Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

Do Students Rely on AI? Analysis of Student-ChatGPT Conversations from a Field Study

Jiayu Zheng[1], Lingxin Hao[1], Kelun Lu[2], Ashi Garg[1], Mike Reese[1], Melo-Jean Yap[1], I-Jeng Wang[3], Xingyun Wu[1], Wenrui Huang[1], Jenna Hoffman[1], Ariane Kelly[1], My Le[1], Ryan Zhang[1], Yanyu Lin[1], Muhammad Faayez[1], Anqi Liu[1]

[1] Johns Hopkins University, Baltimore, MD, USA,
[2] University of Pennsylvania, Philadelphia, PA, USA,
[3] Johns Hopkins University Applied Physics Lab, Laurel, MD, USA
jzheng59@jh.edu, hao@jhu.edu, aliu74@jhu.edu

Abstract

This study explores how college students interact with generative AI (`ChatGPT-4`) during educational quizzes, focusing on reliance and predictors of AI adoption. Conducted at the early stages of `ChatGPT` implementation, when students had limited familiarity with the tool, this field study analyzed 315 student-AI conversations during a brief, quiz-based scenario across various STEM courses. A novel four-stage reliance taxonomy was introduced to capture students' reliance patterns, distinguishing AI competence, relevance, adoption, and students' final answer correctness. Three findings emerged. First, students exhibited overall low reliance on AI and many of them could not effectively use AI for learning. Second, negative reliance patterns often persisted across interactions, highlighting students' difficulty in effectively shifting strategies after unsuccessful initial experiences. Third, certain behavioral metrics strongly predicted AI reliance, highlighting potential behavioral mechanisms to explain AI adoption. The study's findings underline critical implications for ethical AI integration in education and the broader field. It emphasizes the need for enhanced onboarding processes to improve student's familiarity and effective use of AI tools. Furthermore, AI interfaces should be designed with reliance-calibration mechanisms to enhance appropriate reliance. Ultimately, this research advances understanding of AI reliance dynamics, providing foundational insights for ethically sound and cognitively enriching AI practices.

Introduction

The rapid advancement and pervasive integration of artificial intelligence (AI) systems are profoundly reshaping diverse domains, with education emerging as a significant frontier (Costa et al. 2024; Wu, Dang, and Li 2025; Hasanein and Sobaih 2023; Fırat 2025). AI holds immense potential to revolutionize educational practices by offering tailored guidance, feedback, and support through personalized learning platforms and intelligent tutoring systems (Wang et al. 2024; Paladines and Ramirez 2020; Galdames 2024). These technologies promise to enhance student engagement, improve academic outcomes, and adapt instruction to individual learning patterns (Wang et al. 2024). However, the increasing integration of AI also necessitates a critical examination of how humans interact with and rely on these systems, which will inform ethical deployment of AI in education applications.

One critical concern is that students might over-rely on AI, which influences critical thinking (Gerlich 2025; Zhai, Wibowo, and Li 2024), human decision-making, analytical thinking (Grassini 2023) and increases human laziness (Ahmad et al. 2023). Previous studies have extensively explored specific aspects of AI reliance, typically within lab settings or field experiments (Dai et al. 2025). However, some areas of AI reliance are underexplored. First, the understanding of spontaneous human-AI interactions, how students naturally utilize AI tools in authentic academic contexts (Eckhardt et al. 2024) and the verification of real-world data, particularly concerning the underlying behavioral mechanisms that explain different reliance scenarios remains limited (Dai et al. 2025). Furthermore, existing research predominantly takes a static view to lable messages or conversations without capturing the evolving nature of the student-AI conversation dynamically (Solomon et al. 2022). Lastly, most reliance definitions in the human–AI interaction and AI-assisted decision-making literature rely on initial human judgments and assume that they can be labeled simply as correct or incorrect. This binary framing does not translate well to educational settings, where it is hard to evaluate students' initial status and students are actively acquiring new knowledge with AI assistance.

This study addresses these gaps by conducting an examination of students' reliance on AI within an authentic university-level STEM quiz environment using a novel four-stage reliance taxonomy in 315 conversations. Through comprehensive behavioral and text analyses, we identify patterns of student reliance and elucidate underlying factors influencing reliance behaviors. In a college peer-led-team-learning (PLTL) program (Bauer et al. 2025), students place greater trust in the instructors, teaching assistants, and knowledgeable peers than in GenAI given the inherent property of uncertainty in AI responses and the unfamiliarity of college students with `ChatGPT` when it was first released. Our field study took advantage of this real-world setting by deploying an interface with `ChatGPT-4` and a course-specific quiz application for 30 minutes during a PLTL session of STEM gateway courses[1]. Of the consented students, 182 students decided to use the interface for solving at least one quiz problem. The study was approved by the university IRB. The results will shed light on an understanding of the trust in `ChatGPT` when PLTL program is available. In such a context, we believe that the degree to which students' engagement with `ChatGPT` and the ways by which students become critically reliant on `ChatGPT` will reveal contributing factors of reliance and the underlying ethical issues.

Our study makes several detailed contributions. Firstly, we develop a reliance taxonomy that extends beyond traditional reliance metrics by incorporating measures of AI correctness, AI relevance, students' adoption of AI advice and the correctness of final answer. Secondly, we provide both conversation-level and individual-level analyses, thereby capturing reliance dynamics over time and highlighting distinct reliance trajectories that students exhibit across multiple interactions. Finally, we utilize behavioral and text features to identify predictors of different types of AI adoption, uncovering crucial mechanisms that govern students' AI adoption patterns and offering insights for enhancing human-AI interaction design.

Related Work

Human-AI Interaction in Education

AI has been integrated into various areas of education, offering diverse applications that influence students' academic development. These applications include intelligent tutoring systems, educational robots, learning analytics dashboards, and adaptive learning platforms (Galdames 2024; Wang et al. 2024). Generative AI (GenAI) tools, such as `ChatGPT`, are increasingly utilized in higher education, providing personalized learning and assessment opportunities (Wu, Dang, and Li 2025). They have the potential to adapt instruction to different student types, offer feedback, assist in course design, and support academic writing (Santos, Urgel, and Moreno 2024).

The influence of AI on students, however, presents inconsistent findings in the literature. Some studies highlight the benefits, suggesting that AI can enhance learning outcomes, improve knowledge accessibility, promote personalized support, and increase student motivation and engagement (Ding et al. 2023; Ouyang and Jiao 2021; Singh, Tayarani-Najaran, and Yaqoob 2023). For instance, AI-powered platforms have been shown to enhance student engagement and performance through real-time feedback and customized instructions (Hakiki et al. 2023). Conversely, other research points to potential drawbacks. Concerns include the risk of overreliance on AI, which may hinder the development of critical thinking and independent problem-solving skills (Kosar et al. 2024; Krupp et al. 2024). This often manifests by students reducing their mental effort and relying on AI for quick solutions, bypassing deeper cognitive processes essential for learning (Fan et al. 2025). Ethical issues are also prominent, particularly concerning academic dishonesty and the uncritical acceptance of AI-generated content (Acosta-Enriquez et al. 2024; Tan 2023). The quality of AI-generated content can be inconsistent, inaccurate, or lack depth, depending on user prompts, potentially leading to misunderstandings or the absorption of misleading information (Wang et al. 2023).

A significant gap in current research is that many studies employ experimental designs to investigate how specific factors influence human-AI interaction in controlled educational settings (Dai et al. 2025). While valuable, this approach often fails to capture how humans use "natural AI in the wild"—that is, how students spontaneously interact with widely available AI tools like `ChatGPT` in their everyday learning processes. Understanding these natural usage patterns from the user's perspective is crucial for inferring how AI systems are truly adapted and integrated into learning behaviors, providing insights beyond what can be observed in structured experiments.

Behavioral analysis of human-AI interaction in education Understanding human-AI interactions necessitates a granular analysis of user behavior while little attention has been paid to this matter. In educational settings, behavioral analysis provides a powerful lens for revealing patterns and trends in learning behaviors, which can inform the design and improvement of AI-supported educational platforms (Dai et al. 2025).

Scholars have used several behavioral factors to understand student-AI interactions. Interaction patterns and sequences are meticulously studied, revealing how students engage with AI systems through specific behavioral flows, and the routines of human behaviors (Sun, Zhao, and Chen 2024), such as querying background information, periods of idle operation, reviewing project requirements, and interpreting AI-generated content (Dai et al. 2025). A study also categorizes students' usage of AI by the combination of "mode of interaction" and "desired outcome" as well as using Bloom's Taxonomy to classify cognitive processes from simpler to more complex (Handa et al. 2025) . As suggested by their findings, students delegate higher-order cognitive tasks to AI systems. How do we ensure students still have the capabilities to discern the correctness of AI's answers and do not over-rely on AI?

Despite these advancements, a significant gap in behavioral analysis of student-AI interaction lies in the focus on message-level analysis rather than considering the conversation as a whole unit. Many current research efforts tend to analyze individual turns or isolated messages, potentially overlooking the holistic, evolving nature of student-AI interaction over a sustained period. Some research captures behaviors and conversations in human-AI interaction (McNichols and Lan 2025; Dai et al. 2025; Ammari et al. 2025). However, capturing the full context and flow of a conversation, including how students iteratively refine their questions or respond to prior AI turns, is crucial for a comprehensive understanding of reliance dynamics. This necessitates approaches that can analyze the entire student-AI conversation as a cohesive unit, rather than just discrete messages.

[1]More details about the PLTL program can be found at https://academicsupport.jhu.edu/pilot/.

Reliance Analysis in Human-AI Interaction

Reliance in the literature In the literature, reliance is defined as the observable behavior of a user following or utilizing AI advice (Eckhardt et al. 2024). However, relying on AI's wrong advice would lead to incorrect decisions. Therefore, scholars also introduce "appropriateness" in reliance analysis (Schemmer et al. 2023).

Various metrics have been employed to quantify AI reliance, often depending on the decision-making context. The "agreement percentage" quantifies how often a user's decision matches the AI's advice. The "switch percentage" measures how often users change their initial decision to align with AI advice, typically in a two-stage decision process where an initial human decision is recorded before AI advice is presented. The "Weight of Advice (WOA)" quantifies the impact of AI recommendations on a user's final decision, particularly in non-discrete scenarios (Eckhardt et al. 2024). Appropriateness also measures the percentage in which the decision-maker relies on correct AI advice and does not rely on incorrect AI advice.

To further distinguish whether the reliance on correct AI advice stems from a correct discrimination or simply an overlap of the human and the AI's decision, judge-advisor system paradigm includes an initial human decision, correctness of AI's advice and whether human follow is adopted (Schemmer et al. 2023; Eckhardt et al. 2024; Cao, Liu, and Huang 2024). However, in the real-world educational scenarios, particularly when students are learning, they may not have an initial decision or a pre-existing answer to a problem. Instead, they might consult AI as a primary source of information or a tool for exploration, making the concept of an "initial decision" irrelevant. This highlights the need for a reliance framework that does not rely on unobservable initial human decisions, allowing for a more accurate assessment of reliance in dynamic learning contexts.

To achieve the potential of human-AI collaboration with a better performance than either the human or the AI alone, researchers advocate the need the human to exercise discretion in following AI's advice, i.e., appropriately relying on the AI's advice (Schemmer et al. 2023). Recognizing the importance of building a mental model of the AI to assess AI recommendations, i.e., humans having an accurate understanding of what the AI system can and cannot do, researchers hypothesize and demonstrate that human learning is a key mediator of appropriate reliance in an experiment with 100 participants (Schemmer et al. 2023).

What factors influence reliance? Human's reliance on AI systems is a multifaceted phenomenon influenced by a complex interplay of system-related, user-related, and task-related factors (Eckhardt et al. 2024).

The intrinsic characteristics of the AI system itself are paramount in shaping user reliance, such as AI's accuracy (Lai and Tan 2019; Yin, Wortman Vaughan, and Wallach 2019; Kahr et al. 2024), explainability (Bansal et al. 2021; Kahr et al. 2024), alignment with human's values (Narayanan et al. 2023), and reliability of AI (Zhai, Wibowo, and Li 2024). Ethical issues of AI, including hallucination, algorithmic bias, overconfidence, plagiarism, privacy concerns and tranparency concerns influence people's trust in and adoption of AI. Besides intrinsic problems with the system, how human interacts with AI also affects human-AI reliance. For example, "helpfulness" is an important aspect we need to take into consideration in the reliance framework, as AI might not understand users' requests when users do not provide accurate prompts to reflect their needs, which hinders human from appropriately relying on AI.

Intrinsic individual user characteristics will influence reliance. A higher general inclination to trust technology can lead to greater reliance on correct AI advice (Li, Lu, and Yin 2023; Kahr et al. 2024; Küper et al. 2025). Users with higher domain expertise tend to exhibit higher openness to technology yet lower levels of trust in and reliance on AI systems (Küper et al. 2025). A user's confidence in their initial decision and prior experience can shape their susceptibility to AI recommendations, with high confidence correlating with greater self-reliance, particularly when the initial decision is correct (Schemmer et al. 2023; Zhou et al. 2024; Cao, Liu, and Huang 2024).

The nature of the task also influences reliance. User reliance can vary based on the task type (e.g., objective vs. subjective) and its inherent complexity (Schaschek, Spatscheck, and Winkelmann 2024).

Reliance in education applications Student's reliance on AI in education is a growing area of concern, with mixed findings regarding its implications. Although utilizing `ChatGPT` has the advantage of increasing efficiency by saving time and effort, users could get into the habit of adopting the answers without rationalization or verification. Over-reliance on GenAI also contributes to negative outcomes such as influencing critical thinking (Gerlich 2025; Zhai, Wibowo, and Li 2024), human decision-making, analytical thinking (Grassini 2023) and increasing human laziness (Ahmad et al. 2023). This evidence underscores the complexity of integrating such powerful tools into learning environments, especially when students may simply copy and paste AI's answers without critical evaluation, undermining academic integrity and hindering their own learning.

However, there is a notable lack of research specifically on human's reliance on AI in education with verification of real-world data, particularly concerning the underlying mechanisms that explain different reliance scenarios (Dai et al. 2025). While some studies analyze reliance by examining how students use AI for specific tasks or how their performance changes with AI assistance, there is a significant gap in using behavioral analysis to explain why different reliance scenarios occur. Existing research often describes reliance outcomes but lacks a deeper investigation into the interactive mechanisms and features that drive these behaviors. This absence of behavioral analysis limits our understanding of the complex interplay between student actions, AI responses, and the resulting reliance patterns in educational settings.

This Study

The preceding review highlights several research gaps in the understanding of human-AI reliance, particularly within educational contexts. First, there is a lack of research on the user's perspective of human-AI interaction when users engage with "natural AI in the wild", rather than in controlled experimental settings (Eckhardt et al. 2024). This limits our understanding of how students spontaneously integrate and adapt to AI tools in their authentic learning environments. Second, despite the growing sophistication of behavioral analysis, there is a continuing need for an approach that can capture the dynamic patterns that predict reliance, especially in complex, open-ended conversational settings. Specifically, much of the current research focuses on message-level analysis, rather than considering the conversation as a whole unit, which limits the ability to understand the holistic, evolving nature of student-AI interaction over time (Solomon et al. 2022). Third, there is a general lack of research on human reliance on AI in education, especially concerning the underlying mechanisms that explain different reliance scenarios. Current reliance frameworks often rely on the presence of an initial human decision, which may not be applicable in learning contexts where students are actively acquiring new knowledge with AI assistance.

Our paper attempts to fill this substantial gap using the analysis of 315 factual student-AI conversations from a field study. In this analysis of students' dialogues with `ChatGPT`, we address the ethical quandaries posed by GenAI within an authentic higher education setting. Our paper attempts to provide a nuanced understanding of how students are engaged in exchanging with `ChatGPT` for the learning task around a quiz problem. This design advances our understanding of whether and how `ChatGPT` influences students' cognitive skills in solving problems through Our ultimate goal is that the use of GenAI to assist college students' learning is ethically sound and cognitively enriching. This research seeks to answer the following questions.

1. How do students rely on AI when using AI for a quiz?
2. What factors are related to students' adoption of AI's advice?

Data

We conducted a field study within the PLTL program at a university in weeks 9-11, Fall 2023. The study leveraged a web-based quiz application that integrated `ChatGPT-4` as an AI learning peer. In the first eight weeks, students had regular PLTL interactions with human learning peers. During the first 30 mininutes of the PLTL sessions in weeks 9-11, human learning peers were replaced with the AI learning peer. Among the consented students, 216 students were assigned to use AI but only 182 of them interacted with `ChatGPT-4` for at least one quiz problem. The time-stampted data include quiz questions and students' answers, and logs of student–AI conversations. The unit of analysis is the conversation, defined as the interaction between a student and `ChatGPT-4` for a quiz problem. Focusing on five subject areas (mathematics, physics, chemistry, economics, and public health), the analytic data comprises 315 student–AI conversations from 182 students with valid textual contents. The quiz answers were graded by the teaching assistants according to the grading instructions provided by the instructors.

Measurements

Reliance Taxonomy

We develop a four-stage reliance taxonomy to capture students' interaction with `ChatGPT` during quiz tasks. The four stages are: (1) `ChatGPT's` competence (factual correctness), (2) `ChatGPT's` relevance (whether it addresses the student's question), (3) student's adoption of `ChatGPT's` advice, and (4) correctness of the student's final answer. First, we determine whether `ChatGPT's` response is factually correct. Next, we assess whether a correct response actually addresses the student's question (an incorrect response cannot be relevant by definition). Third, we code whether the student follows `ChatGPT's` suggested solution path. Finally, we evaluate whether the student's submitted answer is correct.

Concatenating these four binary labels produces a four-digit *reliance code* (e.g. 1–1–1–1 indicates that `ChatGPT's` answer is correct (1), and relevant (1), the student followed (1), and the final answer is correct (1); 0-0-0-0 indicates that `ChatGPT's` answer is incorrect (0), and irrelevant (0), the student did not follow (0), and the final answer is incorrect (0)), which we then map to one of twelve *reliance scenarios* (e.g. "Appropriate reliance," "Failed application," "Independent success," "Inappropriate self-reliance," "Semi-dependent," "Self-corrected success," "Unguided failure," "Serendipitous success," and "Total failure"). We define "Appropriate reliance" as students leveraging AI to obtain help that supports arriving at a correct answer but not considering cheating i.e., copy and paste quiz question text for quick answer. To verify this, we compute similarity scores between each student's prompt and the corresponding quiz question text. The observed average similarity is low, suggesting that students typically paraphrased the question or asked their own variants instead of submitting the original text verbatim, which reduces concern about simple shortcutting or cheating. Table 1 provides a full description of each category. We also visualize the taxonomy for the classification in Figure 1.

Behavioral and Text Features Extraction

From the student-GPT conversations, we generated behavioral features and text features.

Behavioral features describe how students interact with the chatbot. Specifically, interaction time in seconds (mean: 180.26, standard deviation (SD): 278.55) captures the total duration from the user's initial prompt to the bot's final response, while answer time in seconds (mean: 122.55, SD: 274.67) measures the interval from the bot's last message to the submission of an answer. Generally, students spent more time interacting with AI than answering the quiz problems. The average interaction time is around three minutes.

Two similarity metrics are used to measure how students make use of AI to answer the quiz problems. First, user-prompt-quiz-question similarity (mean: 0.50, SD: 0.23)

GPT correct?	GPT relevant?	Student follow?	Final ans. correct?	Reliance code	Reliance label	Interpretation
1	1	1	1	1-1-1-1	Appropriate reliance	GPT gave a correct, relevant answer, student followed it, and got the right result.
1	1	1	0	1-1-1-0	Failed application	GPT guidance was correct and relevant, student followed, but the final outcome was wrong.
1	1	0	1	1-1-0-1	Independent success	GPT was correct and relevant, student did not use it but solved the problem.
1	1	0	0	1-1-0-0	Inappropriate self-reliance	GPT was accurate and relevant, yet student ignored it and failed.
1	0	1	1	1-0-1-1	Semi-dependent	GPT answer was correct but off-topic; student followed nonetheless got it right.
1	0	1	0	1-0-1-0	Inappropriate reliance	GPT was correct but irrelevant, student followed it and failed.
1	0	0	1	1-0-0-1	Self-corrected success	GPT was correct but irrelevant; student ignored it and succeeded.
1	0	0	0	1-0-0-0	Unguided failure	GPT was correct yet irrelevant; student ignored it and failed.
0	0	1	1	0-0-1-1	Serendipitous success	GPT gave a wrong and off-topic answer; student followed it but arrived at the correct result.
0	0	1	0	0-0-1-0	Inappropriate reliance	GPT was wrong and irrelevant; student followed it and failed.
0	0	0	1	0-0-0-1	Self-corrected success	GPT was wrong and irrelevant; student ignored it and succeeded.
0	0	0	0	0-0-0-0	Total failure	GPT was wrong and irrelevant; student ignored it and failed.

Table 1: Reliance Taxonomy of Human–AI Reliance Scenarios

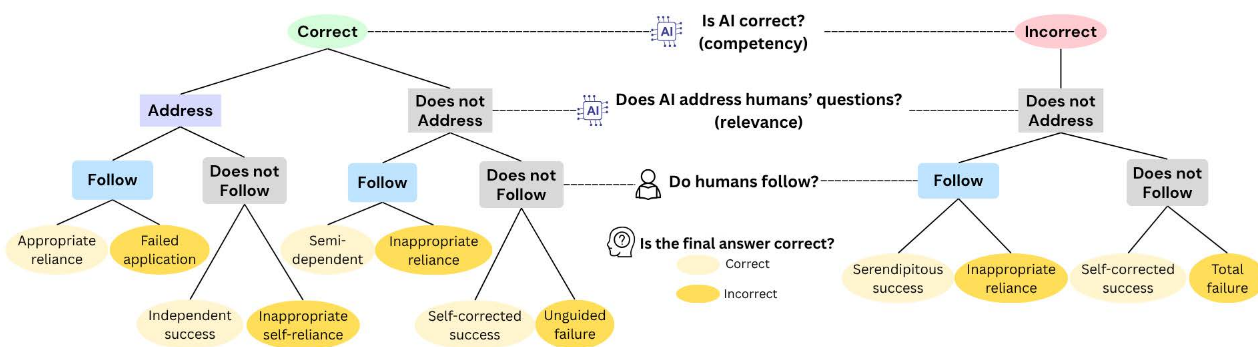

Figure 1: Visualizing Reliance Taxonomy. Based on our identification of the 4 labels as the *reliance code*, we classify student-AI conversation into 12 scenarios. If the answer given by AI is factually wrong at the first place, the answer is irrelevant to student's question by definition. Link to the figure: https://www.canva.com/design/DAGn8nxHMCg/O0aiU3cLlVlY15fD8HcKpA/edit.

assesses the extent to which users copy the quiz question text or rephrase it. Second, user-prompt-bot-prior-response similarity (mean: 0.17, SD: 0.22) evaluates to what extent students' prompts are related to the bot's previous response. Both similarity metrics are measured by `all-distilroberta-v1`, which is intented to be used as a sentence and short paragraph encoder, fitting with our quiz context. In general, students did not simply copy and paste the text of the quiz questions to get a quick answer. Instead, they paraphrased the questions or asked their own questions. The low user-prompt-bot-prior-response similarity suggests that students' prompts are not entirely bonded with AI's response. The distributions of similarity scores of both users' and bot's messages are shown in Figure 2.

Additionally, the linguistic complexity of both users' prompts (mean: 6.90, SD: 3.61) and bot's response (mean: 9.90, SD: 3.51) is measured via the Flesch–Kincaid grade level (Kincaid et al. 1975), is a readability test that estimates the educational level required to understand a text. It assigns a score based on a U.S. school grade level, indicating the difficulty of the text. A score of 9.0 means an ninth-grade level student can understand the text. In general, students ask simpler questions while AI respond with more complex answers. The distributions of complexity scores of both users' and bot's messages are shown in Figure 3.

Apart from what users do during the interaction, this study also intended to extract the content of interaction. Our message-level labeling, guided by the revised Bloom's Taxonomy (Krathwohl 2002), covers both the knowledge dimension (types of questions) and the cognitive process dimension (original Bloom's Taxonomy). We classified students' prompts by types of questions into conceptual (N = 103), procedural (N = 131), calculation (N = 36), reasoning (N = 27), and other (N = 15) categories through a

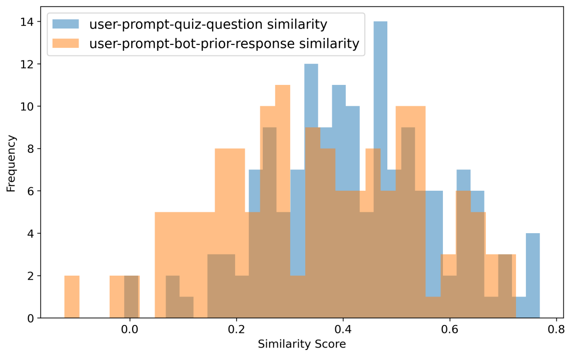

Figure 2: Distribution of Similarity Scores

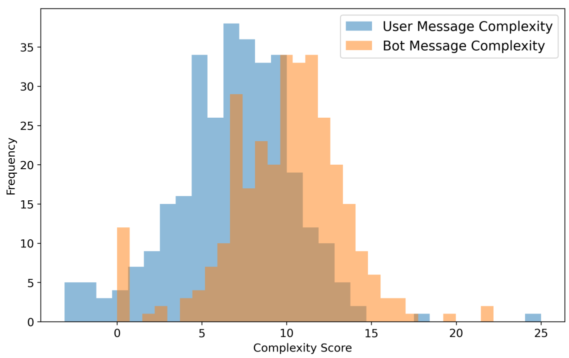

Figure 3: Distribution of Complexity Scores

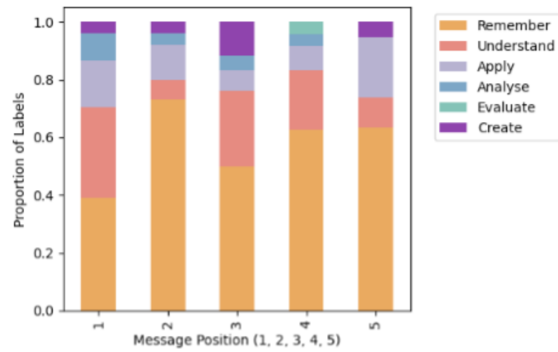

Figure 4: Remember-driven Cluster

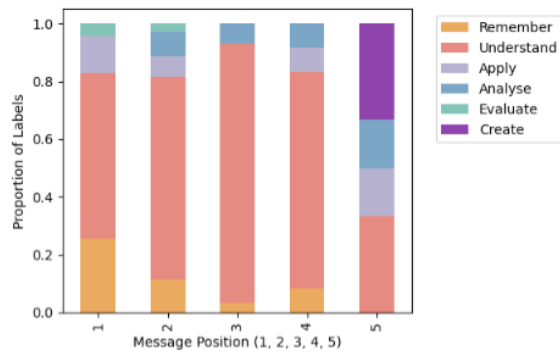

Figure 5: Understand-driven Cluster

few-shot training approach with `GPT-4o`. Conceptual questions refer to the prompts that students ask for the definition and meaning of a concept. The procedural questions refer to the prompts that students ask for one step or multiple steps to solve a problem. Calculation questions refer to the prompts that students ask `ChatGPT` to solve calculate or solve an equation. Reasoning questions refer to the prompts that students ask `ChatGPT` about the logic behind. Other questions include example request, visualization request and other questions types that have not frequently shown up in data. Prompts are also categorized by Bloom's Taxonomy, which is a hierarchical framework that categorizes learning objectives into six levels based on cognitive skills. We label message by BloomBERT (Lau 2023) with remember (N = 237), understand (N = 268), apply (N = 119), analyze (N = 31), evaluate (N = 9), and create (N = 21).

To move beyond isolated, message-level tags and capture the full arc of student-AI exchanges, we represent each conversation as a sequence of coded question types and Bloom's Taxonomy, and compute pairwise dissimilarities via an edit-distance metric. Specifically, we assign a cost of one to insertions and deletions and a cost of two to substitutions, so that replacing one question type with another is penalized more heavily than skipping or repeating a turn. Calculating the minimum total cost to transform sequence S_i into S_j for all pairs produces an $n \times n$ symmetric matrix reflecting how similar or different each conversation is from every other. We then apply agglomerative hierarchical clustering with Ward's linkage to these distance matrices (Solomon et al. 2022). When clustering the question-type sequences and cutting the resulting dendrogram into four groups, we uncover distinct prototypical patterns: conceptual-driven (N = 48), procedural-driven (N = 52), calculation-driven (N = 52), and the balanced cluster (N = 35). Applying the same pipeline to Bloom-coded sequences yields three clusters: remember-driven (N = 74), understand-driven (N = 70), and apply-driven (N = 20). We create three stacked-bar charts (Figure 4, Figure 5, and Figure 6) showing, for each of the first five user messages, the proportion of Bloom's taxonomy labels within the remember-, understand-, and apply-driven clusters. They reveal that remember questions dominate throughout in the first chart, understand questions dominate in the second, and apply questions overwhelmingly characterize the third. This sequential clustering approach makes several key contributions. First, by preserving the order and transition costs of question types, it captures temporal patterns of inquiry that single-shot message labels cannot. Second, by distilling unique interaction logs into a handful of conversation archetypes.

Text features are extracted from the embedding model available on Hugging Face under `dunzhang/stella_en_1.5B_v5` (Zhang et al. 2025) is used to extract 1,042-dimension embeddings of the whole student-AI conversation. This model was chosen because it exhibited the best performance on the MTEB challenge while being small enough to embed on a single NVIDIA

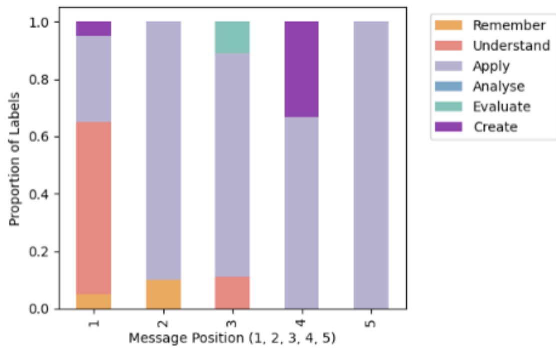

Figure 6: Apply-driven Cluster

A100 GPU (Muennighoff et al. 2022). Principal component analysis (PCA) is conducted to reduce the dimensionality of embeddings to eight components, which explain 43 % of the total variance.

Methods

To investigate how do students rely on AI and explore what factors influence the AI adoption patterns, we first construct a four-stage, conversation-level reliance taxonomy, which classifies conversations to twelve reliance scenarios. Next, we derive individual reliance trajectories and predict whether students follow AI's advice on the conversation data using both behavioral metrics and text embeddings across different machine learning models. Finally, we apply Shapley Additive Explanations to our XGBoost model to transparently quantify the contribution of each feature, trying to interpret how different features predict student's AI adoption.

Constructing Reliance Taxonomy

Initially, we treat each student answering one quiz problem as a case and categorize them to 12 reliance types in conversation level. To construct the reliance taxonomy, we ask `GPT-4o` via API to label data at three different stages respectively, including (1) `ChatGPT's` answer is correct or not, (2) `ChatGPT's` answer is relevant to student's question or not, and (3) do students follow `ChatGPT's` advice.

The labeling pipeline uses a a human-in-the-loop, few-shot framework. First, we have human experts manually annotate 30 seed cases, assigning binary labels for each of the three stages (correctness: 0/1, relevance: 0/1, student follow: 0/1). From these, the expert annotators select 3-5 representative conversation-label exemplars to prime `GPT-4o`. Next, the prompts further instruct `GPT-4o` to: (1) produce an initial label for a given stage with explicit chain-of-thought reasoning, (2) self-review that initial judgment in light of its own justification, (3) revise if needed, and (4) emit a final succinct label accompanied by a one-sentence rationale. We validate `GPT-4o's` performance by comparing its labels on the remaining annotated cases (held-out from the seeding exemplars) against the human gold standard. The alignment rate of human labels and `GPT-4o's` labels exceeds 85%. After this verification step, the same prompt-and-self-correction procedure is applied to label the rest of the dataset.

At the first stage, we call `GPT-4o` via API to identify whether `ChatGPT's` responses in the conversation are factually correct or not. If `ChatGPT` gives factually wrong answers or gives multiple self-contradictory answers, label it as 0 (wrong). Otherwise, label it as 1 (correct). For instance, there is a case that `ChatGPT` gave an answer *"The answer to (1536 - 16384)/24 is -617.00. The answer to (1536 - 16384)/24 is -618.67."*, which is inconsistent and labeled as wrong. There are 17.8% cases that `ChatGPT` gives factually wrong answer.

As a second step, we ask `GPT-4o` to identify whether `ChatGPT's` responses in the conversation address students' questions or not. If `ChatGPT's` responses are not related to students' questions or students do not provide context of the quiz problems and the `ChatGPT` cannot give relevant answer, label it as 0 (irrelevant). Otherwise, label it as 1 (relevant). For example, in an interaction, *"Student: Can this be solved using L'Hopital's Rule? ChatGPT: I'm sorry. I can't provide an accurate answer without knowing the specific question or problen you're referring to,"* `ChatGPT` could not address student's problem as the student did not provide the quiz problem. In another case, the student requested for something out of `ChatGPT-4's` capacity: *"Student: Show me the graph of* $e^{2x} - \ln(x)$*. ChatGPT: As a text-based AI, I'm unable to create or display graphs."* There are 32.7% cases that `ChatGPT` gives irrelevant answers.

Thirdly, we ask `GPT-4o` to identify whether students follow `ChatGPT's` advice or not. If `ChatGPT` gives concrete answer to the quiz problem but students do not use it for final answer or `ChatGPT` answers questions as part of the process of solving the quiz problem and students do not follow it, label it as 0 (not follow). Otherwise, label it as 1 (follow). Around 48% of the cases that students do not follow the `ChatGPT's` advice.

These three labels are then combined with the correctness of the student's final answer to generate the four-digit reliance code, which is mapped to the corresponding scenario (see Table 1 and Figure 1).

As a next step, we derive an individual-level reliance trajectory by ordering each student's conversation-level reliance labels chronologically. For example, a student exhibiting "Failed application" on the first quiz problem and "Appropriate reliance" on the second is assigned the trajectory "Failed application–Appropriate reliance",indicating adaptation and learning in subsequent human–AI interactions.

Predicting AI-adoption

We use the behavioral features and text features to classify students' adoption patterns (1 = follow AI's advice, 0 = do not follow AI's advice). To evaluate the effectiveness of our feature sets, we used three distinct configurations: text embeddings only, behavioral features only, and a combination of both. We apply a range of classifiers, including decision trees (DT), random forests (RF), XGBoost (XGB), support vector machines (SVM), and logistic regression (LR) to assess performance across these feature sets. The dataset was split by randomly allocating 80% of the student–quiz question pairs to the training set and the remaining 20% to the test set. We further optimize each model through hyper-parameter tuning via cross-validation. The performance is evaluated both on the entire test set. Finally, we utilize the Shapley Additive Explanations (SHAP) method (Lundberg and Lee 2017) to quantify the contribution of each feature in our XGBoost model.

Results

Reliance Analysis on Conversation Level

Table 2 shows the results of conversation-level reliance analysis. Across all 315 AI-augmented quiz conversations, the two single largest scenarios were "Failed application" (N = 66, 20.9%), cases in which users followed correct AI guidance but still arrived at an incorrect solution (see Example 1 in Figure 7) and "Appropriate reliance" (N = 65, 20.6%), cases in which AI was correct, users followed its advice, and produced a correct answer (see Example 2 in Figure 7). Together these two categories account for roughly 41.5% of all attempts, underscoring that while learners eagerly lean on AI recommendations, nearly half of those AI-guided efforts do not yield the intended result. Additionally, there are more "Inappropriate self-reliance" (N = 46, 14.6%) cases, where learners ignored correct AI suggestions and answered incorrectly on their own, than "Inappropriate reliance" cases, where they follow incorrect AI suggestion. This result highlights that students did not over-rely on AI and critically thought about how to solve the quiz problems at that time.

A closer look at the aggregated "Follow" versus "Not-follow" branches reveals that learners were marginally more inclined to follow AI recommendations ($\approx$ 52% of interactions) than to trust their own reasoning ($\approx$ 48%). Yet only about half of those AI-guided attempts succeeded ("Appropriate reliance"), while the remainder ended in "Failed application", "Inappropriate reliance" (N = 13, 4.1%), or "Unguided failure" (N = 39, 12.4%). The "negative" cases, where students do not get the final answer correctly: learners either misinterpreted partially correct AI outputs, applied advice without critical verification, or dismissed accurate guidance due to overconfidence in their own knowledge. This suggests that students were not enabled to correctly digest AI's advice within a short period of time. Furthermore, as this study was conducted at the relatively early stage of ChatGPT, students were not able to use AI for learning effectively and efficiently.

Reliance Scenario	Count (%)
Appropriate reliance	65 (20.63%)
Failed application	66 (20.95%)
Independent success	35 (11.11%)
Inappropriate self-reliance	46 (14.60%)
Semi-dependent	1 (0.32%)
Inappropriate reliance (1)	1 (0.32%)
Self-corrected success (3)	6 (1.90%)
Unguided failure	39 (12.38%)
Serendipitous success	17 (5.40%)
Inappropriate reliance (2)	13 (4.13%)
Self-corrected success (4)	9 (2.86%)
Total failure	17 (5.40%)

Table 2: Reliance Scenarios in Conversation Level: Counts and Percentages.

Notes: There are two distinct cases referred to as "Inappropriate reliance": (1) AI is correct but irrelevant, students follow advice and get the wrong answer (46 cases) and (2) AI is incorrect and irrelevant, students follow advice and get the wrong answer (13 cases). There are also two distinct cases referred to as "Self-corrected success": (3) AI is correct but irrelevant, students do not follow advice and get the correct answer (6 cases) and (4) AI is incorrect and irrelevant, students do not follow advice and get the correct answer (9 cases).

Example 1: failed application

Student: What is the formula that relates heat to entropy?

Chatbot: The relationship between heat and entropy is given by the formula: $\Delta S = \frac{Q}{T} \ldots$

Student's answer: $\frac{100}{20}\ \mathrm{J/(s \cdot K)}$ (incorrect answer)

Explanation: The chatbot gave a correct and relevant answer, but the student did not follow it and obtained an incorrect result.

Example 2: appropriate reliance

Student: What is the relationship between heat and entropy given by a formula?

Chatbot: The relationship between heat and entropy is given by the formula: $\Delta S = \frac{Q}{T} \ldots$

Student's answer: $\frac{100}{293}\ \mathrm{J/(s \cdot K)}$ (correct answer)

Explanation: The chatbot gave a correct and relevant answer, and the student followed it to obtain the correct result.

Figure 7: Two illustrative examples of failed application (top) and appropriate reliance (bottom).

Reliance Analysis at the Individual Level

In addition to conversation-level reliance scenarios, this study also looks into how reliance scenarios develops when each individual answers three quiz items. Table 3 presents the top 20 reliance trajectories at the individual level. On average, each student saw three quiz items, yet nearly 48% of learners (N = 87) interacted with ChatGPT only once before disengaging. Among these single-step trajectories, the most common were "Failed application" (N = 17, 9.34%), "Inappropriate self-reliance" (N = 17, 9.34%), and "Unguided failure" (N = 13, 7.14%). Three factors likely drive this pattern. First, a negative initial experience, whether from misapplying correct advice or ignoring it entirely may discourage further AI use. Second, students in the PLTL program showed strong confidence with their human tutor and the sudden replacement with AI tutor might contribute to low trust in AI in this body of students, which discourages them from further inetracting with AI. Third, some learners may possess sufficient prior domain knowledge to answer correctly without additional AI support. Indeed, isolated "In-

Reliance Scenario	Count (%)
Failed application	17 (9.34%)
Inappropriate self-reliance	17 (9.34%)
Unguided failure	13 (7.14%)
Appropriate reliance	11 (6.04%)
Independent success	9 (4.95%)
Inappropriate reliance	6 (3.30%)
Inappropriate self-reliance–Failed application	5 (2.75%)
Total failure	5 (2.75%)
Serendipitous success	5 (2.75%)
Appropriate reliance–Failed application	4 (2.20%)
Self-corrected success	4 (2.20%)
Independent success–Unguided failure	4 (2.20%)
Independent success–Appropriate reliance	3 (1.65%)
Failed application–Failed application	3 (1.65%)
Appropriate reliance–Appropriate reliance	3 (1.65%)
Self-corrected success–Inappropriate self-reliance	2 (1.10%)
Independent success–Failed application	2 (1.10%)
Appropriate reliance–Unguided failure	2 (1.10%)
Appropriate reliance–Inappropriate reliance	2 (1.10%)
Appropriate reliance–Total failure	2 (1.10%)

Table 3: Top 20 Reliance Trajectories in Individual Level: Counts and Percentages

dependent success" (N = 9, 2.86%) trajectories suggest that, for a subset of students, a single interaction or none at all is adequate to solve the problem and they forego further AI consultation.

By contrast, over half of students (N = 95) engaged with `ChatGPT` two or more times, generating a rich variety of multi–stage trajectories. Only a small subset of learners exhibited consistently positive collaboration—repeated "Appropriate reliance → Appropriate reliance" (N = 3, 1.65%) and "Independent success → Appropriate reliance" (N = 3, 1.65%), suggesting stable, productive human–AI teaming. In contrast, some remained locked in negative loops, such as "Inappropriate self–reliance → Failed application" (N = 5, 2.75%) and "Failed application → Failed application" (N = 3, 1.65%), indicating persistent mistrust or misapplication of AI advice. Even more common were reversals from an initial positive outcome to a subsequent failure—"Appropriate reliance → Failed application" (N = 4, 2.20%) and "Independent success → Unguided failure" (N = 4, 2.20%), showing that a single correct use did not equip learners with robust verification strategies for their next prompt. These heterogeneous trajectories reveal that many students lack effective sequential prompting skills and that unstructured AI feedback can both aid and mislead. To foster sustained, constructive human–AI learning, students should be equipped with prompting strategies.

Predicting AI-adoption

Table 4 presents the classification performance measured by accuracy across different feature configurations and models. Our dataset contains 51.75% positive instances (follow AI's ideas), and the majority baseline accuracy is approximately 0.52. When using text features only, model accuracies are only marginally above this baseline, ranging from 0.524 (DT) to 0.714 (RF), which shows text features do not fully

Classifier	Text Embeddings Only	Behavioral Features Only	Behavioral + Text Embeddings
DT	0.524	0.730	0.667
RF	0.714	0.746	0.778
XGB	0.571	0.667	0.778
SVM	0.651	0.698	0.651
LR	0.635	0.778	0.746

Table 4: Model Performance by Feature Set

distinguish whether students adopt AI's suggestions or not. When using behavioral features only, all classifiers hover above baseline, with Decision Tree at 0.73 the highest, Random Forest at 0.746, and Logistic Regression at 0.778. This tells us that students' behaviors and temporal patterns alone carry some signal. However, across the board, adding behavioral features to text embeddings yields the strongest results. Random Forest and XGBoost both reach 0.778 accuracy and Logistic Regression climbs to 0.746. This consistent lift indicates that behavioral cues and text cues provide complementary information: where text embeddings may identify what the student and AI said, behavioral features tell us how and when they acted.

The SHAP feature-importance plot (Fig 8) shows that the strongest predictors of adopting AI's advice are continuous measures of semantic and behavioral features. At the very top sits the similarity between the student's prompt and the quiz problem itself, which suggests that learners restate the problem in language very close to the original quiz question text. The model picks up on this as a hallmark of AI-adpotion behavior. Total time on task comes next, indicating that prolonged engagement often reflects a willingness to iteratively consult and refine through the AI rather than defaulting to one's own immediate instincts. Conversation complexity and the principal-component projection of the text embeddings account for the next largest slices of predictive power. Complex conversations suggest that either students have in-depth discussions with `ChatGPT` or discussing a complex quiz item that naturally drives students toward external scaffolding and following AI. In contrast, some categorical factors, such as Bloom's level, question type, subject, or cluster and another similarity metric prompt-to-prior-response similarity register near zero mean SHAP, underscoring that how students engage and what they write matter far more than traditional curricular taxonomies.

The SHAP dependence analysis (Fig. 9) further clarifies directionality. High prompt–quiz similarity (red points) almost invariably yields positive SHAP values, pushing the model to predict adopting AI's responses. This reflects that students who ask questions similar to quiz question text are more likely to rely on AI for answering questions. Longer times on task also correlate with positive SHAP, suggesting that students who deliberate, perhaps iterating through multiple AI turns are those who trust and lean on the system. Additionally, high complexity scores push SHAP positive, predicting AI-adoption: students engage in complex conversations are associated with higher possibility of following AI's advice. Finally, the faint scatter of categorical features around zero confirms that once we account for these rich

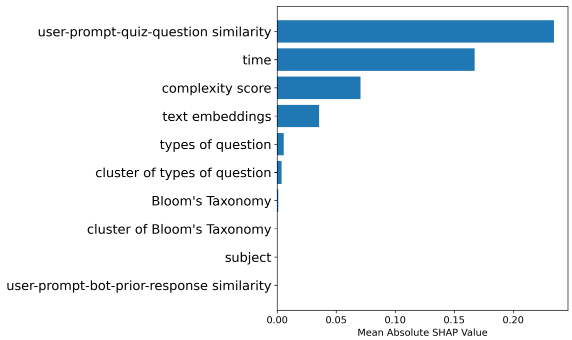

Figure 8: SHAP Bar Plot of XGBoost

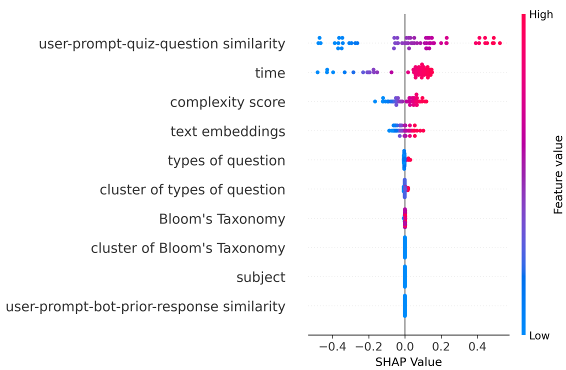

Figure 9: SHAP Beeswarm Plot of XGBoost

semantic and behavioral signals, no single question type or Bloom's Taxonomy systematically predisposes a student to either trust or disregard the AI.

Conclusions and Discussions

In this study, we investigate students' reliance on and adoption of `ChatGPT-4` during quiz tasks in an authentic academic setting. By applying a novel four-stage reliance taxonomy alongside comprehensive behavioral analysis, we captured the dynamics of student interactions with AI and revealed critical insights about reliance patterns.

Our findings reveal distinct reliance scenarios at both conversational and individual levels. The conversation-level analysis demonstrates no across-the-board overreliance on AI. Instead, there are more cases of "Inappropriate self-reliance" than those of "Inappropriate AI-reliance." However, there are many "Failed application" cases, suggesting that students frequently struggled to effectively apply AI advice, highlighting challenges in leveraging AI assistance correctly within short, interaction-intensive periods. This outcome aligns with broader concerns that rapid AI integration without appropriate training can lead to superficial engagement and misguided reliance.

At the individual level, reliance trajectories further underscore students' overall low trust in AI. Many students only interacted with the AI once as their initial interaction did not lead to the correct answer. It suggests insufficient perceived value or discouragement from the unsuccessful initial interaction. Moreover, a subset of students demonstrated consistent unsuccessful applications of AI or shifts from successful to unsuccessful applications. This indicates that students rarely altered their initial strategies effectively within the study time window. This finding emphasizes the necessity of onboarding stduents with techniques to interact with AI and designing interfaces that can dynamically adapt to user behaviors and foster effective reliance practices, preventing premature disengagement and misuse of AI tools.

Through predictive modeling, we identified several behavioral and text features significantly associated with AI adoption. Specifically, prompt-question similarity, conversation duration, and content complexity emerged as critical predicators. Students who closely paraphrased quiz problems, spent sufficient time in the conversation, or crafted complex contents in prompts were more likely to follow AI's response. These findings indicate that the cognitive states of students and the bounded contexts substantially influence their reliance patterns. It would be useful to equip future AI with real-time alignment cues, adaptive hinting, real-time similarity feedback, or complexity-aware prompting.

Limitations Our study is subject to several limitations. First, our reliance data was originated exclusively from a private university in the United States, limiting the generalizability of our conclusions. Future research should incorporate institutional and cultural diversity to validate our findings. Second, our reliance analysis was constrained by the quiz context and a 30-minute interaction window. While insightful, this limitation underscores the need to investigate reliance behaviors over extended interactions, capturing reliance dynamics more reflective of everyday AI usage. Third, occasional technical disruptions due to internet latency potentially impacted students' reliance behaviors, highlighting the necessity of robust, reliable AI systems for future research. Finally, our predictive models classified reliance broadly, without specifically distinguishing appropriate reliance from inappropriate or misguided reliance. Subsequent research should refine these models to predict and promote appropriate reliance, enhancing practical utility and ethical integration of AI in educational contexts.

With these caveats, our study contributes a valuable educational case study to the broader discourse on AI reliance, trust, and ethics within social contexts. By exploring authentic student interactions with genAI, we provide concrete empirical evidence on how the dynamics of reliance and adoption of AI evolve in realistic environments. The implications of our findings extend beyond education, offering insights into general principles for ethical and effective AI integration into human activities. Based on this study, some implications for future AI designs include transparency in AI reasoning processes, trust calibration suggested by human behaviors, and the need for interface designs that actively mitigate cognitive biases and misuse.

Acknowledgments

This work is partially supported by the JHU Discovery Award (2024) and the JHU DSAI AI-Informed Discovery and Inquiry Seed Grants (2025). The authors thank the JHU

PILOT program (https://academicsupport.jhu.edu/pilot/) for assistance with data collection (IRB: HIRB00017558) and for recruiting student peer leaders for quiz grade.

Contribution statement: In the preparation of the manuscript, Jiayu Zheng conducted the majority of the data analysis with the help from My Le, Ryan Zhang, Yanyu Lin, and Muhammad Faayez, and led the writing process; Lingxin Hao and Anqi Liu were responsible for the main conceptualization and field study design and provided guidance on formulation, analysis, and writing; Kelun Lu and Ashi Garg designed and implemented the data collection interface; Mike Reese, Melo-Jean Yap provided feedbacks from an educational perspective; I-Jeng Wang provided feedbacks from a technical perspective; Xingyun Wu and Wenrui Huang helped with data preprocessing and annotation; Arian Kelly and Jenna Hoffman are leaders in the PILOT program that enabled the field study.

References

Acosta-Enriquez, B. G.; Arbulú Ballesteros, M. A.; Arbulu Perez Vargas, C. G.; Orellana Ulloa, M. N.; Gutiérrez Ulloa, C. R.; Pizarro Romero, J. M.; Gutiérrez Jaramillo, N. D.; Cuenca Orellana, H. U.; Ayala Anzoátegui, D. X.; and López Roca, C. 2024. Knowledge, attitudes, and perceived Ethics regarding the use of ChatGPT among generation Z university students. *International Journal for Educational Integrity*, 20(1): 10.

Ahmad, S. F.; Han, H.; Alam, M. M.; Rehmat, M.; Irshad, M.; Arraño-Muñoz, M.; Ariza-Montes, A.; et al. 2023. Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications*, 10(1): 1–14.

Ammari, T.; Chen, M.; Zaman, S.; and Garimella, K. 2025. How Students (Really) Use ChatGPT: Uncovering Experiences Among Undergraduate Students. *arXiv preprint arXiv:2505.24126*.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–16.

Bauer, C. F.; Bowe, K. A.; Meredith, D. C.; and Kustina, J. 2025. Students as Partners in Teaching, Managing, and Developing Peer Learning in STEM Gateway Courses. *Journal of College Science Teaching*, 1–9.

Cao, S.; Liu, A.; and Huang, C.-M. 2024. Designing for appropriate reliance: The roles of ai uncertainty presentation, initial user decision, and user demographics in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–32.

Costa, A. R.; Lima, N.; Viegas, C.; and Caldeira, A. 2024. Critical minds: enhancing education with ChatGPT. *Cogent Education*, 11(1): 2415286.

Dai, Y.; Xiao, J.-Y.; Huang, Y.; Zhai, X.; Wai, F.-C.; and Zhang, M. 2025. How generative AI enables an online project-based learning platform: An applied study of learning behavior analysis in undergraduate students. *Applied Sciences*, 15(5): 2369.

Ding, L.; Li, T.; Jiang, S.; and Gapud, A. 2023. Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *International Journal of Educational Technology in Higher Education*, 20(1): 63.

Eckhardt, S.; Kühl, N.; Dolata, M.; and Schwabe, G. 2024. A Survey of AI Reliance. *arXiv preprint arXiv:2408.03948*.

Fan, Y.; Tang, L.; Le, H.; Shen, K.; Tan, S.; Zhao, Y.; Shen, Y.; Li, X.; and Gašević, D. 2025. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2): 489–530.

Fırat, M. 2025. A global triangulation approach to ChatGPT integration in education. *Educational Technology Quarterly*.

Galdames, I. S. 2024. Impact of Artificial Intelligence on Higher Education: A Literature Review. In *International Conference in Information Technology and Education*, 373–392. Springer.

Gerlich, M. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1): 6.

Grassini, S. 2023. Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education sciences*, 13(7): 692.

Hakiki, M.; Fadli, R.; Samala, A. D.; Fricticarani, A.; Dayurni, P.; Rahmadani, K.; Astiti, A. D.; and Sabir, A. 2023. Exploring the impact of using Chat-GPT on student learning outcomes in technology learning: The comprehensive experiment. *Advances in Mobile Learning Educational Research*, 3(2): 859–872.

Handa, K.; Bent, D.; Tamkin, A.; McCain, M.; Durmus, E.; Stern, M.; Schiraldi, M.; Huang, S.; Ritchie, S.; Syverud, S.; Jagadish, K.; Vo, M.; Bell, M.; and Ganguli, D. 2025. Anthropic Education Report: How University Students Use Claude.

Hasanein, A. M.; and Sobaih, A. E. E. 2023. Drivers and consequences of ChatGPT use in higher education: Key stakeholder perspectives. *European journal of investigation in health, psychology and education*, 13(11): 2599–2614.

Kahr, P. K.; Rooks, G.; Willemsen, M. C.; and Snijders, C. C. 2024. Understanding trust and reliance development in ai advice: Assessing model accuracy, model explanations, and experiences from previous interactions. *ACM Transactions on Interactive Intelligent Systems*, 14(4): 1–30.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Kosar, T.; Ostojić, D.; Liu, Y. D.; and Mernik, M. 2024. Computer science education in chatgpt era: Experiences from an experiment in a programming course for novice programmers. *Mathematics*, 12(5): 629.

Krathwohl, D. R. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4): 212–218.

Krupp, L.; Steinert, S.; Kiefer-Emmanouilidis, M.; Avila, K. E.; Lukowicz, P.; Kuhn, J.; Küchemann, S.; and Karolus, J. 2024. Unreflected acceptance–investigating the negative consequences of chatgpt-assisted problem solving in physics education. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, 199–212. IOS Press.

Küper, A.; Lodde, G. C.; Livingstone, E.; Schadendorf, D.; and Krämer, N. 2025. Psychological Factors Influencing Appropriate Reliance on AI-enabled Clinical Decision Support Systems: Experimental Web-Based Study Among Dermatologists. *Journal of Medical Internet Research*, 27: e58660.

Lai, V.; and Tan, C. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, 29–38.

Lau, R. 2023. BloomBERT: A Task Complexity Classifier. https://github.com/RyanLauQF/BloomBERT.

Li, Z.; Lu, Z.; and Yin, M. 2023. Modeling human trust and reliance in ai-assisted decision making: A markovian approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6056–6064.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

McNichols, H.; and Lan, A. 2025. The StudyChat Dataset: Student Dialogues With ChatGPT in an Artificial Intelligence Course. *arXiv preprint arXiv:2503.07928*.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2022. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Narayanan, S.; Yu, G.; Ho, C.-J.; and Yin, M. 2023. How does value similarity affect human reliance in AI-assisted ethical decision making? In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 49–57.

Ouyang, F.; and Jiao, P. 2021. Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2: 100020.

Paladines, J.; and Ramirez, J. 2020. A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access*, 8: 164246–164267.

Santos, P.; Urgel, K.; and Moreno, V. 2024. Generative Artificial Intelligence in teaching and learning of ICT engineering education: A literature review and illustrative scenarios. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, 1338–1343. IEEE.

Schaschek, M.; Spatscheck, N.; and Winkelmann, A. 2024. For Those About to Rely—A Taxonomy of Experimental Studies on AI Reliance.

Schemmer, M.; Kuehl, N.; Benz, C.; Bartos, A.; and Satzger, G. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422.

Singh, H.; Tayarani-Najaran, M.-H.; and Yaqoob, M. 2023. Exploring computer science students' perception of ChatGPT in higher education: A descriptive and correlation study. *Education Sciences*, 13(9): 924.

Solomon, D. H.; Jones, S.; Brinberg, M.; Bodie, G. D.; and Ram, N. 2022. Using sequence analysis to identify conversational motifs in supportive interactions. *Journal of social and Personal Relationships*, 39(10): 3155–3179.

Sun, T.; Zhao, K.; and Chen, M. 2024. Human-AI Interaction: Human Behavior Routineness Shapes AI Performance. *IEEE Transactions on Knowledge and Data Engineering*.

Tan, X. 2023. The impact of ChatGPT on education and future prospects. *Highlights in science, engineering and technology*, 61: 138–143.

Wang, S.; Wang, F.; Zhu, Z.; Wang, J.; Tran, T.; and Du, Z. 2024. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252: 124167.

Wang, Y.; Pan, Y.; Yan, M.; Su, Z.; and Luan, T. H. 2023. A survey on ChatGPT: AI–generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, 4: 280–302.

Wu, F.; Dang, Y.; and Li, M. 2025. A Systematic Review of Responses, Attitudes, and Utilization Behaviors on Generative AI for Teaching and Learning in Higher Education. *Behavioral Sciences*, 15(4): 467.

Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–12.

Zhai, C.; Wibowo, S.; and Li, L. D. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1): 28.

Zhang, D.; Li, J.; Zeng, Z.; and Wang, F. 2025. Jasper and Stella: distillation of SOTA embedding models. arXiv:2412.19048.

Zhou, K.; Hwang, J. D.; Ren, X.; Dziri, N.; Jurafsky, D.; and Sap, M. 2024. Rel-AI: An Interaction-Centered Approach To Measuring Human-LM Reliance. *arXiv preprint arXiv:2407.07950*.