Computers & Education 234 (2025) 105329

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Measuring undergraduate students' reliance on Generative AI during problem-solving: Scale development and validation

Chenyu Hou [a], Gaoxia Zhu [b,*], Vidya Sudarshan [c], Fun Siong Lim [d], Yew Soon Ong [c]

[a] Nanyang Technological University, Singapore
[b] National Institute of Education (NIE), Nanyang Technological University, Singapore
[c] College of Computing & Data Science, Nanyang Technological University, Singapore
[d] Applications of Teaching and Learning Analytics for Students, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Human-AI collaboration
Problem-solving
Generative AI
Higher education
Reliance on AI
Scale development

ABSTRACT

Reliance on AI describes the behavioral patterns of when and how individuals depend on AI suggestions, and appropriate reliance patterns are necessary to achieve effective human-AI collaboration. Traditional measures often link reliance to decision-making outcomes, which may not be suitable for complex problem-solving tasks where outcomes are not binary (i.e., correct or incorrect) or immediately clear. Therefore, this study aims to develop a scale to measure undergraduate students' behaviors of using Generative AI during problem-solving tasks without directly linking them to specific outcomes. We conducted an exploratory factor analysis on 800 responses collected after students finished one problem-solving activity, which revealed four distinct factors: reflective use, cautious use, thoughtless use, and collaborative use. The overall scale has reached sufficient internal reliability (Cronbach's alpha = .84). Two confirmatory factor analyses (CFAs) were conducted to validate the factors using the remaining 730 responses from this activity and 1173 responses from another problem-solving activity. CFA indices showed adequate model fit for data from both problem-solving tasks, suggesting that the scale can be applied to various human-AI problem-solving tasks. This study offers a validated scale to measure students' reliance behaviors in different human-AI problem-solving activities and provides implications for educators to responsively integrate Generative AI in higher education.

1. Introduction

Artificial intelligence (AI), defined as digital machines capable of performing cognitive tasks similar to human minds, such as learning and problem-solving, has revolutionized the ways people teach and learn (Chiu et al., 2023; Tahiru, 2021). AI, particularly Generative AI, has emerged as a tool that can support learning, offering students assistance in exploring solutions and generating ideas (Grassini, 2023). In higher education, problem-based learning (PBL) is increasingly recognized for its emphasis on real-world and interdisciplinary challenges (Strobel & Barneveld, 2009), which can actively engage students in complex problem-solving processes, encourage them to explore diverse perspectives and rich information, and develop practical solutions through iterative inquiries

* Corresponding author. Learning Sciences and Assessment Department, National Institute of Education (NIE), Nanyang Technological University, Singapore.
E-mail addresses: chenyu004@e.ntu.edu.sg (C. Hou), gaoxia.zhu@nie.edu.sg (G. Zhu), vidya.sudarshan@ntu.edu.sg (V. Sudarshan), lim_fun_siong@ntu.edu.sg (F.S. Lim), asysong@ntu.edu.sg (Y.S. Ong).

https://doi.org/10.1016/j.compedu.2025.105329
Received 5 November 2024; Received in revised form 31 January 2025; Accepted 18 April 2025
Available online 19 April 2025

(Wijnia et al., 2019). However, introducing AI in PBL poses significant challenges, particularly concerning potential students' over-reliance on AI tools which could undermine their problem-solving skills. Studies have shown that students might accept automated feedback provided by AI tools with minimal cognitive engagement (Kasneci et al., 2023; Koltovskaia, 2020). Inappropriate reliance on AI may erode students' ability to critically evaluate information and engage in metacognitive practices in the PBL context (Hamid et al., 2023). Therefore, understanding how college students interact with and depend on Generative AI during PBL is crucial for ensuring that learning goals are not being compromised because of students' inappropriate reliance behaviors.

Reliance on AI systems, the behavioral patterns of when and how individuals depend on AI suggestions (Wang et al., 2008), is a pivotal aspect of human-AI collaboration that can influence whether AI systems enhance or diminish human performance. Introducing AI to assist complex tasks can enhance overall performance, but individuals need to develop the ability to identify and reject erroneous AI advice (Bansal et al., 2019). Several measures have been developed to quantify reliance behaviors, typically emphasizing collective performance between AI and humans. For instance, some instruments measure reliance based on the decision-making outcomes (e.g., Schemmer et al., 2022), while others take a prospective method to measure reliance as the probability that humans would adopt AI suggestions to maximize potential benefits (e.g., Fok & Weld, 2024). These measures are limited in their ability to capture students' inobservable reliance behaviors on Generative AI during problem-solving tasks that involve implicit shifts in cognitive strategies based on interactions with the environment or other participants (Chiou & Lee, 2023). Indeed, problem-solving is a dynamic and multi-faceted activity that requires ongoing collaboration and adaptation among participants. They need to develop shared mental models by communicating and integrating diverse knowledge to achieve a common goal (Larson Jr & Christensen, 1993). This process encompasses multiple interrelated stages, such as problem identification, information exploration, and solution generation and presentation (Liu & Pásztor, 2022), each requiring cognitive skills to carry out (Behfar et al., 2008). Effective problem-solving is nonlinear and a more intricate process than decision-making, which typically involves explicit choices and outcomes (Hesse et al., 2015).

Recognizing the limitations of existing measures, the current study aims to develop a scale that measures reliance behaviors during human-AI problem-solving. As Generative AI tools are incorporated into collaborative problem-solving tasks in classrooms, there is a pressing need to assess how students make use of Generative AI's suggestions, as inappropriate reliance on AI could exacerbate inequalities in learning and compromise the rigorousness of academic works (Al-kfairy et al., 2024). The self-report questionnaire aims to address this need by providing a nuanced, user-centered measure of AI reliance in the human-AI problem-solving context. By enumerating use cases of Generative AI across the three stages (i.e., problem identification, information exploration, and solution generation and presentation; Liu & Pásztor, 2022) of problem-solving, the current study seeks to uncover patterns of reliance that might not be easily observed through objective measures. Furthermore, leveraging a large-scale survey in a course taken by undergraduates across a university, this research will provide an overview of how undergraduate students interact with Generative AI in problem-solving activities. Specifically, the research questions we aim to address are:

1) What dimensions of undergraduate students' reliance behaviors emerge from questionnaire items?
2) To what extent do the developed items demonstrate reliability across different types of problem-solving tasks involving Generative AI?
3) How do undergraduate students' reliance behaviors on Generative AI differ across problem-solving tasks?

2. Literature review

2.1. Problem-based learning and problem-solving

PBL is an instructional approach widely adopted in higher education, focusing on student-driven collaborative exploration of real-world, ill-structured problems. These problems do not have clearly defined solutions or paths to resolution but provide real and complex scenarios for inquiry, collaboration between students, sharing of diverse perspectives, integration of ideas, and deep learning (Barrows, 1996; Wijnia et al., 2019). PBL fosters the development of procedural knowledge, such as reasoning, and declarative knowledge, which involves understanding the principles underlying a problem (Schmidt et al., 2009).

The central focus of PBL is problem-solving. While there are various models of PBL, they all tend to follow a structured problem-solving process involving three key stages: problem identification, information exploration, and solution generation and presentation (Liu & Pásztor, 2022). During problem identification, students need to define the problems within the given problem scope and brainstorm initial ideas or hypotheses to identify knowledge gaps (Schmidt, 1993, 2012). In the information exploration stage, students engage in iterative research to address these gaps, sharing and analyzing information while connecting it to broader concepts (Wijnia et al., 2019). Finally, in the solution generation and presentation stage, students consolidate their findings, develop solutions, and share their solutions with the larger community, reflecting on both the solutions and their learning process (Koschmann, 1994).

2.1.1. Human-AI problem-solving

Ideally, human-AI collaboration, where humans and AI systems work together by leveraging their unique strengths, can enhance human cognition (Dellermann et al., 2019). Specifically, AI can quickly process vast amounts of data and identify patterns, while humans can contribute nuanced reasoning, creativity, and contextual understanding (Braga & Logan, 2017). In educational settings, AI can enhance learning by offering personalized instruction, modeling learning behaviors, supporting formative assessments, and providing timely feedback (Tan et al., 2022). In problem-solving, AI technologies have evolved from simply reducing labor costs to providing cognitive support in tasks such as data analysis, pattern recognition, and even generating creative solutions (Joksimovic et al., 2023; Seeber et al., 2020). In addition, AI tools can provide feedback on the problem-solving process to help learners monitor and

reflect on their problem-solving strategies (Joksimovic et al., 2023).

Furthermore, Generative AI, a category of AI that can produce diverse outputs such as text, images, music, and code based on users' natural language input (i.e., prompt engineering), is reshaping educational practices (Zhao et al., 2023). Generative AI has demonstrated impressive capabilities across complex tasks like arithmetic problems, conceptual mappings, and multidisciplinary understanding (Wei et al., 2022). Hence, Generative AI has the potential to facilitate students' problem-solving during PBL. Generative AI can aid the problem-solving process by acting as a tool for knowledge generation, idea exploration, and feedback (Joksimovic et al., 2023). In the problem identification stage, Generative AI may help brainstorm potential hypotheses and simulate different problem scenarios. For example, Valdivia and colleagues (2024) used Generative AI to generate different problem scenarios concerning unit conversion and thermodynamic processes for students to solve. Their results show that students demonstrate higher-order cognitive levels, including analyzing, synthesizing, and evaluating in these problem-solving tasks. During information exploration, Generative AI can assist in gathering relevant information, summarizing complex and rich information, and even suggesting new directions for inquiry. For example, the study on the GPT-Assisted Summarization Aid (GASA, Lin et al., 2024) demonstrated how Generative AI helped students in STEM education by summarizing group discussions, providing real-time feedback on technical questions, and guiding reflective thinking. GASA can support students in identifying and organizing key concepts during collaborative tasks and suggest advanced approaches for integrating knowledge. Lastly, in the solution generation and presentation stage, Generative AI can offer alternative perspectives, help students refine their solutions, and help them prepare coherent solutions. For example, in Naik's et al. (2024) study on collaborative programming learning, a GPT-based tool was integrated into a cloud-based platform to provide dynamic and personalized reflection triggers. These triggers, tailored to the Structured Query Language (SQL) commands entered by student groups, prompted alternative solutions and facilitated discussions on key programming concepts, such as datatype selection, indexing strategies, and denormalization trade-offs. While the integration of these tailored reflection triggers influenced how students allocated their time, changing completion rates per problem, it did not diminish the overall amount learned.

Integrating Generative AI into PBL has the potential to provide students with tailored feedback and engage them in deeper problem exploration, making the learning process more dynamic and efficient. For example, Hamid and colleagues (2023) investigated the use of ChatGPT in pharmaceutical chemistry PBL sessions with 18 participants. Qualitative analysis of the collaborative process showed that integrating GPT in the process enhanced collaboration, engagement, and motivation, but the accuracy and depth of AI-generated information required additional human verification (Hamid et al., 2023). These findings underscore the dual-edged nature of integrating Generative AI in education. Issues such as misinformation, biases embedded in AI algorithms, and academic dishonesty compound the risks of AI integration in education (Al-kfairy et al., 2024). Generative AI tools may inadvertently reinforce stereotypes or factual inaccuracies due to biases in training data, creating inequities in learning outcomes (Hacker et al., 2023). In addition, AI generated content raises questions about authorship and intellectual integrity, necessitating transparency in AI usage and the development of mechanisms for attribution and verification (Malik et al., 2023).

Indeed, inappropriate reliance on Generative AI could undermine students' high-order competencies such as critical thinking and problem-solving skills (Kasneci et al., 2023). Research has shown that students often accept automated feedback with little cognitive engagement, highlighting the risk of thoughtless reliance on AI tools (Koltovskaia, 2020). Such passive acceptance of AI-generated responses can erode critical thinking and analytical processes central to PBL (Kadaruddin, 2023; Sun et al., 2025). Thus, it is vital for educators to actively monitor students' interactions with Generative AI to ensure it supplements rather than supplants their cognitive effort.

Thus, more research is needed to systematically examine the use of Generative AI in PBL (Hamid et al., 2023). The first step is to develop proper measures to uncover how students use Generative AI during problem-solving tasks. The measure should evaluate not only the frequency of AI use but also the depth of students' cognitive engagement with AI-generated content. Emphasis must be placed on distinguishing between user behaviors that support learning and detrimental overreliance that stifles independent reasoning.

2.2. Measures for reliance on AI systems

Based on the definition of reliance behaviors on AI systems, previous studies have sought to quantify these behaviors through various empirical measures (e.g., Buçinca et al., 2021; Vasconcelos et al., 2023). These measures typically focus on observable traces of reliance on AI suggestions. For instance, the Switch Fraction measures the ratio of users changing their answers to match AI suggestions (Kim et al., 2023). Another common measure, the Weight of Advice, evaluates how much influence AI's recommendations have on users' final decisions relative to their initial independent choices. If the final decision aligns more closely with the AI's recommendation than the user's initial judgment, it indicates a higher reliance on AI (Passi & Vorvoreanu, 2022). However, these measures fail to differentiate between correct and incorrect AI suggestions, neglecting the fact that AI can also make errors. Users need to discern the quality of AI advice and adjust their reliance accordingly.

To address the limitations of the measures, Schemmer et al. (2022) proposed that measures for reliance on AI should also emphasize users' ability to discriminate between correct and incorrect AI advice and act accordingly. Their measure involves a sequential decision-making process where users make an initial decision, receive AI advice, and then decide whether to follow or reject the advice. This model introduces two dimensions: relative AI reliance, which tracks how often users rightfully follow AI recommendations, and relative self-reliance, which measures how often users correctly reject incorrect AI advice. Although this approach provides a more nuanced understanding of reliance, it remains limited to binary decision-making contexts. This measure relies heavily on immediate feedback about the correctness of decisions, which is not always available in real-world settings. For example, in financial decisions like investments, the appropriateness of relying on AI predictions may fluctuate over time, depending on market conditions. Thus, the two-dimensional reliance construct is restricted to structured, outcome-dependent decision-making tasks and does not

account for more complex, non-binary human-AI collaboration scenarios (Fok & Weld, 2024).

To overcome the above-mentioned limitations, Fok and Weld (2024) introduced the concept of strategy-graded reliance, which focuses on reliance behaviors based on expected AI performance rather than outcome correctness. Unlike outcome-graded reliance, which evaluates decisions retrospectively, strategy-graded reliance prospectively looks at cognitive strategies that users might use to maximize overall performance. This shift from focusing on outcomes to strategies offers a more dynamic view of reliance, considering human and AI interactions before the final decision is made. Building on strategy-graded reliance, Guo et al. (2024, pp. 221–236) advanced the understanding of reliance by applying decision theory to assess reliance on AI based on expected comparative performance. Their measure separates mis-reliance (over- or under-reliance on AI) and discrimination loss (the inability to identify when AI advice is better than human judgment), offering a more refined measurement. The reliance level is quantified as the probability of humans following AI advice if AI's recommendations differ from human judgment, providing insights into how users balance their own judgments with AI suggestions (Guo et al., 2024).

While these measurements offer valuable insights into human reliance levels on AI, they still fall short of capturing the intricacies of problem-solving. Strategy-graded reliance assumes that users can accurately estimate AI's performance, which does not account for cognitive biases or task complexity. Additionally, the measure presupposes a shared understanding between humans and AI, which is often absent in complex tasks. In real-world problem-solving, reliance behaviors might involve subtle cognitive shifts or dynamic adjustments that these models cannot fully capture. Therefore, while useful in structured decision-making, existing measures do not fully address the complexity of human-AI problem-solving, where reliance behaviors are implicit and context-dependent.

2.3. The current study

The complexities of problem-solving call for a more comprehensive measure that can account for the diverse reliance behaviors that arise in the context (Joksimovic et al., 2023). Problem-solving tasks, unlike simpler decision-making tasks, require more than binary responses to AI suggestions. They involve exploration, iterative adjustments, and creative reasoning, where reliance behaviors may not always be explicit or immediately observable. In such settings, humans may adopt AI suggestions implicitly, gradually integrating them into their cognitive strategies (Vasconcelos et al., 2023).

Given this gap, this study aims to develop and validate a self-report questionnaire to quantify reliance behaviors during human-AI collaboration in problem-solving. The rationale behind this approach is that self-reported data can capture aspects of reliance behaviors that are difficult to observe directly. Potentially, this measure will provide insights into how students interact with Generative AI during PBL, helping educators better understand the balance between human reasoning and AI assistance.

Table 1
Questionnaire items mapped to the problem-solving stages and reliance categories.

Problem-solving stages	Reliance categories		
	AI reliance	Self-reliance	Peer reliance
Problem identification	R1: I copied the task descriptions or problem statements to Generative AI for help. R6: I feel that Gen AI is more capable of doing these activities than me.	R5: I used Generative AI only when I (we) had limited or no clue how the problem could be solved. R9: I designed my own prompts for Generative AI so it could give me the output I need. R15: I revised my prompting questions so Generative AI can give more appropriate outputs.	R20: We discussed how to ask prompting questions so Generative AI can give more appropriate outputs. R21: I asked my peer(s) questions during the problem-solving activities.
Information exploration	R3: For each activity I did in this class, I spent most of the time interacting with Generative AI. R8: I feel that Generative AI played a more important role in our problem-solving activities than me.	R7: I asked Generative AI to improve my initial draft/solutions for the problem-solving activities. R10: I read Generative AI's output critically when the output was generated. R11: I modified Generative AI's output to make it more applicable to the problem-solving activities. R12: I spotted errors in the Generative AI's output. R13: I found Generative AI's output is not perfect.	R17: We discussed the Generative AI's output when the output was generated. R22: My peers answered my questions and supported my understanding.
Solution generation and presentation	R2: I copied the Generative AI's output as part of our solution. R4: I adopted Generative AI's output without major changes.	R14: I discarded Generative AI's output. R16: I finished a subtask in problem-solving activities with no or limited help from the Generative AI.	R19: My peers helped me modify Generative AI's output. R18: My peer(s) helped me spot errors in the Generative AI's output. R23: I adopted a solution initially proposed by peer(s).

3. Methods and materials

3.1. Questionnaire items development

The study developed questionnaire items that describe specific behavioral cases of how students utilize Generative AI during human-AI problem-solving. The development process was grounded in a systematic approach that sought to map the behavioral traces of human-AI problem-solving across problem-solving stages (i.e., problem identification, information exploration, and solution generation and presentation) and reliance subjects (i.e., Generative AI, self, or peers).

The research team referred to a study (Zhu et al., 2024) that we conducted in a problem-solving context and with samples similar to the current study to develop the initial items. Specifically, Zhu and colleagues (2024) collected 79 responses from students after the problem-solving tasks on two reflection questions: "How would you describe your interaction with ChatGPT during the maze design task? In terms of contributions, did you feel that you, ChatGPT, or both equally contributed to the design of the maze? Please elaborate" (Zhu et al., 2024). After reviewing these reflective responses, we enumerate the behavioral cases of using Generative AI during the problem-solving tasks (e.g., We discussed how to ask prompting questions so Generative AI can give more appropriate outputs; I asked Generative AI to improve my initial draft/solutions for the problem-solving activities). Accordingly, we developed 23 items in total to capture possible reliance behaviors during human-AI problem-solving as shown in Table 1. For example, item R3 (For each activity I did in this class, I spent most of the time interacting with Generative AI) was developed because some students spent considerable time testing the capacity of Generative AI but ignored the problem-solving objectives. Similarly, item R1 (I copied the task descriptions or problem statements to Generative AI for help) was developed because the researchers observed that many students started the task by copying the task description to GPT.

As shown in Table 1, we mapped these questionnaire items to the three main problem-solving stages and the possible subjects that an individual can rely on (i.e., AI, self, and peers). Specifically, during problem identification, AI reliance includes tasks such as copying prompts directly into the Generative AI, while self-reliance involves formulating custom prompts, and peer reliance focuses on collaborative clarification of goals or strategies. In the information exploration stage, AI reliance includes extensive interaction and direct use of AI outputs, self-reliance includes critical evaluation and revision of outputs, and peer reliance includes seeking feedback and clarifications from peers. Finally, during solution generation and presentation, AI reliance involves minimal modification of AI outputs, self-reliance reflects independent task completion or rejecting AI suggestions, and peer reliance involves group revisions of AI-generated content. All the questions were constructed on a five-point Likert-type scale to measure the frequency of the behaviors during human-AI problem-solving (1 = Almost never, 5 = Almost always).

After the initial questionnaire was developed, we invited a panel of experts (2 professors with expertise in human-AI collaboration and 3 graduate students who observed the students' interaction with Generative AI during the study; Zhu et al., 2024) to further review the items to ensure that the items were easy to interpret and choices were easy to make. We improved the wording of the questions to make them clear and easy to interpret. For example, one item, "I reiterated the task requirements to GPT multiple times to let GPT generate good outputs.", was reworded to "I revised my prompting questions so Generative AI can give more appropriate outputs" to demonstrate students' autonomy in using Generative AI. In addition, two items (i.e., R6 and R8) were flagged at this stage because they did not describe specific behaviors, but all 23 items were administered to students after the two problem-solving activities, respectively.

3.2. Participants

The participants of this study were first-year and second-year students enrolled in a digital literacy course at an Asian university. In total 3030 students across various schools in the university enrolled in the digital literacy course, and they were divided into 70 tutorials, with around 20 to 50 students in each. In each tutorial, 5 to 6 students from different schools were purposefully assigned to a group that would work on various problem-solving activities in class. We collected 1818 and 1406 responses in two consecutive weeks towards the end of the semester. The number of responses collected from the second week was lower because some of the instructors did not allow students to use Generative AI that week.

The project received approval from the university's ethics review board to waive consent forms from the participants because the course took place in a regular educational setting, the risk was minimal, the participation was voluntary, and the survey did not take much time from participants and could be considered as their reflections after the problem-solving activities. As such, all students enrolled in this course were encouraged to complete the questionnaire, and no personal information other than their student ID was collected to minimize the risks involved. Based on open data from the university, of all the undergraduate students enrolled in 2023, 54.6 % are male, and 45.4 % are female; of students enrolled in 2022, 54.2 % are male and 45.8 % are female.

3.3. Problem-solving activities

As mentioned earlier, students in this course participated in various in-class problem-solving activities to improve their digital literacy, with two activities involving Generative AI. During the first activity (3D Maze week), students co-designed and co-solved 3D mazes on a game-based programming platform called 3D Maze. In the first part, they created challenging mazes by placing environmental assets (e.g., rocks, trees, rivers) on a map. In the second part, they solved the maze using pseudo-codes like "turn left" and "if-else loops." For both activities, students are encouraged to use Generative AI to design challenging mazes and solve them with pseudo-codes (See ***Appendix B*** for details).

In the second activity (Policy Drafting week), students drafted a policy proposal addressing the fair and ethical use of Generative AI in undergraduate education. They reviewed university guidelines, brainstormed common misuse cases, and evaluated their proposed policies from interdisciplinary perspectives. Students were allowed to use Generative AI freely throughout the problem-solving activity (See Appendix A for details).

3.4. Procedures

The flipped classroom approach was applied in this course. Before class, students were required to watch instructional videos on related topics and complete the subsequent quizzes. Then, they dedicated 2 hours per week to complete the two in-class problem-solving activities involving Generative AI. When coming into the class, all the students reviewed key concepts with the instructors. Then, the instructors gave them handouts that included problem-solving activity requirements and emphasized the need for inter-disciplinary collaboration among group members. After the problem-solving activities, the instructors reminded students to finish the reliance behaviors questionnaire. However, not all the enrolled students finished the problem-solving activities either because they were absent, time was used up, or their respective instructors adopted different designs rather than using the standard materials provided by the curriculum committee. Thus, some students also did not complete the questionnaire.

3.5. Data analyses

The data analyses include three steps to ensure that the instrument is reliable and valid for measuring reliance behaviors during human-AI problem-solving. First, during data cleaning, we addressed the issue of careless responses by identifying and removing responses that exhibited patterns indicative of careless responses (see details in the Careless Response below). Following the data cleaning, we conducted an Exploratory Factor Analysis (EFA) using part of the responses from the 3D Maze week to explore the dimensions that emerged naturally from the data to answer what dimensions of reliance behaviors emerge (RQ1). Furthermore, to examine the questionnaire across problem-solving tasks involving Generative AI (RQ2), we conducted two Confirmatory Factor Analyses (CFAs) to validate the factor structure identified in the EFA and assess the model's reliability and validity across two problem-solving tasks: 3D Maze and Policy Drafting tasks. Also, we conducted a qualitative validation (Ross et al., 2012; Torrance, 2012) by examining the alignment between students' self-reported questionnaire responses and their actual behaviors using Generative AI, as reflected in their chat histories. Specifically, we analyzed eight chat histories submitted by student groups from two problem-solving activities. We purposefully chose groups that had average questionnaire scores exceeding 4.5 for each factor identified from EFA over the two weeks. We focused on high-scoring groups because their behaviors can be more representative of the factors and can provide explicit and clear evidence of the distinctions among different factors. Lastly, to investigate how undergraduate students' reliance behaviors on Generative AI differ across problem-solving tasks (RQ3), we conducted a descriptive analysis and *t*-test of students' reliance behaviors across the two problem-solving activities to examine undergraduate students' reliance patterns on Generative AI.

3.5.1. Careless responses

After an initial review of the data collected, we found some cases of careless responses—answers provided without proper attention or thought, often characterized by random, repetitive, or inconsistent patterns (Kam, 2019). These responses can compromise the quality and accuracy of the data, weaken the relationships among survey variables, and impact factor loading and structure (Meade & Craig, 2012). Therefore, we first screened all the responses to exclude careless responses. For each response, we computed the longstring index, which is the maximum number of consecutive identical responses, to examine the invariability. There is no unified recommended threshold on the longstring index, as participants can respond with identical answers for similar items (Kupffer et al., 2024). Hence, we plotted the distribution of the average length of consecutive identical responses for students' responses to the two problem-solving activities separately. As the reflection surveys administered at the end of both weeks' problem-solving activities included some items other than the reliance behaviors items, we calculated the longstring index of all the survey questions. For the 3D

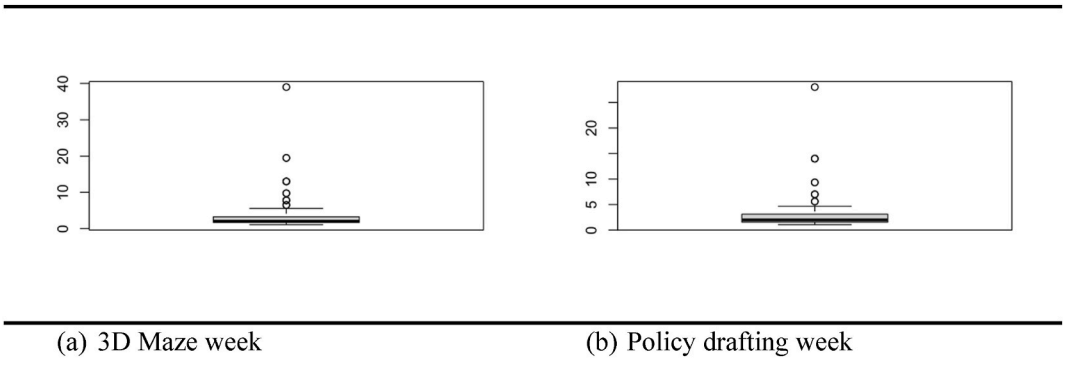


Fig. 1. The Boxplots of the Average Longstring Index of Survey Responses
Notes. The longstring index was calculated based on complete reflection surveys for both weeks, with 39 and 28 items, respectively.

Maze week (39 items in the reflection survey), the average longstring index ranges from 1.11 to 39, with a mean of 3.75 and a standard deviation of 3.98 (See Fig. 1a). For the Policy Drafting week (28 items), the average longstring index ranges from 1.08 to 28, with a mean of 3.84 and a standard deviation of 4.23 (See Fig. 1b). We identified outliers, which are data points that exceed 1.5 times the interquartile range above the third quartile of the average longstring index and excluded these responses from further analyses. In the end, 1540 surveys were retained from the 3D Maze week, and 1173 surveys were retained from the Policy Drafting week.

3.5.2. The exploratory factor analysis and confirmatory factor analysis

We first used almost half of the data from the 3D Maze week (800 survey responses) for an EFA. After a preliminary review of correlation among all the factors, the two flagged questionnaire items (i.e., R6 and R8) were weakly correlated with all other items and, therefore, were excluded in the following analyses (See Fig. 2 for the correlation visualization). EFA was used to uncover the underlying factor structure that emerges naturally from the data as it can determine the number of latent factors in the factor of interest. This study performed EFA with the Maximum Likelihood (ML) factoring method and oblique rotation. The ML method is suitable for determining latent factors (Fabrigar et al., 1999), and oblique rotation can identify correlated latent factors. Oblique rotation is preferred when developing a new scale (Schmitt & Sass, 2011). This study used a .30 cut-off point for evaluating factor loading (Cudeck & O'Dell, 1994). An item was defined as a cross-loading item and was removed if its factor loading was greater than .3 on two factors and the differences between the loadings were less than .15 (Schmitt & Sass, 2011).

After the EFA, to validate the structure of the factor models, we conducted two CFAs with the remaining 740 survey responses from the 3D Maze week and 1173 survey responses from the Policy Drafting week. Multiple indices could be used for evaluating the final factor model: the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standard root mean residuals (SRMR; Kline, 2005). The CFI is a goodness-of-fit measure that compares the fit of the factor model to a null model, accounting for sample size and model complexity. A CFI value over .90 is considered acceptable (Hu & Bentler, 1999). The RMSEA is a fit index to assess how well a hypothesized model, with its estimated parameters, approximates the population covariance matrix. An RMSEA value below .08 is considered acceptable (Marsh et al., 2004). The SRMR is a fit index to measure the average discrepancy between the observed correlation and the model's predicted correlations. An SRMR value below .08 is considered evidence of a good fit (Marsh et al., 2004). To assess the internal reliability of the developed scale, we calculated Cronbach's alpha, with .70 or higher indicating adequate reliability and .6 or higher indicating acceptable reliability (e.g., Azmi et al., 2024). We also calculated the mean inter-item correlations (MICs), with .20–.40 indicating the optimal range (Piedmont, 2014).

4. Results

4.1. RQ1: what dimensions of undergraduate students' reliance behaviors emerge from questionnaire items?

A total sample of 800 was used for the EFA to determine the latent structure of reliance behaviors during human-AI problem-solving. Fig. 2 illustrates the visualization of the correlation matrix among questionnaire items. We adopted polychoric correlation because the scale is ordinal on a Likert scale from 1 to 5 (Holgado–Tello et al., 2010). Then, we performed the Bartlett test of Sphericity and concluded that the variables were correlated enough to proceed with factor analysis (X^2 (df = 190) = 6760.16, $p < .05$). We also performed the Kaiser-Meyer-Olkin test to assess the adequacy of the sample size for factor analysis and concluded that the factor analysis is appropriate with an overall RMSA value of .92. To determine the number of latent factors, we performed a parallel analysis in R studio using *psych* package (Henson & Roberts, 2006). The parallel analysis suggested 4 factors.

Table 2 illustrates the specific statistics of the 4-factor model. R5 was removed as the factor loading was lower than .30 on all four

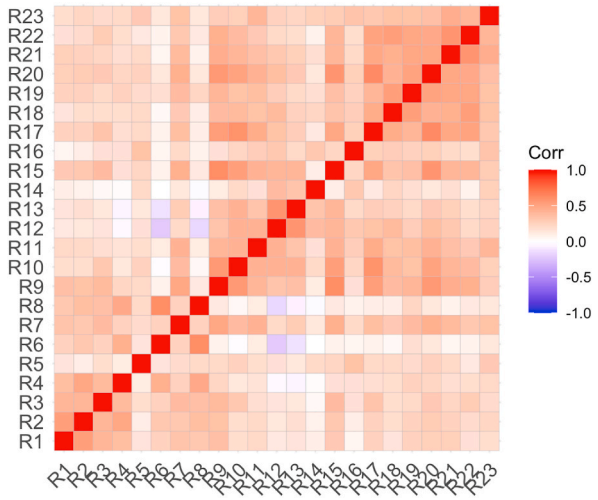


Fig. 2. The correlation matrix of questionnaire items.

Table 2
Exploratory factor analysis results.

	Factor Loading					
	Factor 1	Factor 2	Factor 3	Factor 4	Communality (H^2)	Uniqueness (U^2)
R1	−.16	.11	**.68**	.02	.44	.55
R2	−.15	.03	**.79**	.03	.55	.44
R3	−.04	.25	**.50**	−.08	.38	.62
R4	.14	−.17	**.70**	−.18	.45	.54
R5	.16	−.03	.11	.20	.13	.86
R7	.14	**.32**	**.26**	.04	.37	.62
R9	.11	**.67**	.12	−.02	.63	.36
R10	.12	**.59**	−.10	.19	.54	.46
R11	.19	.27	.02	.26	.36	.63
R12	−.11	.25	−.10	**.74**	.60	.39
R13	−.26	**.35**	−.05	**.67**	.51	.48
R14	.00	−.20	.06	**.62**	.36	.66
R15	.02	**.65**	.08	.11	.57	.42
R16	.24	−.14	.07	**.36**	.25	.74
R17	**.43**	**.49**	−.08	−.04	.56	.43
R18	**.67**	.02	−.09	.10	.49	.51
R19	**.64**	.04	.05	−.01	.46	.53
R20	**.40**	**.47**	.00	−.02	.57	.42
R21	**.64**	.15	−.02	−.03	.50	.49
R22	**.83**	.17	−.12	−.24	.57	.42
R23	**.50**	−.15	.13	.19	.38	.61

factors. R7, R17 and R20 cross-loaded on two factors, and the absolute difference between the factor loading was less than .15. All four factors equally include 4 question items each, totaling 16 items on the scale.

4.2. RQ2: to what extent do the developed items demonstrate reliability across different types of problem-solving tasks involving Generative AI?

We then conducted two separate CFAs with the remaining survey responses collected from the two problem-solving activities to assess whether the scale can be applied to different problem-solving scenarios. With the 3D Maze week data, the fit indices suggested adequate model fit with $X^2(98) = 450.59$, CFI = .89, which is close to .90, RMSEA = .06 < .08, and SRMR = .06 < .08. With the Policy

Table 3
The finalized scale of reliance behaviors during Human-AI problem-solving with descriptive results.

Categories	Cronbach's Alpha	MICs	M(SD)	Correlation				Questionnaire items
				1	2	3	4	
1. Reflective Use	.74	.41	4.01(.58)	–				R9: I designed my own prompts for Generative AI so it could give me the output I needed R10: I read Generative AI's output critically when the output was generated R15: I revised my prompting questions so Generative AI can give more appropriate outputs R7: I asked Generative AI to improve my initial draft/solutions for the problem-solving activities
2. Cautious Use	.65	.37	3.45(.69)	.40*	–			R12: I spotted errors in the Generative AI's output R13: I found Generative AI's output is not perfect R14: I discarded Generative AI's output R16: I finished a subtask in problem-solving activities with no or limited help from Generative AI
3. Thoughtless Use	.72	.30	3.63(.68)	.28*	.16*	–		R1: I copied the task descriptions or problem statements to Generative AI for help R2: I copied the Generative AI's output as part of our solution R3: For each activity I did in this class, I spent most of the time interacting with Generative AI R4: I adopted Generative AI's output without major changes
4. Collaborative Use	.76	.42	3.93(.61)	.53*	.37*	.31*	–	R18: My peer(s) helped me spot error in the Generative AI's output R19: My peer(s) helped me modify Generative AI's output R21: I asked my peer(s) questions during the problem-solving activities R22: My peers answered my questions and supported my understanding

*$p < .001$.

Drafting week data, the fit indices also suggest adequate model fit with $X^2(98) = 604.23$, CFI = .91 > .90, RMSEA = .06 < .08, SRMR = .06 < .08. As the results indicated adequate model fit, we kept all the items in each factor. Factor 1 is associated with questions concerning the revision of prompts or solutions for problem-solving. Thus, it is categorized as *reflective use* (Cronbach's alpha = .74). Factor 2 includes behaviors related to recognizing errors produced by Generative AI and reacting to them. Thus, it is named *cautious use* (Cronbach's alpha = .65). Although Cronbach's alpha for this factor is slightly below the conventional threshold of .70, the mean inter-item correlation (MIC) falls within the optimal range of .20–.40, indicating that the items are sufficiently homogeneous while still reflecting unique aspects of the construct (Piedmont, 2014). Furthermore, removing any item from Factor 2 does not improve the Cronbach's alpha. Thus, all items are retained, as the factor contributes critical conceptual and theoretical value despite its slightly lower reliability. Factor 3 includes behaviors such as copying and pasting and relying on Generative AI for problem-solving, leading us to identify it as *thoughtless use* (Cronbach's alpha = .72). Factor 4 encompasses behaviors related to peer communication and collaboration, so is designated as *collaborative use* (Cronbach's alpha = .76). The Cronbach's alpha of the overall scale is .84. Table 3 presents the finalized scale with Spearman correlation among each variable.

Moreover, to validate whether the questionnaire items accurately captured reliance patterns, we cross-referenced participants' self-reported responses with their actual interaction patterns in the chat histories. Participants scoring high on reflective use demonstrated autonomy in adapting Generative AI to their needs. For instance, one group critically evaluated GPT's output and provided follow-up prompts like, "I think you misunderstood my request," while another requested revisions to enhance a maze design: "Help me make the following maze more challenging." Fewer groups had high cautious use, and these groups typically identified AI errors or engaged minimally with the tool, suggesting they preferred independent problem-solving. Those scoring high on thoughtless use often copied example prompts from course materials directly into the chat, with some even not replacing placeholders in the example prompts. However, validating collaborative use from chat histories was more challenging, as such behaviors often involved offline peer interactions. This qualitative validation supports the scale's ability to measure reliance behaviors in human-AI problem-solving.

4.3. RQ3: how do undergraduate students' reliance behaviors on Generative AI differ across problem-solving tasks?

We then calculated the descriptive statistics, Spearman correlations, and Welch two-sample t-tests to explore undergraduate students' reliance behaviors across the two problem-solving activities. As shown in Table 4, reflective use was the highest among the four factors, whereas cautious use was the least frequent among students. Regarding correlations among these factors, a strong correlation between collaborative use and reflective use was found, which might suggest that reflective uses of Generative AI and peer collaborations might reinforce each other. Surprisingly, thoughtless use is moderately related to reflective use, suggesting that students who generally engage in reflective practices may also adopt some thoughtless behavior during human-AI collaboration.

The *t*-test results revealed significant differences across the two problem-solving activities for reflective, thoughtless, and collaborative uses. Specifically, reflective use was significantly higher in the 3D Maze activity ($t(2284.3) = -9.95$, $p < .001$; M = 4.01, SD = .59 vs. M = 3.77, SD = .67), as were thoughtless use ($t(2220.2) = -15.26$, $p < .001$; M = 3.61, SD = .69 vs. M = 3.15, SD = .82) and collaborative use ($t(2387.4) = -7.27$, $p < .001$; M = 3.92, SD = .62 vs. M = 3.74, SD = .67) than in the Policy Drafting activity. These findings highlight that reliance behaviors vary across different problem-solving tasks.

5. Discussion

The primary rationale for developing the Reliance Behaviors Scale is to uncover how undergraduate students use Generative AI during problem-solving activities and the extent to which they rely on it. With Generative AI becoming increasingly integrated into educational environments and its advanced computational capabilities aiding teams in solving complex issues, such a scale is urgently needed. Previous measures of reliance on AI predominantly depend on decision-making accuracy. In contrast, the current instrument focuses on behaviors during human-AI collaboration without directly linking them to problem-solving outcomes and allows for measuring reliance behaviors in different activities.

Through EFA and CFAs, we verified that the scale includes four factors, each representing a different type of reliance behavior in problem-solving activities: reflective use, cautious use, thoughtless use, and collaborative use. Each factor encompasses specific behaviors that illustrate how students engage with AI during problem-solving tasks. Reflective use includes behaviors where students critically engage with Generative AI. They think about how to frame their questions to get the desired output, critically read and revise AI-generated outputs, and use AI to enhance their initial drafts or solutions. This reflective use suggests a deep interaction with AI, focusing on improving and refining the problem-solving process. Reflective uses of Generative AI might be associated with reflective

Table 4
The difference in undergraduate students' reliance behaviors in the two problem-solving activities.

Categories	3D maze week	Policy Drafting week	*t*-test
	M(SD)	M(SD)	
1. Reflective Use	4.01(.59)	3.77(.67)	−9.95*
2. Cautious Use	3.46(.69)	3.46(.65)	.19
3. Thoughtless Use	3.61(.69)	3.15(.82)	−15.26*
4. Collaborative Use	3.92(.62)	3.74(.67)	−7.27*

*$p < .001$.

thinking, which involves self-regulation and continuous evaluation of one's strategies and outcomes (Gürol, 2011). Reflective thinking encourages individuals to think critically about their strategies, evaluate their progress, and take appropriate steps in problem-solving (Gencel & Saracaloglu, 2018). Reflective thinking is positively associated with cognitive presence and academic achievement (Kim & Lim, 2019). Similarly, engaging with AI in a reflective manner encourages students to control and regulate their problem-solving processes rigorously, which might foster deeper cognitive involvement and higher academic performance. Future research can examine the correlations between students' reflective use of Generative AI and their problem-solving performance and other types of academic performance.

Cautious use captures behaviors where students recognize and react to errors produced by Generative AI. They spot inaccuracies, acknowledge the imperfection of AI outputs, and sometimes discard AI suggestions. This cautious use indicates a healthy skepticism and a critical evaluation of AI contributions. AI hallucinations—instances where AI produces content that diverges from the input or contradicts established facts—are serious concerns when incorporating Generative AI in educational settings (Zhang et al., 2023). Research shows that LLMs frequently exhibit overconfidence with seemingly authoritative content (Kim et al., 2024, pp. 822–835). However, if students are aware of the hallucinations and overconfidence, they are more likely to recognize that even the most fluent and convincing responses can be incorrect. Thus, the cautious use of Generative AI can be seen as behavioral traces of users' awareness of hallucinations inherent in Generative AI. This awareness encourages a more vigilant approach, where students actively verify AI-generated information to avoid potential errors and maintain critical engagement with AI outputs.

Thoughtless use describes behaviors where students rely heavily on AI without much critical engagement. They might copy and paste task descriptions to Generative AI or AI-generated output directly into their solutions, indicating a passive reliance on AI. Previous research has frequently labeled behaviors that accept AI suggestions with no or limited considerations as overreliance. This passive dependence can inhibit students from conducting thorough research and developing their own insights (Kasneci et al., 2023). Li and Little (2023) argued that AI users, regardless of their subject matter expertise, are susceptible to overreliance. At the same time, relying on Generative AI for routine and basic tasks during problem-solving, such as checking grammar issues, or articulating ideas concisely and formally can be beneficial and efficient (Zhai et al., 2024). Hence, it is worth further empirical investigation to determine how the thoughtless use of Generative AI is associated with problem-solving performances and outcomes of other types of learning tasks.

Collaborative use involves behaviors related to peer communication and collaboration in conjunction with Generative AI use, such as students collaboratively spotting errors, modifying AI output, and supporting each other during problem-solving activities. This collaborative approach highlights the social dimension of learning with AI. Collaborative learning is common in higher education (Laal & Laal, 2012). Generative AI can facilitate this by acting as a learning partner, helping students to engage in meaningful dialogues, share diverse perspectives, and improve the collective knowledge of the group (Joksimovic et al., 2023). Students can use AI to explore ideas in their personal space, bring these insights into a shared learning environment, and collectively assess and refine these ideas (Tan et al., 2022). Thus, it is essential that students are capable of collaborating with their group mates, as well as Generative AI tools in the collaborative process. How students interact with their peers and Generative AI in the learning process to optimize the learning process, outcomes, and experiences is a complex issue and warrants further research.

Furthermore, we conducted two CFAs on data collected from two distinct problem-solving tasks. PBL involves diverse and authentic problem-solving activities based on specific learning objectives and encourages collaborative efforts (Wijnia et al., 2019). Thus, to accommodate various problem-solving tasks in undergraduate classrooms, we applied CFAs to two tasks: one focused on programming games and the other on policy drafting. Both analyses showed a good model fit, indicating the scale's potential applicability across various PBL contexts. Together, these categories and items provide a comprehensive measure of reliance behaviors, capturing the nature and frequency of students' interactions with Generative AI across tasks.

The Reliance Behavior Scale focuses on behaviors. We acknowledge that a particular behavior may be beneficial in some situations and potentially detrimental in others. As indicated by the descriptive analysis of reliance behaviors across the two problem-solving activities (Please see Table 4), it is evident that most of the correlations between the factors are moderate or strong, except between thoughtless use and cautious use, which are conceptually distinct. The results imply that people may demonstrate different reliance behaviors in different stages of problem-solving. For instance, reflective behaviors might be highly desirable in contexts requiring deep understanding and critical thinking, while in routine tasks, thoughtless use behaviors might not significantly impact the learning outcomes. By avoiding labeling each behavior as inherently good or bad, the scale provides a flexible tool for examining the diverse ways students use Generative AI across various stages of different learning tasks.

Furthermore, collaborative, reflective, and thoughtless uses are significantly higher in 3D Maze week compared to Policy Drafting week, which highlights the nuanced, context-dependent nature of reliance behaviors. The difference in reliance behavior between Policy Drafting week and 3D Maze week might be explained by cognitive and emotional factors, such as students' confidence in problem-solving, prior exposure to AI tools, perceived ease of use, or attitudes toward collaboration (Van Dongen & Van Maanen, 2013). While this was beyond the scope of the current study, it might be promising to examine how reliance behaviors vary across tasks requiring different levels of cognitive load, and other related cognitive or emotional factors.

5.1. Limitations

Several limitations of this study need to be acknowledged. First, the scale was developed and validated with a relatively homogeneous sample of undergraduate students from a single Asian university, which may limit its generalizability. Cultural and environmental factors could influence how students perceive and use Generative AI, potentially impacting reliance behaviors. For instance, cultures with higher levels of technological adaptation may foster greater familiarity and comfort with AI, while economically

disadvantaged regions may exhibit lower AI literacy due to limited access to technology and resources (Pinski & Benlian, 2024). Furthermore, the current study's findings may not be directly generalized to K–12 education. Generative AI tools might not yet be widely adopted among younger students due to ethical concerns and limited access. However, adapting the scale to K–12 contexts could offer insights into the reliance behaviors of students at different educational levels. Future research should include a more diverse sample, encompassing students from different educational levels, cultural backgrounds, and learning environments, to broaden the applicability of the findings.

Second, the scale focuses specifically on behaviors related to Generative AI tools, such as those that generate text, images, or code. While Generative AI is a rapidly expanding field, the scale does not account for reliance behaviors with other types of AI, such as intelligent tutoring systems, recommendation systems, or adaptive learning technologies. This limits the scale's applicability across all AI tools. Further research should expand the scale to include behaviors associated with other AI technologies to provide a more comprehensive understanding of human-AI collaboration.

Third, a notable limitation of this study is that, after removing some items based on EFA results, certain categories of the problem-solving stages were represented by a single question in the questionnaire. For instance, AI reliance for the problem identification stage only includes "I copied the task descriptions or problem statements to Generative AI for help", and peer reliance for the information exploration stage was limited to the single question "My peers answered my questions and supported my understanding". This reduction of questionnaire items may insufficiently capture the complexity of these problem-solving stages. At the same time, while the current questionnaire yielded a good overall Cronbach's alpha, one of the dimensions, Cautious Use, is slightly below the commonly used benchmark ($.65 < .70$). Thus, further refinement of the survey items is necessary. Future research should expand the number of items in each subcategory to provide more comprehensive coverage of these dimensions and improve the instrument's ability to capture nuanced students' behaviors using Generative AI while collaborating with peers for problem-solving tasks.

Finally, while we had conducted a qualitative validation to compare self-reported behaviors with behaviors observable in the chat histories, such validation was limited to high-scoring groups and did not systematically address the full range of self-reported responses. Thus, future studies could adopt a more comprehensive approach, integrating purposeful sampling and analyzing a broader range of responses to further validate and refine the scale.

5.2. Implications

Developing the Reliance Behaviors Scale has several important implications for subsequent research in human-AI collaboration and educational practice. It enables educators to measure students' reliance on Generative AI tools unobtrusively after problem-solving activities, providing authentic data on how students use AI and can help identify potential issues in students' use behaviors, such as high levels of thoughtless use. Also, educators can leverage the scale results to encourage critical thinking and productive reliance behaviors in problem-based learning contexts. For example, instructors could use insights from the scale to adjust teaching methods by tailoring problem-solving tasks that encourage balanced and critical reliance on both AI and human collaboration. Additionally, the scale can help researchers link specific reliance behaviors to learning outcomes and 21st-century skills, such as problem-solving performance, critical thinking and reflective thinking skills, and communication among students and AI. This information can guide the design of AI tools and instruction that encourage productive use behaviors while reducing overreliance.

CRediT authorship contribution statement

Chenyu Hou: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gaoxia Zhu:** Writing – review & editing, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Vidya Sudarshan:** Writing – review & editing, Resources, Project administration, Investigation. **Fun Siong Lim:** Writing – review & editing, Resources. **Yew Soon Ong:** Writing – review & editing, Resources, Project administration.

Availability of data and materials

Because of confidentiality agreements and ethical concerns, the data used in this study will not be made public. These data will be made available to other researchers on a case-by-case basis.

Funding

This project is supported by Ministry of Education (MOE) Academic Research Fund Tier 1 (RG133/24) and Tier 2 (MOE-T2EP40124-0004).

Declaration of competing interest

The authors declare no potential conflict of interest in the work.

Acknowledgments

The authors are indebted to the students who participated in this study. We also want to express our sincere gratitude towards Dr. Shen Yong Ho for all his back-end administrative support, which made the study possible, as well as Drs. Kin Yew Low, Lay Poh Tan and all instructors for their support in implementing this study.

Appendix

Appendix A

Problem-solving Activity Requirement and Sample Prompts for Students.

		Requirement	Sample Prompts for Generative AI
3D Maze week	Part 1: Maze Creation	Your group will design a maze in 3D Maze that meets the following design requirements (also stated in the Lesson Overview above): • The width and the length of the maze should be at least 10 units. • The maze should contain all environmental assets (green tree, cherry blossom, rock, block A, block B, river, bridge) • Try your best to make the maze challenging but engaging. Optional: the maze can contain two gems (gems need to be collectible in maze solving)	*I have created a 10 x10 solvable maze. it includes the following elements:* • *1 x starting point represented by "S"* • *1 x end point represented by "E"* • *Walkable paths represented by "0"* • *Obstacles represented by "1"* • *Monsters represented by "M"* • *Gem represented by "G"* • *Heart that must be collected represented by "H"* *The following is the maze that I created in a table format. You are an expert on maze design. Can you redesign the following maze to be more challenging by including more elements, and have the path difficult to navigate?*
	Part 2: Maze Solving	Your group will solve the maze you created. You can decide how and when you want to use Gen AI, e.g., to generate the pseudo-codes or correct your pseudo-codes. Again you can experiment with different Gen AI tools and decide which one to use. Here is an example prompt that you can use to help you solve the maze.	*In this programming game, I need to solve the maze using pseudo-code to learn programming language. You are a programming teacher. I will tell you in natural language how to navigate the maze, and you need to present me with the correct pseudo codes to solve to maze.* *The pseudo-codes that are available:* *Five actions:* 1. *Turn left.* 2. *Turn right.* 3. *Move forward.* 4. *Turn back* 5. *attack* *Four conditions:* 1. *can move "forward/backward/left/right"* 2. *not reach destination* 3. *repeat < #* 4. *Enemy in front* *Three action blocks:* 1. *while loop* 2. *if else loop* 3. *if loop* *Do you understand?*
Policy Drafting week	–	The university has released its position on using Gen AI on campus in 2023. After a semester's practice, the Teaching, Learning, and Pedagogy Division at the university is having an open call for suggestions about the fair and safe use of Generative AI (Gen AI) during undergraduate learning activities. The Division is concerned with Gen AI misuse such as students' excessive use of and over-reliance on AI. They aspire to foster students' critical thinking, creativity, and agency while allowing them to explore latest AI innovations. As students, the focus of teaching and learning activities, your observations and experiences of how Gen AI can be (mis)used, and your suggestions on how to minimize such misuse across campus are critical. You and your teammates are undergraduate representatives	*"The team is working on a problem-solving activity. They need to give one suggestion to the university about how to use Generative AI in their assignments while minimizing misuse. You are an instructor at the university. You need to ask 3 Socratic questions to help them think deeply and evaluate their plan critically. Socratic questions are defined as questions that help learners distinguish things they know from things they don't understand. Some sample Socratic questions are "What is the central aim of the task in this line of thought?"; "How can people from a different viewpoint react to the response?"; and "On what information are you basing that comment?*\\ *The team's preliminary suggestion is:* \\\"

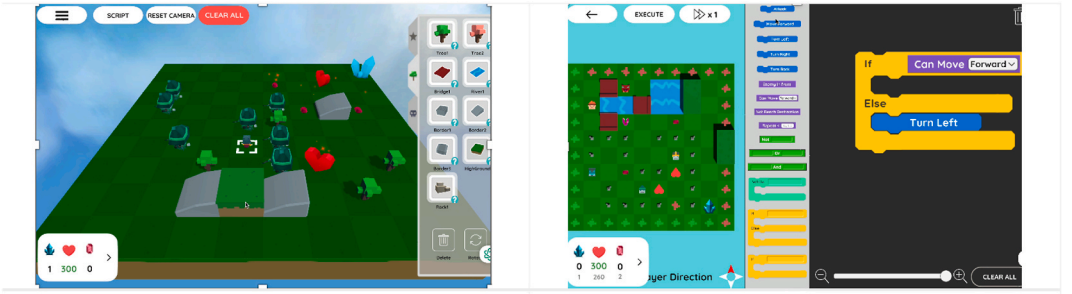
(*continued on next page*)

(*continued*)

	Requirement	Sample Prompts for Generative AI
	elected from different schools to **provide specific suggestions for the stakeholders (e.g., students themselves, the university, and instructors) to promote ethical use and minimize misuse of Gen AI on campus**. When considering the use of Gen AI among undergraduate students, you can think about the learning activities you usually have in your respective schools. When your group has developed the suggestion, you should also critically evaluate the suggestion from the perspectives of your respective schools.	

Appendix B

An Illustration of the 3D Maze Platform.



(a) Maze Creation (b) Maze Solving

Data availability

Data will be made available on request.

References

Al-kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., & Alfandi, O. (2024). Ethical challenges and solutions of generative AI: An interdisciplinary perspective. In *Informatics, 11* p. 58). MDPI. No. 3.

Azmi, A. D., Wahab, S., Abdul Kadir, N. B. Y., Abdul Wahab, N. A., & Razak, R. A. (2024). The development and validation of a knowledge, attitude, and practice questionnaire of methamphetamine use. *Scientific Reports, 14*(1), Article 21835.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing, 7* pp. 2–11). https://doi.org/10.1609/hcomp.v7i1.5285

Barrows, H. S. (1996). Problem-based learning in medicine and beyond: A brief overview. *New directions for teaching and learning, 1996*(68), 3–12.

Behfar, K. J., Peterson, R. S., Mannix, E. A., & Trochim, W. M. K. (2008). The critical role of conflict resolution in teams: A close look at the links between conflict type, conflict management strategies, and team outcomes. *Journal of Applied Psychology, 93*(1), 170–188. https://doi.org/10.1037/0021-9010.93.1.170

Braga, A., & Logan, R. K. (2017). The emperor of strong AI has No clothes: Limits to artificial intelligence. *Information, 8*(4). https://doi.org/10.3390/info8040156. Article 4.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 188:1–188:21. https://doi.org/10.1145/3449287

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers & Education: Artificial Intelligence, 4*, Article 100118. https://doi.org/10.1016/j.caeai.2022.100118

Cudeck, R., & O'Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin, 115*(3), 475–487. https://doi.org/10.1037/0033-2909.115.3.475

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering, 61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272.

Fok, R., & Weld, D. S. (2024). In Search of verifiability: Explanations rarely enable complementary Performance in AI-advised decision making. *(arXiv:2305.07722). arXiv*. https://doi.org/10.48550/arXiv.2305.07722

Gencel, I. E., & Saracaloglu, A. S. (2018). The effect of layered curriculum on reflective thinking and on self-directed learning readiness of prospective teachers. *International Journal of Progressive Education, 14*(1), 8–20.

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. https://doi.org/10.3390/educsci13070692

Guo, Z., Wu, Y., Hartline, J. D., & Hullman, J. (2024). A decision theoretic framework for measuring AI reliance. *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. https://doi.org/10.1145/3630106.3658901

Gürol, A. (2011). Determining the reflective thinking skills of pre-service teachers in learning and teaching process. *ENERGY EDUCATION SCIENCE AND TECHNOLOGY PART B-SOCIAL AND EDUCATIONAL STUDIES, 3*(3). https://avesis.yildiz.edu.tr/yayin/095713e8-6b58-4988-8d7d-a29683a9a542/determining-the-reflective-thinking-skills-of-pre-service-teachers-in-learning-and-teaching-process.

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1112–1123).

Hamid, H., Zulkifli, K., Naimat, F., Che Yaacob, N. L., & Ng, K. W. (2023). Exploratory study on student perception on the use of chat AI in process-driven problem-based learning. *Currents in Pharmacy Teaching and Learning, 15*(12), 1017–1025. https://doi.org/10.1016/j.cptl.2023.10.001

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393–416. https://doi.org/10.1177/0013164405282485

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In *Assessment and teaching of 21st century skills* (pp. 37–56). Dordrecht: Springer. https://doi.org/10.1007/978-94-017-9395-7_2.

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity, 44*(1), 153–166. https://doi.org/10.1007/s11135-008-9190-y

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Joksimovic, S., Ifenthaler, D., Marrone, R., De Laat, M., & Siemens, G. (2023). Opportunities of artificial intelligence for supporting complex problem-solving: Findings from a scoping review. *Computers & Education: Artificial Intelligence, 4*, Article 100138. https://doi.org/10.1016/j.caeai.2023.100138

Kadaruddin, K. (2023). Empowering education through Generative AI: Innovative instructional strategies for tomorrow's learners. *International Journal of Business, Law, and Education, 4*(2), 618–625.

Kam, C. C. S. (2019). Careless responding threatens factorial analytic results and construct validity of personality measure. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01258

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, S. S. Y., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Vaughan, J. W. (2024). "I'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. https://doi.org/10.1145/3630106.3658941

Kim, J. Y., & Lim, K. Y. (2019). Promoting learning in online, ill-structured problem solving: The effects of scaffolding type and metacognition level. *Computers & Education, 138*, 116–129. https://doi.org/10.1016/j.compedu.2019.05.001

Kim, A., Yang, M., & Zhang, J. (2023). When algorithms err: Differential impact of early vs. Late errors on users' reliance on algorithms. *ACM Transactions on Computer-Human Interaction, 30*(1), 1–36.

Kline, T. J. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage publications.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by *grammarly*: A multiple case study. *Assessing Writing, 44*, Article 100450. https://doi.org/10.1016/j.asw.2020.100450

Koschmann, T. D. (1994). Toward a theory of computer support for collaborative learning. *The Journal of the Learning Sciences, 3*(3), 219–225. https://doi.org/10.1207/s15327809jls0303_1

Kupffer, R., Frick, S., & Wetzel, E. (2024). Detecting careless responding in multidimensional forced-choice questionnaires. *Educational and Psychological Measurement*. , Article 00131644231222420. https://doi.org/10.1177/00131644231222420

Laal, M., & Laal, M. (2012). Collaborative learning: What is it? *Procedia - Social and Behavioral Sciences, 31*, 491–495. https://doi.org/10.1016/j.sbspro.2011.12.092

Larson Jr, J. R., & Christensen, C. (1993). Groups as problem-solving units: Toward a new meaning of social cognition. *British Journal of Social Psychology, 32*(1), 5–30. https://doi.org/10.1111/j.2044-8309.1993.tb00983.x

Li, M. D., & Little, B. P. (2023). Appropriate reliance on artificial intelligence in radiology education. *Journal of the American College of Radiology, 20*(11), 1126–1130. https://doi.org/10.1016/j.jacr.2023.04.019

Lin, C. J., Lee, H. Y., Wang, W. S., Huang, Y. M., & Wu, T. T. (2024). Enhancing reflective thinking in STEM education through experiential learning: The role of generative AI as a learning aid. *Education and Information Technologies*, 1–23.

Liu, Y., & Pásztor, A. (2022). Effects of problem-based learning instructional intervention on critical thinking in higher education: A meta-analysis. *Thinking Skills and Creativity, 45*, Article 101069. https://doi.org/10.1016/j.tsc.2022.101069

Malik, T., Hughes, L., Dwivedi, Y. K., & Dettmer, S. (2023). Exploring the transformative impact of generative AI on higher education. In *Conference on e-Business, e-Services and e-Society* (pp. 69–77). Cham: Springer Nature Switzerland.

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*(3), 275.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437.

Naik, A., Yin, J. R., Kamath, A., Ma, Q., Wu, S. T., Murray, C., … Rose, C. P. (2024). Generating situated reflection triggers about alternative solution paths: A case study of generative AI for computer-supported collaborative learning. In *International conference on artificial intelligence in education* (pp. 46–59). Cham: Springer Nature Switzerland.

Passi, S., & Vorvoreanu, M. (2022). Overreliance on AI literature review. *Microsoft Research*.

Piedmont, R. L. (2014). Inter-item correlations. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 3303–3304). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-0753-5_1493.

Pinski, M., & Benlian, A. (2024). AI literacy for users–A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans*, Article 100062.

Ross, L., Lundstrøm, L. H., Petersen, M. A., Johnsen, A. T., Watt, T., & Groenvold, M. (2012). Using method triangulation to validate a new instrument (CPWQ-com) assessing cancer patients' satisfaction with communication. *Cancer Epidemiology, 36*(1), 29–35.

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *(arXiv:2204.06916). arXiv*. http://arxiv.org/abs/2204.06916.

Schmidt, H. G. (1993). Foundations of problem-based learning: Some explanatory notes. *Medical Education, 27*(5), 422–432. https://doi.org/10.1111/j.1365-2923.1993.tb00296.x

Schmidt, H. G. (2012). A brief history of problem-based learning. In G. O'Grady, E. H. J. Yew, K. P. L. Goh, & H. G. Schmidt (Eds.), *One-day, one-problem: An approach to problem-based learning* (pp. 21–40). Springer. https://doi.org/10.1007/978-981-4021-75-3_2.

Schmidt, H. G., Van Der Molen, H. T., Te Winkel, W. W. R., & Wijen, W. H. F. W. (2009). Constructivist, problem-based learning does work: A meta-analysis of curricular comparisons involving a single medical school. *Educational Psychologist, 44*(4), 227–249. https://doi.org/10.1080/00461520903213592

Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement, 71*(1), 95–113. https://doi.org/10.1177/0013164410387348

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management, 57*(2), Article 103174. https://doi.org/10.1016/j.im.2019.103174

Strobel, J., & Barneveld, A. van (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. *Interdisciplinary Journal of Problem-Based Learning, 3*(1). https://doi.org/10.7771/1541-5015.1046. Article 1.

Sun, D., Xu, P., Zhang, J., Liu, R., & Zhang, J. (2025). How self-regulated learning Is affected by feedback based on large language models: Data-driven sustainable development in computer programming learning. *Electronics, 14*(1), 194.

Tahiru, F. (2021). AI in education: A systematic literature review. *Journal of Cases on Information Technology, 23*(1), 1–20. https://doi.org/10.4018/JCIT.2021010101

Tan, S. C., Lee, A. V. Y., & Lee, M. (2022). A systematic review of artificial intelligence techniques for collaborative learning over the past two decades. *Computers & Education: Artificial Intelligence, 3*, Article 100097. https://doi.org/10.1016/j.caeai.2022.100097

Torrance, H. (2012). Triangulation, respondent validation, and democratic participation in mixed methods research. *Journal of Mixed Methods Research, 6*(2), 111–123.

Valdivia, A. E. O., Osorio, C. M., Velasco-Bejarano, B., & Vargas-Rodríguez, Y. M. (2024). Evaluating cognitive and affective development in chemistry students using AI-supported problem-based learning. *American Journal of Educational Research, 12*(11), 447–454.

Van Dongen, K., & Van Maanen, P. P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies, 71*(4), 410–424.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 129:1–129:38. https://doi.org/10.1145/3579605

Wang, L., Jamieson, G. A., & Hollands, J. G. (2008). Selecting methods for the analysis of reliance on automation. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting, 52*(4), 287–291. https://doi.org/10.1177/154193120805200419

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., & Metzler, D. (2022). Emergent abilities of large language models. *arXiv Preprint arXiv:2206.07682*.

Wijnia, L., Loyens, S. M. M., & Rikers, R. M. J. P. (2019). The problem-based learning process. In *The wiley handbook of problem-based learning* (pp. 273–295). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119173243.ch12.

Zhai, C., Wibowo, S., & 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. https://doi.org/10.1186/s40561-024-00316-7

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv.Org*. https://arxiv.org/abs/2309.01219v2.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). A survey of large language models. *(arXiv:2303.18223). arXiv*. http://arxiv.org/abs/2303.18223.

Zhu, G., Sudarshan, V., Kow, J. F., & Ong, Y. S. (2024, June). *Human-Generative AI Collaborative Problem Solving Who Leads and How Students Perceive the Interactions* (pp. 680–686). IEEE.