifying chatbot outputs using multiple resources [64].

Chatbot Experience & Perceptions. A high level of trust in the chatbot may cause some learners to accept chatbot outputs uncritically, while those with low trust may develop algorithm aversion [16, 39, 86]. Users who frequently engage with chatbots often develop folk theories to guide their detection and correction of errors [18, 37].

Demographics. Factors such as age [4, 88], gender [11, 74], education level [85], and English proficiency [66] may influence trust, error-attribution patterns, and cognitive load during chatbot interactions. For example, females may attribute failures internally more frequently [11], and non-native English speakers may struggle with cognitively demanding chatbot responses [66].

3 Learning Context and Platform Development

3.1 Learning Context and Procedure

We situated our study in the STEM learning context of introductory Statistics. This subject is accessible to diverse learners and contains independent, easy-to-evaluate concepts. We selected four distinct topics: (1) *mean vs. median*, (2) *simple random sampling vs. stratified sampling*, (3) *correlation vs. causation*, and (4) *standard deviation vs. standard error.*

Participants aimed to "understand" factual and conceptual knowledge in two of the four topics, the second level of Bloom's taxonomy [5]. Overall, a participant completed a pre-test, a learning session, and a post-test to achieve a set of learning objectives for each topic. They repeated this process twice, once for each topic. The experimental intervention, a chatbot that may make errors of a certain type, took place in the learning sessions. The first two authors, both with extensive backgrounds in Statistics, iteratively designed the learning and evaluation materials with feedback from a Statistics instructor at our university.

3.2 Learning Environment

We iteratively designed and implemented an open-ended online learning environment where participants had access to an imperfect chatbot using React and Django. We prioritized ecological validity when designing the learning experience and platform. Therefore, we made the following design decisions.

First, we used three learning objectives as an anchor to guide learners through each topic. They appeared as guiding questions at each step of the process (e.g., *"What is correlation, and what is causation?"*)

Second, we provided multiple learning resources offered in popular open-ended online learning environments along with the chatbot during the learning sessions (Figure 1). We encouraged participants to use the reading materials and a Google search engine alongside the chatbot to learn to answer the learning objective guiding questions. They were free to choose whether to use them, and when and how. Participants on average sent 2.67 messages to the chatbot, spent 3.43 minutes on the readings, and used Internet search 0.66 times per session. In addition, we designed a set of T/F practice problems to guide their learning. Upon submission, they immediately received answer keys and explanations as feedback.

Finally, we provided incentives to motivate participants' learning. We offered performance-based bonuses on the practice ($0.1/problem) and post-test problems ($0.4/problem) as extrinsic motivation. To stimulate intrinsic motivation, participants started each topic with a 2-minute video that used real-life examples to explain how the topic is relevant to daily lives. The median time spent on each learning session was 8 minutes, showing reasonable engagement.

3.3 The Imperfect Chatbot

3.3.1 The Error Types. This paper focused on two types of factual errors. Initially, we designed and collected data for a third error type, information omission errors (IO). Due to a design flaw in the related materials, we excluded the data points related to IO errors from the analysis.

The first type was a response with **factually incorrect information supported by a formal logical fallacy (FE with LF)** such as affirming the consequent, denying the antecedent, or undistributed middle-term. For example, the chatbot might state: "*Stratified sampling produces a more representative sample for a heterogeneous*

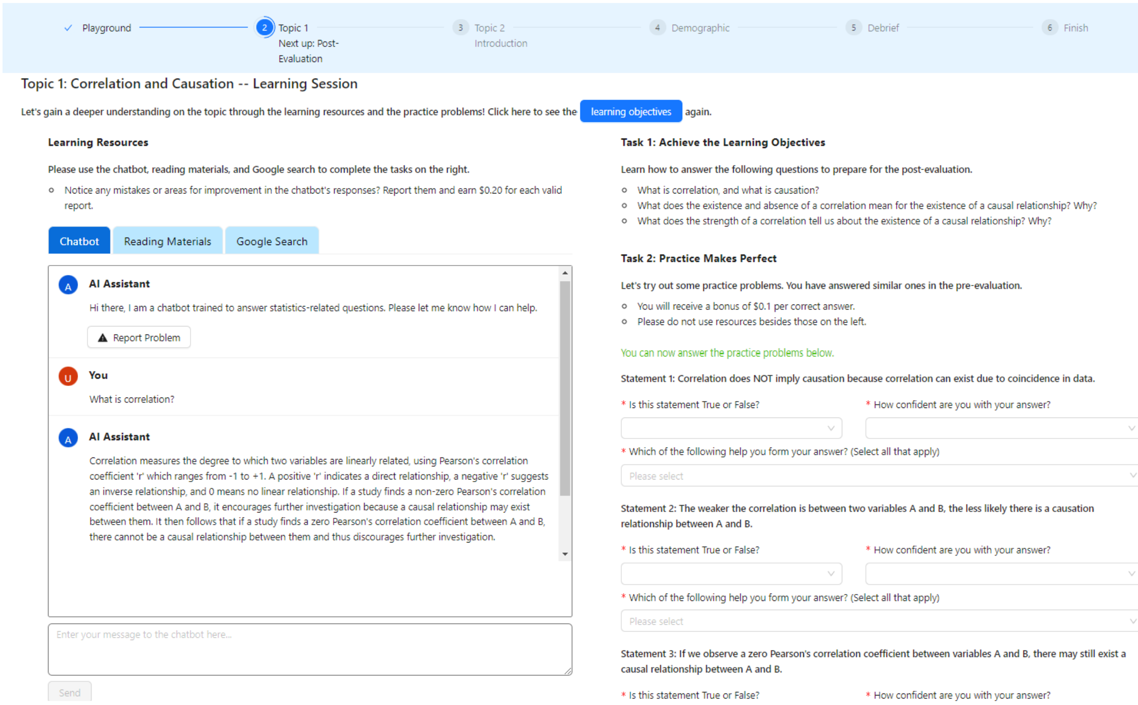

Figure 1: The learning session page with the chatbot, reading materials, and Google search engine on the left, and learning objective guiding questions and T/F practice problems on the right.

population than simple random sampling. Therefore, if the representativeness of a data sample improves after switching from simple random sampling to stratified sampling, then this sample must come from a heterogeneous population." Here, let's call "*the population is heterogeneous*" [A] and "*a sample from stratified sampling is more representative than that from simple random sampling*" [B]. This statement then follows the form "*If [A], then [B]. Therefore, if [B], then [A]*". This is an affirming-the-consequent fallacy because [B] cannot necessarily imply [A], i.e. there could be other reasons why a simple random sample is not representative.

Another type of factual error we studied was a response containing **incorrect factual information *without* a formal logical fallacy (FE without LF)**. For example, "*Sample size affects the magnitude of standard deviation (SD) and standard error (SE) in opposite ways. As the sample size increases, SE decreases because there is less uncertainty. But the SD of a sample increases as the sample size increases because more data points mean more variability.*" It incorrectly implies that SD always increases with sample size.

3.3.2 *Curating Chatbot Errors.*

We manually curated two unique factual chatbot errors per topic, one of each type. This resulted in 8 unique factual errors for 4 topics. These error statements were chosen from an initial set of 26 for their average difficulty based on a pilot study, where 95 participants on Amazon Mechanical Turk rated the correctness of the statements (see Appendix A.2[1]).

3.3.3 *How Participants Experienced Chatbot Errors.*

For each chatbot error statement, we created a corresponding learning objective, pre-test problem, practice problem, and post-test problem. This yielded three sets per topic, one for each error type[2]. In each learning session on a topic, a participant worked on all three practice problems, but they were randomly assigned to experience one of the error types or no error. We used rule-based randomization to customize prompts, instructing the chatbot to present up to one unique error of the randomly assigned type in each learning session. The same error may appear more than once per session. It was also possible that the assigned error did not appear if the participant did not ask any relevant questions. We categorized sessions in this case in the "no error" condition during the analysis.

3.3.4 *Implementing The Imperfect Chatbot.*

We used a combination of semantic cache and prompt chaining to create the imperfect chatbot. The chatbot first searched for a cached inquiry with high semantic similarity from a bank of inquiry-response pairs designed for the assigned error condition and topic. We used GPT-4 Turbo (0125-Preview) to generate the responses and variations of the inquiries with human oversight. This approach enhanced consistency in how errors were introduced. The overall cache hit rate was 87.6%.

If the program found nothing in the cache, it prompted GPT-4o (2024-05-13) for a response. The response then went through a chain of prompts uniquely designed for each error statement to incorporate the error and remove contradictions[3]. The first author

[1] All appendixes in the paper are available at https://osf.io/5wj3x/?view_only=c810f8f9418a4068a6a363c6483e1994

[2] One of them was for the IO error, which we excluded in the analysis.

[3] Prompts available in the "Chatbot Prompts" file here: https://osf.io/5wj3x/?view_only=c810f8f9418a4068a6a363c6483e1994

reviewed and categorized all dynamic responses to the appropriate error condition. We saw no unintended errors in the responses. The average delay for these dynamically generated responses was 5.7 seconds (standard deviation = 3.9 seconds). We added an artificial delay of 5 to 15 seconds to cached responses to reduce their distinguishability.

4 Study Design

Our experiment employed a within-between-subjects design. Participants experienced two chatbot error conditions out of four at random (three error types + a no error condition), as specified in Section 3.3. Note that while we collected data on a third type of error, information omission error, we excluded it from the analysis in this paper due to design flaws. We implemented the experiment via an online survey setting and an interview setting. In this section, we first explain the study procedures for the two settings, focusing on how we measured the variables of our interest. Then, we describe our data collection process and the samples.

4.1 Online Survey Procedure

This section outlines the step-by-step procedure for survey participants. Please refer to Appendix A.1 for the survey questions.

4.1.1 Part 1: Before Learning. **Prior chatbot experience survey:** Upon informed consent, we used Likert questions to survey participants' *prior familiarity, usage, perceived reliability,* and *perceived helpfulness of LLM and LLM-powered chatbots (RQ3).* If they had never heard of or used LLM-powered chatbots, we asked the same questions concerning general chatbot experiences instead.

Introducing the learning platform: We then introduced the study procedure and incentive structure. To familiarize participants with the learning environment, we explained how to use the resources and guided participants through practice tasks in a playground environment. We specifically emphasized that the chatbot *"may make mistakes or provide sub-optimal answers, even though it can answer many questions well".*

4.1.2 Part 2: During Learning. Overall, participants completed a pre-test, a learning session, and a post-test for both topics. Next, we describe the procedure for each topic.

Pre-learning assessment and survey: After viewing the introductory video, participants answered and reported their confidence in six True/False questions to demonstrate their *prior knowledge of the topic (RQ3).* Participants then reported their *self-efficacy in mastering the topic (RQ3)* on a 10-point scale adapted from Bandura et al. [9]. We also measured *intrinsic and extrinsic motivations*, and *perceived task value (RQ3)* using a subset of the Motivated Strategies for Learning Questionnaire (MSLQ) [72].

Learning session: Participants worked towards the learning objectives and prepared for the post-learning assessment at their own pace using the chatbot, practice problems, and other resources we provided (described in Section 3.2). We encouraged participants to submit a "Report Problem" form if they identified any issues with the chatbot's response with a bonus of $0.2/report. Participants explained what was wrong, why, and how they noticed it in the report form. Based on these reports, we derived *error management outcomes (EMO) for RQ1.*

Post-learning survey and assessment: First, we measured participants' *post-learning self-efficacy (RQ2)* [9] and attitudes towards the chatbot. Then, participants completed three assessment problems, each a proxy of *learning outcomes (RQ2)* to a learning objective. All problems contained a short-answer component to capture nuanced learning progress.

4.1.3 Part 3: After Learning. **Demographic survey:** After learning both topics, participants reported their *age, gender, education level, English proficiency,* and *STEM background (RQ3).*

Debrief: Finally, participants reviewed their chat history, which included annotations highlighting and correcting the known chatbot errors. This step ensured participants did not leave with misconceptions due to the chatbot's mistakes.

4.2 Interview Procedure

Interview participants completed the same surveys and tasks as online survey takers in the presence of an interviewer. We interviewed participants at three different points during their learning.

- **When participants opened a chatbot response report form,** we asked them to explain the issue, how they first noticed it, and how they decided to report it.
- **After the second learning session,** we asked participants why and how they used the various learning resources. We inquired about their thoughts on the quality of the chatbot responses and the strategies they used to assess them.
- **After the second post-learning survey,** we asked participants to explain the rationale behind the changes in their self-efficacy before and after learning.

4.3 Data Collection

Data collection took place between June and August 2024, following approval from our Institutional Review Board (IRB24-0024). We used a stratified sampling approach by setting a quota for participants with varying levels of statistical background to recruit a balanced sample.

We recruited 196 U.S.-based survey participants on Prolific [4] in two batches, half in June (with 3 learning topics), and half in July (with all 4 topics). Since each participant completed two topics, we started with 392 participant-topic pairs. We removed 48 pairs where participants failed two or more attention checks (N = 6) or showed either minimal effort (N = 30) or possible cheating behaviors using outside resources (N = 17) in the post-test[5]. The first two authors made exclusion decisions based on the language style and log data evidence (details in Appendix D). After removing 82 learning sessions in the information omission error condition, we got 262 valid participant-topic pairs from 177 participants. The median completion time for the study was 50 minutes. Participants received a base pay of $8/hour with a bonus of up to $3.5.

We recruited 26 interview participants through physical flyers, our institution's mailing list, and local social media groups (Reddit). 16 participants attended in person ($28 base pay + bonus) and 10 attended via Zoom ($20 base pay + bonus). The first two authors conducted the interviews with a median length of 77 minutes.

[4]https://www.prolific.com/
[5]Some instances overlapped across the removal categories.

4.4 Sample Descriptive Statistics

Our online survey sample (177 participants) closely matched the U.S. population in gender and ethnicity[6] but skewed younger and more educated. The sample included a balanced distribution of statistics knowledge: 24.71%, 36.34%, 26.45%, and 12.5% scored a weighted total[7] of ≤ 1, 2, 3, ≥ 4 points out of 6, respectively, in pre-tests. The interview sample skewed older and more educated compared to the survey sample. Additionally, more interviewees had a STEM background and had taken one or more Statistics classes.

Around 70% of both samples reported being slightly or moderately familiar with LLMs. About 20% had never used an LLM chatbot, while approximately 14% used it daily. Roughly 70% perceived chatbots as sometimes or often reliable, and 60-65% viewed them as moderately or very helpful. Further details for the samples are available in Appendix D.

5 Data Analysis

5.1 Models And Variables Definitions

Our research questions ask about how well participants managed the two types of factual chatbot errors (RQ1) and how the experience with a chatbot error impacted learning outcomes and change in self-efficacy (RQ2). In addition, we investigated how the effects vary by participants' characteristics (RQ3). In this section, we formally define the outcome variables (DVs) and predictors in the five models we constructed to answer the research questions.

5.1.1 Error management outcome (EMO) in RQ1. **Error management outcome (EMO)** is the stage at which a participant arrived in managing the chatbot error in a learning session. The first author analyzed the chatbot reporting forms and assigned each session one of the four outcomes: (1) no report, (2) reported a response but did not identify a valid error, (3) identified a valid error but failed to correct it, and (4) identified and corrected a valid error (See Appendix B for detailed definitions). We constructed two models.

Model 1 *[EMO (categorical) ∼ Session error type (categorical)]* estimates how likely participants reached each stage when managing each error type (RQ1). *Session error type* captures whether a participant encountered a given error type in a learning session.

Model 2 *[EMO (binary) ∼ Participant characteristics × Session has error (binary)]* estimates how participants' characteristics affect whether they successfully identified a chatbot error, i.e., reached stage 3 or 4 (RQ3), via interaction effects.

5.1.2 Learning outcomes in RQ2. Recall that each chatbot error corresponds to a unique practice problem and post-test problem. To estimate how experiencing a chatbot error impacted participants' learning outcomes, we modeled two metrics: **(1) correctness of the practice problem and (2) performance on the post-test problem** corresponding to the chatbot error.

Model 3 *[Correctness of a practice problem (binary) ∼ Problem has error × Problem's potential error type + Problem has error × Participant characteristics]* estimates a chatbot error's influence on learning outcomes when participants had access to reading materials and Google search. The variable *Problem has error* takes three values: no error, a relevant error, and an irrelevant error (i.e., experienced a chatbot error unrelated to the problem).

Model 4 *[Performance on a post-test problem (ordinal) ∼ Problem has error and its detection status × Problem's potential error type + Problem has error and its detection status × Participant characteristics]* estimates how a detected or undetected chatbot error affects learning outcomes after participants received direct feedback on the learning task (i.e., answer keys and explanations). The variable *Problem has error and detection status* has five possible values: no error, a detected/undetected relevant error, and a detected/undetected irrelevant error.

There are five levels to one's performance on a post-test problem. A participant earned the highest level if they selected the right answer and explained its rationale. An answer received the lowest grade if it aligned with the chatbot error. The first two authors went through three rounds of grading calibration and reached an agreement percentage of over 0.9 before independently grading half of the answers. The codebook is available in Appendix B.

5.1.3 Self-efficacy change in RQ2. We modeled the difference between participants' self-efficacy before and after each learning session in **Model 5** *[Difference in self-efficacy (continuous) ∼ Session error type and its detection status + Session has error and its detection status × Participant characteristics (excluding pre-learning self-efficacy]*. The variable *Session error type and detection status* has 5 categories: no error, detected/undetected factual errors with/without logical fallacies.

5.2 Causal Modeling

We aimed to estimate the causal relationships between the predictors and outcome variables (DVs) in the five models above. To identify confounding variables that affected both the predictors and DVs, we constructed a directed acyclic graph (DAG) (Figure 2) via a two-step process. First, we constructed the nodes (representing the variables) and arrows (representing the causal relationship between variables) directly supported by the research questions, prior literature [1, 4, 11, 16, 18, 33, 37, 39, 64, 66, 74, 77, 85, 86, 88], and experiment design. Then, we tested the implied independence and conditional independence of the DAG after data collection to verify its validity (process detailed in Appendix C). Based on the DAG, we adjusted for the following variables in our models.

Participant's inquiry behaviors that affect the likelihood of experiencing errors: The experiment design determined that the error type a participant experienced (a predictor) was affected by the inquiries the participant sent to the chatbot. In 19.5% of the learning sessions, participants did not encounter any chatbot error even though they were assigned a chatbot that made errors because they did not ask relevant questions. Participant characteristics may affect the inquiries a participant sent to the chatbot, which then affects the error type predictor, as well as the DVs (e.g., learning outcome, EMO). Therefore, we included a binary variable for *whether a participant sent the chatbot an error-inducing inquiry in a session* per error type in the models to block off the path from participant characteristics to the error type predictor.

Factors from the experiment design: We adjusted for individual error statements to tease apart the effect of an error type

[6] https://www.census.gov/

[7] pre-test score weighted by confidence rating on each question

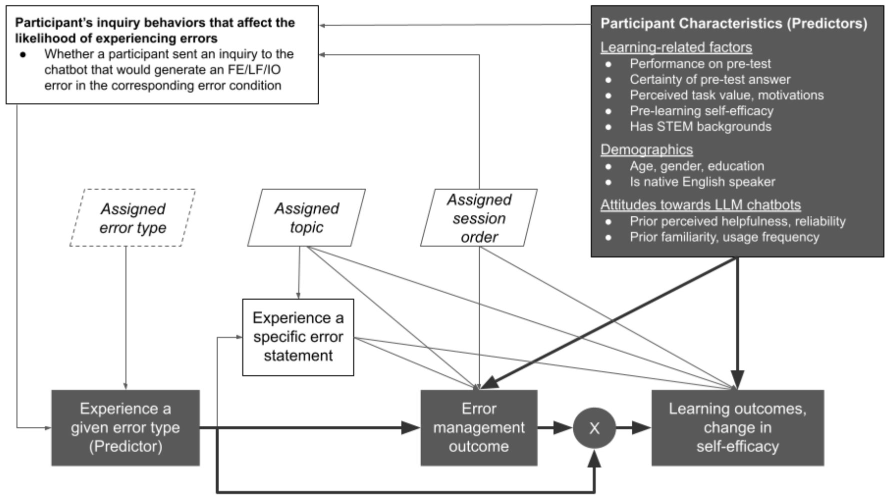

Figure 2: A DAG that describes the causal model we created for the outcome variables (DVs) and predictors in our research questions (filled boxes). We used this causal model to construct our regression models to ensure an unbiased estimation of the causal arrows of our interest (bolded). Parallelogram boxes represent variables randomized in the experiment design. The boxes with dotted edges were not included in the models.

from that of a particular error statement. In addition, we accounted for ordering effects [15, 87] and topic effects.

5.3 Bayesian Inference

We operationalized the models with Bayesian formulations of hierarchical generalized linear models [57] because Bayesian inference in HCI [35] and educational research [42] has a number of advantages. Bayesian modeling requires priors, which embody prior beliefs and evidence about the distribution of modeled variables and effect sizes. We set weakly informative priors by selecting maximum entropy distributions and setting prior distributions for parameters that encode skepticism but not impossibility towards large effect sizes. Given the priors and observed data, Bayesian inference generates posterior distributions that represent the likely values of regression coefficients. See Appendix E for details.

To obtain our posterior distributions, we implemented the models with NumPyro, a popular Bayesian inference framework, and performed sampling using the Markov Chain Monte Carlo (MCMC) technique with the No-U Turn Sampler (NUTS). All parameters achieved a Gelman-Rubin statistic (a measure of MCMC convergence) of 1.0, indicating that the multiple sampling chains converged [24]. We additionally verified the inference process by comparing the predicted values of the outcome variables against the real data and observed a reasonable fit for our data.

5.4 Qualitative Analysis

The first author of the paper performed affinity diagramming and thematic analysis on the interview data and chatbot issue report forms. The analysis focused on participants' self-reported chatbot error detection strategies, rationales behind their strategies, and failures and challenges they encountered in the process.

6 Results - Part I: Main Effects of Chatbot Errors (RQ1, RQ2)

In this section, we present our findings on how well participants identified and corrected the chatbot errors (RQ1) and how these errors impacted participants' learning outcomes and self-efficacy (RQ2) across two types of factual errors. We report the effects on practice problem (logistic regression) and post-test problem performance (ordinal regression) with odds ratios and cumulative odds ratio, respectively. The effects on self-efficacy change (linear regression) are measured with Cohen's d. An effect is considered significant when the 94% Highest Posterior Density Interval (HPDI) excludes the reference value that indicates no effect (odds ratio = 1, Cohen's d = 0). If an effect doesn't meet the significance threshold above, we say it is "close to significant" when there is a greater than 89% probability that the effect is consistently greater or smaller than the reference value.

Overall, participants had difficulty identifying and correcting factual errors in the chatbot. Our survey participants identified 19.77% and 24.62% of the factual errors with and without logical fallacies, respectively. After adjusting for the effect of individual error statements, Model 1 predicted that an average survey participant would identify a typical factual error with and without a logical fallacy 14.20% and 16.85% of the time. Very few survey participants identified but failed to correct factual errors (3 out of 33). Interview participants showed a higher rate of failed corrections upon identifying an FE w/o a LF (2 out of 4).

Among those who failed to identify the target errors, Model 1 predicted that about 5% of survey participants reported a response for reasons unrelated to the actual error (e.g., ambiguity, excessive information). Two uncommon cases stood out: one participant falsely reported an accurate response due to a misconception; one interviewee lost confidence in the chatbot after seeing it err in the first session and reported its responses in the second session despite not knowing what the errors were.

Participants' low accuracy of the practice problems triangulates their struggles in error detection. Model 3 predicted that an average participant had 25.16% and 29.80% probability of answering a practice problem correctly if they experienced a relevant factual error with and without a logical fallacy, respectively. By comparing the practice problem accuracy when participants did and did not encounter a relevant chatbot error, we found that a chatbot error negatively impacted learning outcomes with a large effect size (FE w/ LF: odds ratio mean = 0.18, 94% HPDI = [0.03, 0.40]; FE w/o LF: odds ratio mean = 0.18, 94% HPDI = [0.02, 0.38]) when participants had access to reading materials and a Google search engine but not direct feedback on their work.

Receiving direct feedback on their practice problem eliminated the negative effect of FE with LF and decreased the negative effect of FE without LF to a small effect size (mean = 0.65, 94% HPDI = [0.42, 0.87]). The remaining negative effect happened to participants who did not detect and report the chatbot error. Model 4 predicted that an average participant who did not detect and report an FE w/o a LF had a 54.11% chance of failing to achieve the learning goal (i.e., failure to correctly explain their multiple choice answer in the post-test), as opposed to the 40.43% chance for those who did not encounter a chatbot error.

Participants who successfully detected and reported chatbot errors did not get harmed by factual errors on the post-test. In fact, we observed a slightly better post-test performance for those who detected and reported a FE w/ a LF compared to those who did not experience any chatbot error (62.9% vs. 56.9% achieving learning goals) after controlling for various factors, including prior knowledge. But the effect was not significant (cumulative odds ratio mean = 1.92, 94% HPDI = [0.15, 4.96], ≥ 1 with 67.4% chance).

In addition, successful detection of chatbot errors eliminated chatbot errors' negative impact on self-efficacy. **Those who did not detect the chatbot error had smaller gains or greater drops in their self-efficacy** after the learning session than those who did not experience any chatbot error. The effect sizes ranged from small to large, but were insignificant due to uncertainty (FE w/o LF: Cohen's d mean = -0.34, 94% HPDI = [-0.98, 0.38], ≤ 0 with 82.1% chance; FE w/ LF: Cohen's d mean = -0.28, 94% HPDI = [-0.92, 0.39], ≤ 0 with 77.7% chance).

Chatbot's negative impact on self-efficacy may stem from misattribution of the poor performance on practice problem. We observed some interviewees who failed to identify chatbot errors attributed the responsibility for answering the practice problems wrong to themselves rather than the chatbot. For instance, one interviewee who adopted the chatbot's incorrect response commented: "*I thought I read the chatbot correctly, but I guess maybe I didn't. So I think it's just maybe me rushing things. It's probably more on my end than the chatbot.*"

7 Results - Part II: Effects of Chatbot Errors across Learners (RQ3)

We focus on how learner characteristics affect chatbot error detection and the impact on learning outcomes in this section. For the effects of ordinal predictors on the outcome variables, we report the effect of a one-standard-deviation change from the median in our predictors' data. The effect of a categorical predictor is the difference in outcomes between categories. We analyzed the two types of factual errors as a whole.

7.1 Prior Knowledge

To investigate the impact of prior knowledge, we modeled the interaction effects of chatbot errors with total pretest score, correctness of the relevant pretest problem (for problem-level models), confidence in the pretest answers, and STEM background. Overall, we found that participants with incorrect or less prior knowledge were significantly more likely to struggle with error detection.

When the chatbot made an error, participants who answered the corresponding pretest problem incorrectly were significantly more likely to get the practice problem wrong than those who had the right pretest answer. The effect size ranges from small to medium (odds ratio mean = 1.97, 94% HPDI = [1.01, 3.04]), equivalent to an average decrease in the chance of a correct practice answer from 37.87% to 25.53%. When the chatbot's responses were accurate, the average difference of practice problem accuracies between participants who answered the pretest problem correctly and wrongly was significantly smaller – around 7%, a drop from 66.92% to 60.18%. We found a similar trend for total pretest score. Participants with higher pretest scores had a higher chance of detecting and reporting chatbot errors with a small effect size, though not significant (odds ratio mean = 1.44, 94% HPDI = [0.67, 2.37], ≥ 1 with 84.6% chance). We identified two reasons from our interviews.

First, conflicts with prior knowledge played a key role in participants' chatbot evaluation strategies. Directly comparing with prior knowledge was the second most popular strategy (N = 8). In addition, 6 interviewees said they would further verify a chatbot response with alternative sources only when the response conflicted with prior knowledge. Those who came in with a misconception that aligned with the chatbot error were harmed as well. One survey participant that was in this case chose to agree with the chatbot even after noticing the conflict between the chatbot and the answer key on the practice problem. They dismissed the correct answer key as "absurd": "*I agree with the [chatbot] response: "The weaker the correlation, the less likely there is causation." This is obvious, and pretty self-evident. ... but the test said this was incorrect. This is absurd.*"

Second, lacking prior knowledge discouraged participants from using any chatbot evaluation strategy at all. Four interviewees stated they didn't try to assess the chatbot's responses because they felt they lacked the necessary knowledge and ability. One interviewee remarked: "*I am not very savvy in this field [Statistics]. That's just not something I am good at. So that's probably why I didn't try to give any suggestions.*"

7.2 Prior Experience & Perceptions with Chatbots

Among the four variables related to prior experience and perceptions with chatbots, we found that participants' prior chatbot usage frequency and perceived reliability of chatbots influenced error detection. Prior familiarity with how chatbots work and perceived helpfulness did not impact error detection.

Participants who used chatbots more often prior to the study detected chatbot errors and answered practice problems correctly with significantly higher probabilities despite receiving erroneous chatbot responses. For instance, Model 3 predicted that a participant who frequently vs. sometimes used chatbots on average had a 37.23% and 31.7% chance, respectively, of answering the practice problem correctly when given a wrong chatbot response, indicating a small effect size (odds ratio mean = 1.33, 94% HPDI = [0.99, 1.76]). Interestingly, when the chatbot was correct, prior usage frequency did not affect participants' accuracy on practice problems.

Prior perceived chatbot reliability had a significant negative effect on chatbot error detection with a small to medium effect size (odds ratio mean = 0.64, 94% HPDI = [0.33, 0.89]). Participants who considered chatbots as "often reliable" and "often unreliable" had an average error detection probability of 19.70% and 32.26%, respectively. But their accuracies on practice problems given chatbot errors did not differ significantly.

We identified two sources of unwarranted trust from our interview data. First, the chatbot's arguments often sounded convincing, regardless of accuracy. Three interviewees accepted chatbot responses that contradicted their initial answers, and in two cases, the chatbot was incorrect. One commented: "*There is a danger I see if the chatbot is very good with arguing. It can make a black color turn white, and I don't even know it.*" Second, two interviewees said their unwarranted trust came from the perception that the chatbot knew more than they did. One said: "*I am not too familiar on the topic myself. … So I trust the computer more.*"

7.3 Motivations

Participants' motivations had a limited impact on error detection and learning outcomes. Perceived task value had a close-to-significant but very small effect on error detection (odds ratio mean = 1.15, 94% HPDI = [0.93, 1.41]). Participants with higher self-efficacy did significantly better on the post-test regardless of whether they encountered or detected any chatbot error. But this positive effect weakened when participants encountered a chatbot error. Extrinsic and intrinsic motivations did not affect error detection or performance on practice and post-test problems.

7.4 Demographics

We investigated four demographic factors: gender, age, highest education level, and English proficiency. Before participants received answer keys and feedback on practice problems, non-native English speakers and female participants managed chatbot errors less effectively than their counterparts. Non-native English speakers reported detecting chatbot errors significantly less (13.05% mean probability vs. 39.72% for native speakers), but did not perform worse on practice problems. Male and female participants reported chatbot errors with similar probability, but females did worse on the practice problems in the presence of chatbot errors, having an average accuracy of 25.03%, as opposed to males' 34.54%.

On the post-test, participants with lower educational degrees were harmed more by undetected chatbot errors. For instance, those with high-school degrees performed significantly better than those with graduate degrees when chatbot was accurate with a small to medium effect size (odds ratio mean = 0.52, 94% HPDI = [0.25, 0.83]), but performed worse when they did not detect the chatbot errors with a small effect size (mean = 1.48, 94% HPDI = [0.83, 2.34]). Older participants answered problems with lower accuracy under all conditions, except when they detected chatbot errors, in which case they achieved similar performance.

8 Discussion and Future Work

Participants in our study faced challenges in detecting factual chatbot errors, which negatively impacted their learning outcomes and self-efficacy. In addition, participants with certain characteristics struggled more than others. This section explores the implications of these findings and provides recommendations for chatbot design and learner empowerment to improve learners' interactions with imperfect generative pedagogical chatbots.

8.1 Understanding the Impact of Chatbot Errors

Undetected chatbot errors had clear negative effects on both learning outcomes and self-efficacy, showing the importance of assisting learners with chatbot error detection. The harm to self-efficacy due to incorrect error attribution was concerning, especially in subjects where certain learner groups typically have lower self-efficacy, such as girls in STEM [74]. This harm may further exacerbate among learners who are more likely to attribute failure to internal factors rather than external factors, such as females compared to males [11].

Conversely, participants who detected errors showed a trend of performing better on post-tests than those who experienced no error in some cases (FEs with LFs), aligned with findings from previous studies [16, 34]. The learning gain may stem from resolving contradictions between the chatbot and other sources [47], suggesting that error detection itself could be a valuable learning exercise. Alternatively, learning strategies that helped with error detection may have improved the learning outcomes, regardless of the presence of chatbot errors. Future research is needed to determine whether error detection itself or the learning process that leads to error detection drives improved outcomes.

8.2 Designing Supporting Resources for Chatbot Error Management

Participants' success in detecting and recovering from chatbot errors varied by the available alternative learning resources. Despite knowing the chatbot might be imperfect and having access to reading materials and Google search, many participants failed to detect errors. This suggests that simply warning users about potential errors and providing unstructured resources, as proposed by prior work [48, 49], is insufficient. We observed potential explanations from our interviewees. Some did not attempt to use alternative resources to evaluate chatbot responses because they did not think they had the ability to do so. Others used evaluation strategies that depended on accurate prior knowledge. Designing scaffolds to guide the use of alternative resources as part of the interaction system may resolve this issue. For instance, a guided playground task or training on strategies to evaluate a chatbot response may boost learners' confidence and skills.

Even though many participants failed to detect chatbot errors at first, receiving immediate targeted feedback on the practice problems greatly reduced the learning harms. This shows that providing the right type of alternative learning support at the right time mitigates the harm of failed error detection. Future work can explore what kinds of learning activities and support help learners recover from misconceptions (e.g., explanations of common misconceptions, opportunities to consult human experts), and when and how to incorporate them in an ecologically valid learning environment with resource constraints.

8.3 Addressing Differential Impacts Across Learners

Our study revealed multiple learner groups that are more vulnerable to chatbot errors. First, beginners of a topic and learners who struggle with a topic (e.g., those who have many misconceptions) are one of the most vulnerable groups when using an imperfect chatbot. Our findings suggest that accurate prior knowledge plays a critical role in effective error management, similar to prior findings on expertise and AI-decision making [77]. We should caution learners against using an imperfect chatbot before they have studied the content from a reliable source. If instructors choose to introduce such a chatbot to beginners, they can structure the learning process to first build foundational knowledge through initial activities. Additionally, we should help learners develop alternative error-detection strategies, such as conditioning verification behaviors on confidence in their knowledge rather than on knowledge conflicts.

Second, participants with less experience using chatbots were less likely to detect the errors. This raises fairness concerns for deploying such systems among learners with different access to technology resources. Therefore, it is essential for instructors to introduce the system (e.g., how it works, how well it works) and provide guidance on how to use it. Instead of a one-size-fits-all deployment process, a scaffolded sandbox environment for personalized experimentation before formal use may allow learners with less experience to catch up.

Lastly, we found that non-native English speakers, females, and learners with lower educational levels encountered additional challenges with chatbot errors. Future work should dive deeper into the specific challenges these groups face and develop personalized solutions.

8.4 Empowering Learners to Manage Imperfect Chatbots

When learners seek help from a chatbot, especially those with little prior knowledge, they may perceive the chatbot as having significant expert power [23] over themselves. This power dynamic may explain why some participants had unwarranted trust in the chatbot, did not think they were in a position to evaluate the chatbot's responses, or were persuaded by the chatbot's wrong arguments. One solution is empowering learners to raise their status in this interactive relationship. Even though the skills to manage chatbot errors do not come naturally, they are malleable [73], similar to search literacy [58, 62].

As a first step, future research can study what evaluation strategies work. Based on our study, we believe these strategies are learner-dependent. For example, learners with little prior knowledge require strategies that do not depend on topic expertise. In addition, our findings suggest that strategies may be error-type-dependent. For instance, two of three interviewees who detected factual errors with logical fallacies first noticed the error via the logical reasoning structure. Therefore, trainings that increase logical reasoning proficiency may be more suited for FEs with LFs than FEs without LFs.

8.5 Limitations

Although our study offers valuable insights into how chatbot errors affect STEM learning, several factors limit the generalizability of our findings. First, we focused on introductory statistics and factual/conceptual knowledge. Learners at higher cognitive levels of Bloom's taxonomy (e.g., Bloom's "analyze" or "evaluate") or in more interpretive and subjective domains (e.g., social sciences, humanities) might encounter different error types and outcomes.

Moreover, our controlled online setting using Prolific participants may not fully capture classroom realities, where peer collaboration, instructor intervention, and varied resources could influence error detection and recovery. Future research should validate these findings in classroom settings, where the different contextual and motivational dynamics may influence learning and error management outcomes. In addition, we examined only two short learning sessions. User behavior, trust, and error-management strategies likely evolve with repeated system usage over time. Longitudinal studies could reveal more about how learners adapt to chatbot imperfections.

Finally, we studied two types of factual errors and observed differences across error types. While factual errors are common in generative models, they may not reflect the full spectrum of possible errors, such as those omitting information or involving ambiguous information. Future work can investigate and compare findings to understand what error types to prioritize addressing.

9 Conclusion

As the first comprehensive evaluation of the impact and user detection of factual errors in generative pedagogical chatbots within a non-programming STEM learning context, our study underscores the imminent need to better support learners in identifying and correcting chatbot errors. Providing this support will not only protect learners from the harms posed by undetected errors, but also create potential learning opportunities through successful error detection. Drawing from our findings, we have proposed strategies to support learners' chatbot error management via several avenues, including designing supporting resources, empowering learners, and providing customized support for vulnerable learner groups. We hope our work motivates and informs future research aimed at addressing the challenges of using imperfect generative pedagogical chatbots, moving us closer to realizing their full potential in transforming personalized education at scale.

Acknowledgments

We would like to thank the support of Microsoft Research's Accelerating Foundation Models Research program and the University of Illinois Center for Just Infrastructures. Tiffany Wenting Li was supported by the Google PhD Fellowship during this work.

References

[1] ESMA Aimeur, HUGO Dufort, DANIEL Leibu, and CLAUDE Frasson. 1997. Some justifications for the learning by disturbing strategy. In *Proceedings of the Eighth World Conference on Artificial Intelligence in Education*. Citeseer, 119–126.

[2] Esma Aimeur and Claude Frasson. 1996. Analyzing a new learning strategy according to different knowledge levels. *Computers & Education* 27, 2 (1996), 115–127.

[3] Murtaza Ali, Prerna Rao, Yifan Mai, and Benjamin Xie. 2024. Using Benchmarking Infrastructure to Evaluate LLM Performance on CS Concept Inventories: Challenges, Opportunities, and Critiques. In *Proceedings of the 2024 ACM Conference on International Computing Education Research - Volume 1* (Melbourne, VIC, Australia) *(ICER '24)*. Association for Computing Machinery, New York, NY, USA, 452–468. doi:10.1145/3632620.3671097

[4] Andrew Anderson, Tianyi Li, Mihaela Vorvoreanu, and Margaret Burnett. 2021. Human-AI Interaction for Diverse Humans: What Cognitive Style Disaggregation Reveals. *arXiv preprint arXiv:2108.00588* (2021).

[5] Patricia Armstrong. 2010. Bloom's taxonomy. *Vanderbilt University Center for Teaching* (2010), 1–3.

[6] Amos Azaria. 2022. ChatGPT usage and limitations. (2022).

[7] Rishabh Balse, Viraj Kumar, Prajish Prasad, and Jayakrishnan Madathil Warriem. 2023. Evaluating the Quality of LLM-Generated Explanations for Logical Errors in CS1 Student Programs. In *Proceedings of the 16th Annual ACM India Compute Conference* (Hyderabad, India) *(COMPUTE '23)*. Association for Computing Machinery, New York, NY, USA, 49–54. doi:10.1145/3627217.3627233

[8] Rishabh Balse, Bharath Valaboju, Shreya Singhal, Jayakrishnan Madathil Warriem, and Prajish Prasad. 2023. Investigating the potential of gpt-3 in providing feedback for programming assessments. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*. 292–298.

[9] Albert Bandura et al. 2006. Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents* 5, 1 (2006), 307–337.

[10] Amy L Baylor. 2011. The design of motivational agents and avatars. *Educational Technology Research and Development* 59 (2011), 291–300.

[11] Sylvia Beyer. 1998. Gender differences in causal attributions by college students of performance on course examinations. *Current psychology* 17 (1998), 346–358.

[12] Denys Brand, Matthew D Novak, Florence D DiGennaro Reed, and Samara A Tortolero. 2020. Examining the effects of feedback accuracy and timing on skill acquisition. *Journal of organizational behavior management* 40, 1-2 (2020), 3–18.

[13] Daniel H Chang, Michael Pin-Chuan Lin, Shiva Hajian, and Quincy Q Wang. 2023. Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability* 15, 17 (2023), 12921.

[14] Sam Cunningham-Nelson, Wageeh Boles, Luke Trouton, and Emily Margerison. 2019. A review of chatbots in education: practical steps forward. In *30th annual conference for the australasian association for engineering education (AAEE 2019): educators becoming agents of change: innovate, integrate, motivate*. Engineers Australia, 299–306.

[15] Mikolaj Czajkowski, Marek Giergiczny, and William H Greene. 2014. Learning and fatigue effects revisited: Investigating the effects of accounting for unobservable preference and scale heterogeneity. *Land Economics* 90, 2 (2014), 324–351.

[16] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[17] Benedict du Boulay, Antonija Mitrovic, and Kalina Yacef. 2023. *Handbook of Artificial Intelligence in Education*. Edward Elgar Publishing Limited, Cheltenham, UNITED KINGDOM. http://ebookcentral.proquest.com/lib/uiuc/detail.action?docID=7243036

[18] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" It, Then I Hide It: Folk Theories of Social Feeds. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2371–2382. doi:10.1145/2858036.2858494

[19] Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. An intelligent discussion-bot for answering student queries in threaded discussions. In *Proceedings of the 11th international conference on Intelligent user interfaces*. 171–177.

[20] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. *arXiv preprint arXiv:2402.00367* (2024).

[21] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30 (2020), 681–694.

[22] Jeremiah Folsom-Kovarik, Sae Schatz, and Denise Nicholson. [n. d.]. Plan Ahead: Pricing ITS Learner Models. ([n. d.]).

[23] John RP French. 1959. The bases of social power. *Studies in social power/University of Michigan Press* (1959).

[24] Andrew Gelman and Donald B Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical science* 7, 4 (1992), 457–472.

[25] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2014. Automation bias: empirical results assessing influencing factors. *International journal of medical informatics* 83, 5 (2014), 368–375.

[26] Michael H Graner. 1987. Revision workshops: An alternative to peer editing groups. *The English Journal* 76, 3 (1987), 40–45.

[27] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. 2023. LLM-as-a-tutor in EFL Writing Education: Focusing on Evaluation of Student-LLM Interaction. *arXiv preprint arXiv:2310.05191* (2023).

[28] Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. 2023. Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1 (ICER '23, Vol. 1)*. Association for Computing Machinery, New York, NY, USA, 93–105. doi:10.1145/3568813.3600139

[29] Jason M Hirst and Florence D DiGennaro Reed. 2015. An examination of the effects of feedback accuracy on academic task acquisition in analogue settings. *The Psychological Record* 65, 1 (2015), 49–65.

[30] Silas Hsu, Tiffany Wenting Li, Zhilin Zhang, Max Fowler, Craig Zilles, and Karrie Karahalios. 2021. Attitudes surrounding an imperfect AI autograder. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–15.

[31] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[32] Douglas A Johnson, Jessica M Rocheleau, and Rachael E Tilka. 2015. Considerations in feedback delivery: The role of accuracy and type of evaluation. *Journal of Organizational Behavior Management* 35, 3-4 (2015), 240–258.

[33] Young Ju Joo, Kyu Yon Lim, and Jiyeon Kim. 2013. Locus of control, self-efficacy, and task value as predictors of learning outcome in an online university context. *Computers & Education* 62 (2013), 149–158.

[34] Manu Kapur. 2015. Learning from productive failure. *Learning: Research and practice* 1, 1 (2015), 51–65.

[35] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4521–4532.

[36] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara Jane Ericson, David Weintrop, and Tovi Grossman. 2023. How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*. 1–12.

[37] Willett Kempton. 1986. Two theories of home heat control. *Cognitive Science* 10, 1 (1986), 75–90. doi:10.1207/s15516709cog1001_3

[38] Anjali Khurana, Hariharan Subramonyam, and Parmit K Chilana. 2024. Why and When LLM-Based Assistants Can Go Wrong: Investigating the Effectiveness of Prompt-Based Interactions for Software Help-Seeking. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 288–303. doi:10.1145/3640543.3645200

[39] Jihyun Kim, Kelly Merrill Jr, and Chad Collins. 2021. AI as a friend or assistant: The mediating role of perceived usefulness in social AI vs. functional AI. *Telematics and Informatics* 64 (2021), 101694.

[40] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 822–835.

[41] Yanghee Kim, Amy 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 (2006), 223–243.

[42] Christoph König and Rens van de Schoot. 2018. Bayesian statistics in educational research: a look at the current state of affairs. *Educational Review* 70, 4 (2018), 486–509.

[43] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies* 28, 1 (2023), 973–1018.

[44] Arun HS Kumar. 2023. Analysis of ChatGPT tool to assess the potential of its utility for academic writing in biomedical domain. *Biology, Engineering, Medicine and Science Reports* 9, 1 (2023), 24–30.

[45] Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N. Rafferty, John Stamper, and Michael Liut. 2024. Supporting Self-Reflection at Scale with Large Language Models: Insights from Randomized Field Experiments in Classrooms. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 86–97. doi:10.1145/3657604.3662042

[46] Judine Ladbrook and Elizabeth Probert. 2011. Information skills and critical literacy: Where are our digikids at with online searching and are their teachers helping? *Australasian Journal of Educational Technology* 27, 1 (2011).

[47] Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. 2024. One vs. many: Comprehending accurate information from multiple erroneous and inconsistent AI generations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2518–2531.

[48] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From ChatGPT to FactGPT: A Participatory Design Study to Mitigate the Effects of Large Language Model Hallucinations on Users *(MuC '23)*. Association for Computing Machinery, New York, NY, USA, 81–90. doi:10.1145/3603555.3603565

[49] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A Hallucination Identifier for Large Language Models. 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 482, 13 pages. doi:10.1145/3613904.3642428

[50] Haiying Li, Qinyu Cheng, Qiong Yu, and Arthur C Graesser. 2015. The role of peer agent's learning competency in trialogue-based reading intelligent systems. In *Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings 17*. Springer, 694–697.

[51] Tiffany Wenting Li, Silas Hsu, Max Fowler, Zhilin Zhang, Craig Zilles, and Karrie Karahalios. 2023. Am I wrong, or is the autograder wrong? Effects of AI grading mistakes on learning. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*. 159–176.

[52] Anna Lieb and Toshali Goel. 2024. Student Interaction with NewtBot: An LLM-as-tutor Chatbot for Secondary Physics Education. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 614, 8 pages. doi:10.1145/3613905.3647957

[53] Sook Lim. 2009. How and why do college students use Wikipedia? *Journal of the American Society for Information science and Technology* 60, 11 (2009), 2189–2202.

[54] Suqing Liu, Zezhu Yu, Feiran Huang, Yousef Bulbulia, Andreas Bergen, and Michael Liut. 2024. Can Small Language Models With Retrieval-Augmented Generation Replace Large Language Models When Learning Computer Science?. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1* (Milan, Italy) *(ITiCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 388–393. doi:10.1145/3649217.3653554

[55] Kristi Lundstrom and Wendy Baker. 2009. To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of second language writing* 18, 1 (2009), 30–43.

[56] Sara E McBride, Wendy A Rogers, and Arthur D Fisk. 2014. Understanding human management of automation errors. *Theoretical Issues in Ergonomics Science* 15, 6 (2014), 545–577.

[57] Richard McElreath. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan.* Chapman and Hall/CRC.

[58] Sarah McGrew. 2020. Learning to evaluate: An intervention in civic online reasoning. *Computers & Education* 145 (2020), 103711.

[59] Miriam J Metzger, Andrew J Flanagin, and Lara Zwarun. 2003. College student Web use, perceptions of information credibility, and verification behavior. *Computers & Education* 41, 3 (2003), 271–290.

[60] Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence* 6 (2024), 100199.

[61] Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. WorldBench: Quantifying Geographic Disparities in LLM Factual Recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1211–1228. doi:10.1145/3630106.3658967

[62] Renee Morrison and Georgina Barton. 2018. Search engine use as a literacy in the middle years: The need for explicit instruction and active learners. *Literacy Learning: The Middle Years* 26, 3 (2018), 37–47.

[63] Kathleen L Mosier and Linda J Skitka. 1999. Automation use and automation bias. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 43. SAGE Publications Sage CA: Los Angeles, CA, 344–348.

[64] Sandrine Neuville, Mariane Frenay, and Etienne Bourgeois. 2007. Task value, self-efficacy and goal orientations: Impact on self-regulated learning, choice and performance among university students. *Psychologica Belgica* 47, 1-2 (2007), 95–117.

[65] Zheng Ning, Zheng Zhang, Tianyi Sun, Yuan Tian, Tianyi Zhang, and Toby Jia-Jun Li. 2023. An empirical study of model errors and user error discovery and repair strategies in natural language database queries. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 633–649.

[66] Thomas Oakland and Holly B Lane. 2004. Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing* 4, 3 (2004), 239–252.

[67] Sandra Y Okita. 2014. Learning from the folly of others: Learning to self-correct by monitoring the reasoning of virtual characters in a computer-supported mathematics learning environment. *Computers & Education* 71 (2014), 257–278.

[68] Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. AutoTutor meets Large Language Models: A Language Model Tutor with Rich Pedagogy and Guardrails. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 5–15. doi:10.1145/3657604.3662041

[69] Sumit Pandey and Srishti Sharma. 2023. A comparative study of retrieval-based and generative-based chatbots using deep learning and machine learning. *Healthcare Analytics* 3 (2023), 100198.

[70] Hyanghee Park and Daehwan Ahn. 2024. The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications. 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 271, 21 pages. doi:10.1145/3613904.3642785

[71] Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K Gurbani. 2023. How susceptible are llms to logical fallacies? *arXiv preprint arXiv:2308.09853* (2023).

[72] PR Pintrich. 1991. A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). *National Center for Research to Improve Postsecondary Teaching and Learning* (1991).

[73] Snehal Prabhudesai, Ananya Prashant Kasi, Anmol Mansingh, Anindya Das Antar, Hua Shen, and Nikola Banovic. 2025. " Here the GPT made a choice, and every choice can be biased": How Students Critically Engage with LLMs through End-User Auditing Activity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[74] Ashley D Rittmayer and Margaret E Beier. 2009. Overview: Self-Efficacy in STEM. . *Self* (2009).

[75] Rod D Roscoe, Joshua Wilson, Adam C Johnson, and Christopher R Mayra. 2017. Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior* 70 (2017), 207–221.

[76] Soumyadeep Roy, Aparup Khatua, Fatemeh Ghoochani, Uwe Hadler, Wolfgang Nejdl, and Niloy Ganguly. 2024. Beyond Accuracy: Investigating Error Types in GPT-4 Responses to USMLE Questions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1073–1082. doi:10.1145/3626772.3657882

[77] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.

[78] Christian A Schiller. 2024. The Human Factor in Detecting Errors of Large Language Models: A Systematic Literature Review and Future Research Directions. *arXiv preprint arXiv:2403.09743* (2024).

[79] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. 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 1050, 17 pages. doi:10.1145/3613904.3642414

[80] Brad Sheese, Mark Liffiton, Jaromir Savelka, and Paul Denny. 2024. Patterns of Student Help-Seeking When Using a Large Language Model-Powered Programming Assistant. In *Proceedings of the 26th Australasian Computing Education Conference* (Sydney, NSW, Australia) *(ACE '24)*. Association for Computing Machinery, New York, NY, USA, 49–57. doi:10.1145/3636243.3636249

[81] Shruthi Shekar, Pat Pataranutaporn, Chethan Sarabu, Guillermo A. Cecchi, and Pattie Maes. 2024. People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. doi:10.48550/arXiv.2408.15266 arXiv:2408.15266.

[82] Amit Sheth and Krishnaprasad Thirunarayan. 2021. The Duality of Data and Knowledge Across the Three Waves of AI. *IT Professional* 23, 3 (May 2021), 35–45. doi:10.1109/MITP.2021.3070985 Conference Name: IT Professional.

[83] Abdulhadi Shoufan. 2023. Can Students without Prior Knowledge Use ChatGPT to Answer Test Questions? An Empirical Study. *ACM Trans. Comput. Educ.* 23, 4, Article 45 (dec 2023), 29 pages. doi:10.1145/3628162

[84] Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744* (2023).

[85] Bruce W Tuckman and Gary J Kennedy. 2011. Teaching learning strategies to increase success of first-term college students. *The Journal of Experimental Education* 79, 4 (2011), 478–504.

[86] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. 2022. Trust in human-AI interaction: Scoping out models, measures, and methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[87] Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*. 355–368.

[88] Verena Wagner-Hartl, Ramona Schmid, and Katharina Gleichauf. 2022. The influence of task complexity on acceptance and trust in human-robot interaction–gender and age differences. *Cognitive computing and internet of things* 43 (2022), 118–126.

[89] Amber Walraven, Saskia Brand-Gruwel, and Henny PA Boshuizen. 2009. How students evaluate information and sources when searching the World Wide Web for information. *Computers & education* 52, 1 (2009), 234–246.

[90] Thiemo Wambsganß, Leonie Haas, and Matthias Söllner. 2021. Towards the Design of a Student-Centered Question-Answering System in Educational Settings.. In *ECIS*.

[91] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[92] Jennifer Wiley, Susan R Goldman, Arthur C Graesser, Christopher A Sanchez, Ivan K Ash, and Joshua A Hemmerich. 2009. Source evaluation, comprehension, and learning in Internet science inquiry tasks. *American Educational Research Journal* 46, 4 (2009), 1060–1106.

[93] Rainer Winkler and Matthias Söllner. 2018. Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of Management Proceedings*, Vol. 2018. Academy of Management Briarcliff Manor, NY 10510, 15903.

[94] Defne Yigci, Merve Eryilmaz, Ail K Yetisen, Savas Tasoglu, and Aydogan Ozcan. 2024. Large Language Model-Based Chatbots in Higher Education. *Advanced Intelligent Systems* (2024), 2400429.

[95] Wenting Zhang, Qiaorong Zhang, Mingming Cai, Dongqing Wang, and Yafeng Zheng. 2024. Navigating the Application Challenges of ChatGPT in Education: Promoting Responsible Use and Minimizing Mental Risks. In *Proceedings of the 2024 9th International Conference on Distance Education and Learning* (Guangzhou, China) *(ICDEL '24)*. Association for Computing Machinery, New York, NY, USA, 23–28. doi:10.1145/3675812.3675843