ActiveAI: The Effectiveness of an Interactive Tutoring System in Developing K-12 AI Literacy

Ying-Jui Tseng[1](✉), Gautam Yadav[1], Xinying Hou[2], Muzhe Wu[1], Yun-Shuo Chou[1], Claire Che Chen[1], Chia-Chia Wu[3], Shi-Gang Chen[3], Yi-Jo Lin[3], Guanze Liao[3], and Kenneth R. Koedinger[1]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
{yingjuit,gyadav,muzhew,yunshuoc,clairechen,koedinger}@andrew.cmu.edu
[2] University of Michigan, Ann Arbor, MI, USA
xyhou@umich.edu
[3] National Tsing Hua University, Hsinchu, Taiwan
{candacewu,shigangc,yijolin,gzliao}@mx.nthu.edu.tw

Abstract. As we witness groundbreaking advancements in Artificial Intelligence (AI), it is clear that the next generation must be equipped with AI literacy: the skill to interact, evaluate, and collaborate with AI systems. This study introduces ActiveAI, a scalable web-based tutoring system aligned with AI4K12's five big ideas in AI, designed to foster AI literacy among K-12 students through active learning and interaction with intelligent agents. A controlled classroom study involving 171 middle school learners was conducted to assess the effectiveness of ActiveAI in fostering AI literacy skills and competency toward AI. Results showed that, compared to students in the tell-and-practice control condition, students who used ActiveAI exhibited higher post-test performance in the module about how next-word prediction and temperature work in large language models. Students also developed higher self-reported competence toward AI after using ActiveAI than in the control condition. We conclude by suggesting assessment designs that promote deeper engagement with AI concepts by addressing students' common misconceptions, like "AI thinks just like humans", in K-12 AI literacy education.

Keywords: AI Literacy · Intelligent Agents · Experiential Learning · K-12 · Instructional Design · AI education · Classroom Implementation

1 Introduction

The advancements in Artificial Intelligence (AI), largely due to the popularity of generative AI, have made the need for AI literacy across the educational spectrum more crucial than ever. As AI becomes increasingly integral to a wide array

R. Ferreira Mello et al. (Eds.): EC-TEL 2024, LNCS 15159, pp. 452–467, 2024.
https://doi.org/10.1007/978-3-031-72315-5_31

of applications, it is essential for K-12 students to understand how to interact with these technologies and recognize their capabilities and limitations [7,20]. Initiatives in various parts of the world have aimed to integrate AI education into the K-12 curriculum. For example, Song et al. (2023) developed a comprehensive AI syllabus for compulsory education, showing positive outcomes in primary schools in China [24]. Also, Touretzky et al. (2022) introduced a flexible, nine-week elective AI course in middle schools across Georgia, USA, effectively engaging students with AI's technical aspects [29]. While these initiatives have been groundbreaking in introducing AI to the younger generation of students, they rely on various tools, platforms, and passive instructional material that necessitate significant investments in teacher training and resources. Given the imminent ubiquity of AI in future professional and everyday life scenarios, there is a pressing need for a more accessible and scalable approach to AI literacy education for K-12 learners.

To address this, we designed and developed ActiveAI, a scalable interactive tutoring system that is guided by the five big ideas of the AI4K12 initiative: Perception, Representation & Reasoning, Learning, Natural Interaction, and Societal Impact [30]. These ideas cover how AI systems perceive the world, represent and reason with data, learn from experiences, interact naturally with humans, and impact society. By doing so, we aim to make AI literacy education more accessible and adaptable as well as reduce the burden on teachers to learn multiple, disparate AI Literacy strategies.

In ActiveAI, each module incorporates active learning through intelligent agents and assessments that aim to transfer student learning to the real world. By directly interacting with AI in the tutoring system, students can gain a deep understanding of AI capabilities, including its inherent randomness and unpredictability, without the need for direct programming. In addition, ActiveAI emphasizes experiential learning, encouraging students to actively engage in constructing AI solutions. It further enhances learning with immediate, targeted feedback and on-demand hints, designed to provide an effective and engaging AI literacy learning experience. We also conducted a controlled classroom study with 171 middle school learners to evaluate its effectiveness in fostering students' AI literacy. Our research questions are as follows:

1. Do middle school students achieve higher AI literacy learning outcomes from engaging with ActiveAI than from the tell-and-practice control condition?
2. Do middle school students show higher competence in using AI after engaging with ActiveAI than from the tell-and-practice control condition?
3. Do middle school students show higher levels of engagement when interacting with ActiveAI compared to the tell-and-practice control condition?

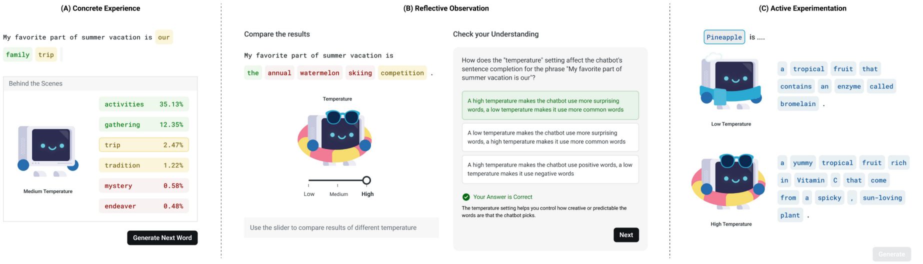
2 Background

Active learning [3] and experiential learning [20] methods are essential for K-12 AI literacy education, effectively engaging learners with new AI concepts [8,23]. For instance, active learning has been applied to teach complex topics like generative adversarial networks and supervised machine learning models in middle school settings [1,9,11]. Experiential learning, which involves hands-on experiences and reflection [14], is also vital. They help students link AI theories to real-world applications through unplugged activities, programming, and intelligent agents.

Unplugged activities teach AI concepts without computers, often through role-playing, simulations, and physical manipulatives, allowing educators to introduce complex AI concepts in an engaging way [1,2,9]. An example is using hand-drawn decision trees in an after-school program to teach about decision algorithms [16]. Programming activities enable students to build AI algorithms using platforms like Colab for teaching Python programming and ML/NLP algorithms. This approach uses constructivist strategies for hands-on learning and real-world application [4,17]. Block-based programming environments like Cognimates are also used, suitable for younger learners [6]. Narrative scripts can be used to educate adolescents about social media algorithms' nuances, thereby cultivating a more informed and critical perspective toward AI technologies [27]. Intelligent agents, such as expert systems and machine learning trainers, are increasingly used for hands-on AI model building and interaction without coding. For example, Google Teachable Machine has been used by sixth graders for creating machine learning applications [28], and tools like ArtBot and VotestratesML engage older students in understanding AI concepts and societal implications [12,32].

Although these pedagogies effectively foster AI literacy in K-12 students, they often require specialized training. Each strategy focuses on specific modules or narrow aspects of AI literacy, demanding frequent educator training to integrate these methodologies into curricula. This aspect, while not diminishing the invaluable contributions of these pedagogies, highlights a challenge in broadly applying these methods across diverse AI concepts.

A scalable way to address these challenges is the use of interactive tutoring systems. Research has shown that intelligent tutoring systems (ITS) are efficient in both well-defined domains like mathematics [19] and programming [5], and ill-defined domains [21]. There is also preliminary work in applying these systems to AI literacy for adult learners [31]. These systems can provide personalized, scalable learning experiences and reduce the need for extensive teacher training by offering direct instructional support to students. Our current research aims to fill the gap by integrating these advanced tutoring systems into K-12 AI literacy education, thereby making AI learning more accessible and scalable.



My favorite part of summer vacation is our family trip

activities	35.13%
gathering	12.35%
trip	2.47%
tradition	1.22%
mystery	0.58%
endeaver	0.48%

My favorite part of summer vacation is the annual watermelon skiing competition .

Use the slider to compare results of different temperature

How does the "temperature" setting affect the chatbot's sentence completion for the phrase "My favorite part of summer vacation is our"?

- A high temperature makes the chatbot use more surprising words, a low temperature makes it use more common words
- A low temperature makes the chatbot use more surprising words, a high temperature makes it use more common words
- A high temperature makes the chatbot use positive words, a low temperature makes it use negative words

✓ Your Answer is Correct

The temperature setting helps you control how creative or predictable the words are that the chatbot picks.

Pineapple is ....

a tropical fruit that contains an enzyme called bromelain .

a yummy tropical fruit rich in Vitamin C that come from a spicky , sun-loving plant .

Fig. 1. A)Concrete Experience B)Reflective Observation C)Active Experimentation.

3 ActiveAI: An Interactive Tutoring System for K-12 AI Literacy

3.1 Instruction Design

The ActiveAI system, designed for teaching the five big ideas of AI4K12 to 7–9 grade students, uses an active learning approach integrating proven learning science mechanisms.

Embedded Intelligent Agents. Each module includes intelligent agents (Fig. 2C) for engaging with AI algorithms without programming. These agents highlight AI randomness and uncertainty while addressing ethical concerns in student-AI interactions. Formative assessments provide feedback during learning.

Experiential Learning. The ActiveAI modules embody experiential learning through four phases: *Concrete Experience, Reflective Observation, Abstract Conceptualization*, and *Active Experimentation* [14]. An example is the *How Temperature Shapes the Response of LLM* module in Fig. 1.

- **Concrete Experience:** Students start by observing a chatbot with a medium temperature setting completing a sentence. They then compare this with the chatbot's responses at high and low-temperature settings. Typically, a low temperature leads to more predictable results, while a high temperature results in more surprising outcomes. Probability of each word within the sentence's context is highlighted using color coding (Fig. 1A).
- **Reflective Observation**: Students are prompted to consider how the "temperature" setting affects the chatbot's sentence completion. They use a slider to experiment with different temperature settings, observing the varying outcomes. This activity is designed to help students discern the relationship between temperature and chatbot output, reducing the cognitive load of memorizing details (Fig. 1B). The expected realization is that a high temperature setting leads to more surprising word choices, whereas a low temperature favors common words.

- **Abstract Conceptualization**: Following this, students apply their newfound understanding by selecting the most appropriate temperature settings for different tasks. For instance, when tasked with scripting a sci-fi movie that requires intriguing dialogues and unexpected plot twists, they should recognize that a high-temperature setting is more suitable.
- **Active Experimentation**: Finally, students apply their knowledge to several practical scenarios they might encounter in daily life. For instance, They might be asked to "gather facts for a biology project", choosing a topic in biology and selecting a temperature setting for the task. An intelligent agent, powered by GPT-3.5, generates content in both high and low-temperature settings based on the student's input in real-time (Fig. 1C). If an inappropriate temperature setting is chosen, targeted feedback is provided to address any gaps in understanding.

3.2 System Design & Implementation

ActiveAI's educational content is scalable, built with React.js, and managed in JSON format for easy updates. The design incorporates specific learner input modalities to optimize AI literacy education by reducing cognitive load and focusing on core AI concepts.

We selected three learner input modalities to interact with intelligent agents: *sliders*, *steppers*, and *collectors*. These reduce the extraneous cognitive load associated with learning new interactions [18] and enhance learning efficiency, supporting scalable module development.

- **Collector:** Enables students to gather or label datasets, fostering active participation in learning and a deeper understanding of data's role in AI [22].
- **Slider:** Allows students to adjust variables like classification model thresholds, providing immediate feedback and facilitating adjustments in real-time to grasp the fluid nature of AI algorithms [10].
- **Stepper:** Controls time-related variables or offers step-by-step explanations, reducing cognitive overload through the segmentation principle of multimedia learning [18].

In addition to these interactive elements, we also integrated features based on principles of intelligent tutor design consistent with best practices across all learning activities within the modules:

- **On-demand Hint:** Offers targeted help when students struggle to foster independence and deeper understanding, as per scaffolding approach [35].
- **Step-based Feedback:** Provides immediate responses to reinforce learning and address misunderstandings, as per the feedback intervention theory [13].
- **Progress Bar:** Visualizes students' advancement, promoting a sense of achievement and motivation, as per self-regulated learning research [36].

4 Classroom Evaluation Study

4.1 Experimental Design

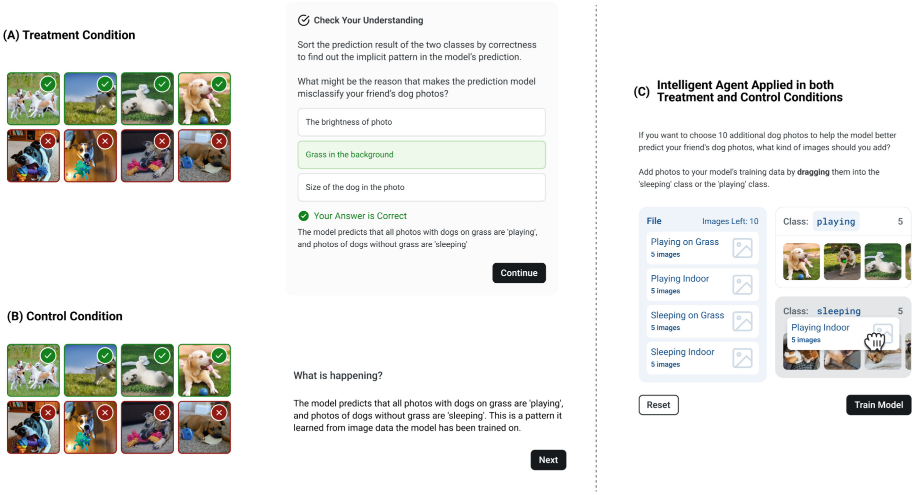
The primary aim of this study was to evaluate the effectiveness of ActiveAI in enhancing students' AI literacy. Specifically, we sought to understand if ActiveAI leads to greater post-test performances, increased competence in using AI technology, and a more engaging learning experience compared to the conventional 'tell and practice' approach often employed in AI education. To achieve this, we conducted a classroom study in middle school, focusing on two AI literacy modules:

- **Module 1** Next Word Prediction & Temperature in Large Language Models
  - 1.1. Articulate how large language models predict the next word in a sequence
  - 1.2. Analyze the role of temperature in chatbot decision-making to predict its influence on sentence completion
- **Module 2** Image Classification: Bias in Training Data & How to Reduce It
  - 2.1. Determine features that predict labels from patterns in labeled data
  - 2.2. Explain how the choice of training data shapes behavior of the classifier, and how bias can be introduced if the training set is not properly balanced

Within each module, we incorporated a structured approach consisting of a pre-test, the intervention followed up with a survey, and a post-test (Table 1). Due to time limitations, we only applied the 12-question survey to measure student competence in using AI [33] once after the first learning objective to compare treatment and control conditions. Additionally, a question regarding engagement was asked after finishing each module.

The control condition we used in this study mirrors the existing common practices in K-12 AI literacy studies involving experiential learning in the form of tell and practice [26], reading about and interacting with AI intelligent agents. The same instructional text and intelligent agent used in the ActiveAI modules were delivered to students. The only difference was that, in the control condition, there were no formative assessments or explanatory feedback (Fig. 2).

Our experimental design, shown in Table 1, counterbalances the instructional conditions across the learning objectives, effectively controlling for test difficulty and eliminating potential bias. This approach guarantees that any observed differences in outcomes can be attributed more confidently to the instructional methods employed than to external factors. Specifically, we developed two test forms for each learning objective (L1a, L1b, and L2a, L2b) containing isomorphic questions to each other. These forms were carefully crafted in collaboration with an AI expert, ensuring that both possessed equivalent difficulty and content coverage. By reversing the order of the tests for each subgroup, our design also mitigates the effects of test difficulty, enabling a fair comparison between the two test forms.


(A) Treatment Condition
Check Your Understanding
(B) Control Condition
(C) Intelligent Agent Applied in both Treatment and Control Conditions

Sort the prediction result of the two classes by correctness to find out the implicit pattern in the model's prediction.

What might be the reason that makes the prediction model misclassify your friend's dog photos?

- The brightness of photo
- Grass in the background
- Size of the dog in the photo

Your Answer is Correct

The model predicts that all photos with dogs on grass are 'playing', and photos of dogs without grass are 'sleeping'

What is happening?

The model predicts that all photos with dogs on grass are 'playing', and photos of dogs without grass are 'sleeping'. This is a pattern it learned from image data the model has been trained on.

If you want to choose 10 additional dog photos to help the model better predict your friend's dog photos, what kind of images should you add?

Add photos to your model's training data by **dragging** them into the 'sleeping' class or the 'playing' class.

Fig. 2. The treatment condition (A) includes formative assessments with hints and feedback, while the control condition (B) does not. Both conditions use the same intelligent agent interactions shown in (C).

Table 1. Overview of Experimental Conditions and Testing Sequence

Subgroup	Pre-test	Instruction	Survey & Post-test	Pre-test	Instruction	Survey & Post-test
1	L1a	L1 Experiment	Competence + Engagement + L1b	L2a	L2 Control	Engagement + L2b
2	L1b	L1 Experiment	Competence + Engagement + L1a	L2b	L2 Control	Engagement + L2a
3	L1a	L1 Control	Competence + Engagement + L1b	L2a	L2 Experiment	Engagement + L2b
4	L1b	L1 Control	Competence + Engagement + L1a	L2b	L2 Experiment	Engagement + L2a

4.2 Participants

171 Taiwanese eighth-grade students, aged 14–15, were recruited to participate this study. The age distribution was carefully balanced across all experimental conditions. Students were randomly assigned to different subgroups. Each learner experienced either the ActiveAI module or the traditional 'tell and practice' condition in one lesson, and then the alternate condition in the subsequent lesson. In the data analysis phase, we focused on students who completed both pre- and post-tests for Learning Objectives 1 (L1) and 2 (L2). This resulted in a final sample of 115 students for L1 and 99 students for L2. After the study, another 8 students, different from the 171 students who took part in the study, were selected by the teachers for the think-aloud and the interview.

4.3 Procedure

The study was conducted during students' scheduled digital literacy classes. Each participant had access to a desktop computer and the Internet. Students were randomly assigned to one of four conditions (Table 1) by their teacher. Pre- and post-tests were administered via Google Forms, and the instructional content for all conditions was delivered through the ActiveAI platform. Participants spent approximately 5 min completing a pretest to assess their existing AI knowledge, followed by 15 min of learning with an instructional module. Afterward, they completed a post-test and surveys to evaluate their learning outcomes and experiences. This procedure was repeated for the subsequent learning objective, totaling 50 min of study time.

We also conducted think-aloud interviews with another 8 students to interact with ActiveAI. Eight participants engaged in the one-on-one sessions with researchers via Zoom, conducted in the teacher's office. These interviews were divided into two phases: first, participants shared their screen and interacted with an ActiveAI module for 15 min, during which they were encouraged to verbalize their thoughts (think-aloud protocol). Following module completion, a 15-minute interview was conducted to understand their learning experiences and suggestions for instructional and system design enhancements. They were asked to rate the likelihood of applying what they learned on a scale from 1 (not likely) to 10 (very likely). Ethical compliance was ensured through Institutional Review Board (IRB) approval for participant recruitment and data collection.

4.4 Data Analysis

For RQ1, we analyzed students' pre-and post-test scores in two learning objectives by converting their responses into percentage scores. For RQ2, we measured students' competence in using AI technology by averaging their responses to 12 survey questions, ranging from 1 to 7, based on the validated AI literacy scale by Wang et al. (2023). For RQ3, student engagement for each module was calculated using the one-item survey question "*How engaged are you in the learning activities of this unit?*", resulting in an engagement score from 1 to 7. Regarding students' think-aloud and interviews, two researchers performed a thematic analysis of the video recordings and transcripts. This approach allowed us to delve deeper into the students' experiences and interactions with the learning material to further unpack our findings in the research questions.

5 Results

5.1 RQ1: Higher Learning Outcomes for L1, No Significant Improvement for L2

Distribution of the Data. We conducted the Shapiro-Wilk test to examine the normality of both learning objectives L1 (pre-test: $p <.001$, post-test: $p <.001$) and L2 (pre-test: $p <.001$, post-test: $p <.001$). Neither the pre- or post-tests for L1 and L2 are normally distributed.

Difficulty Difference Between Two Test Versions. Given that our data did not follow a normal distribution, we conducted a non-parametric statistical

Table 2. Pre- and post-test performance for L1 & L2 by condition

		Learning Objective 1			Learning Objective 2		
Condition	Test	Median	S.D.	*N*	Median	S.D.	*N*
Treatment	Pre	58.82	23.24	61	50.00	15.72	44
	Post	70.59	20.59	61	66.67	22.19	44
Control	Pre	44.12	25.87	54	60.00	22.24	55
	Post	58.82	26.04	54	60.00	27.13	55

test, the Mann-Whitney U test, to determine if there was a significant difference between students within the same treatment condition who received different pre and post-tests in Table 1. Confirming the test versions for Learning Objective 1 (L1) were well matched, we found no significant differences for students in the treatment condition who took different pre-test versions, $U = 29.5$, $p = .316$, CLES = .34, and post-test versions, $U = 36.0$, $p = .582$, CLES = .41. Similarly, we found no significant differences for students in the control condition who took different pre-tests, $U = 31.5$, $p = .204$, CLES = .75, and post-tests, $U = 29.0$, $p = .340$, CLES = .69. Similarly for Learning Objective 2 (L2), we found no significant differences in treatment pre-test versions, $U = 21.5$, $p = 1.0$, CLES = .51, treatment post-test versions, $U = 21.0$, $p = 1.0$, CLES = .50, control pre-test versions, $U = 28.0$, $p = .262$, CLES = .32, and control post-test versions, $U = 29.5$, $p = .313$, CLES = .34. Given consistent student performance across test versions, we aggregate test version groups into a single group for each learning objective and test-time.

Effectiveness of ActiveAI on Learning. Initial analysis using the Wilcoxon Signed-Rank Test indicated significant improvements in scores from pre-test to post-test for both Learning Objectives L1, $W = 1263.0$, $p < .001$, CLES = .38 and L2, $W = 842.0$, $p = .008$, CLES = .41, across all conditions. These results suggest in both treatment and control conditions, students' performance improved in both objectives after the learning session.

Descriptive statistics about student test scores for each learning objective (L1 and L2) in each condition are included in Table 2. We conducted a series of Mann-Whitney U tests to examine differences between the treatment and control groups. The pre-test scores for both L1, $U = 1921.0$, $p = .122$, CLES = .58 and L2, $U = 1182.5$, $p = .846$, CLES = .49 showed no significant differences between the treatment and control groups, indicating that the two groups are comparable. We then found a significant difference between the treatment and control groups on L1 post-test scores, $U = 2049.0$, $p = .023$, CLES = .62, showing students who used ActiveAI achieved significantly higher post-test performance in L1 than those in the control condition. However, there were no condition differences in L2 post-test scores, $U = 1282.5$, $p = .608$, CLES = .53.

To probe possible reasons why ActiveAI treatment was more effective than the tell-and-practice control for L1 but not for L2, we analyzed students' think-aloud transcripts while interacting with the module and the follow-up interviews.

We found that, when completing the AI literacy learning activities, a recurring student misconception was that AI systems "think just like humans". This assumption, while incorrect in general, occasionally leads to correct answers, especially in module L2 which focuses on image classification. For instance, when tasked with identifying relevant features for a fruit image classification model, students were presented with options like the fruit's country of origin, physical attributes, or taste. P2 reasoned: *"Taste is too abstract, and determining a fruit's origin country is challenging. Therefore, 'color, shape, and size' are the most straightforward attributes for 'anyone' to identify a fruit."* This response, though aligning with the correct answer, was based on the assumption that AI "thinks" like humans, rather than an understanding of AI capabilities.

However, this misconception fails in L1 and in more complex scenarios in L2. For example in L2, when faced with an assessment (If a classification model is trained on a dataset with only red and yellow fruits, how will it likely classify a green fruit?), P2 initially wavered between two incorrect options (a. Correctly, as a new category of green fruits; b. It will ask for more data on green fruits before making a decision). She reasoned that distinguishing between green, yellow, and red is easy for humans, hence she chose a. It was only after selecting the wrong answers and engaging with the feedback and hints that she realized the need to shift her perspective: *"I think I start to get it; I have to think from the robot's perspective."* Conversely, L1's focus on how temperature settings in large language models (LLMs) influence chatbot responses, which diverge from human decision-making processes, did not lend itself to the *"AI thinks like humans"* heuristic. The heuristic is not applicable for L1 items in the post-test.

For the L2 post-test, as both the correct understanding of AI principles and the heuristic lead to the correct answer for the majority of items (5 out of 6), it is difficult to discern whether treatment students are relying less on the misunderstanding as a heuristic. However, for one specific item that directly relates to the formative assessment described above, which was designed to reveal reliance on the heuristic, we observed a difference in student performance by condition: 70.45% of students in the treatment group answered correctly, compared to 58.18% in the control group. Given that this item's distractor options were specifically aligned with heuristic thinking, it could suggest a deeper understanding of AI concepts among students in the treatment group.

5.2 RQ2: Improved Self-Reported Competency in AI Technology

The Shapiro-Wilk test indicated that the data on students' competence in AI were not normally distributed. Therefore, we conducted a Mann-Whitney U test and found that the treatment group scored higher in overall competence in using AI technology compared to the control group, $p = .022$ (Table 3).

In terms of the qualitative feedback, participants expressed that what they learned in ActiveAI could help them better leverage AI technology in the future. Specifically, six out of eight students (75%) rated their likelihood of applying what they learned in the future as'very likely' (scores above 7 out of 10). P1 remarked, *"In this AI era, it helps to clear doubts and gain a deeper understanding of emerging*

Table 3. Student competence in AI literacy by condition

Treatment		Control		***U***	p-value	CLES
Median	S.D.	Median	S.D.			
4.4	1.2	4.0	1.2	1237.5	.022	0.38

Table 4. Student engagement for L1 & L2 by condition

Learning Objective	Treatment		Control		***U***	p-value	CLES
	Median	S.D.	Median	S.D.			
L1	4.0	2.1	4.0	1.6	1458.5	.281	.44
L2	5.0	1.4	5.0	1.7	1235.0	.860	.51

technologies." Another student, P6, shared, *"It's highly probable, as AI is becoming increasingly prevalent. I might encounter similar situations, even if I don't pursue a career in AI. The modules exposed me to knowledge I rarely encounter. I really liked it and might use it in the future, possibly in a related job."* However, two out of the eight students rated their likelihood as 'likely' (scores between 5 and 6), possibly due to the inherent difficulty of the content. P8 who gave a score of 5 commented on the complexity of the concepts: *"For now, it seems okay, but in the future, it might be useful. Some of the knowledge is quite difficult to grasp, but understanding the simpler parts is feasible"*.

Notably, although L2 did not explicitly address generative AI, some participants expressed that they were able to apply their understanding of how training data influences classification models to grasp the content generation process of ChatGPT, a tool they frequently use. For example, P3 noted, *"When using ChatGPT, I can now understand how it generates responses."* P2 added, *"In daily life, using ChatGPT might yield answers tailored to my habits... Reducing AI bias could lower the risks associated with AI use, protect user rights, and ensure safer AI interactions. It could also lead to more accurate responses when AI generates content."*

5.3 RQ3: Engagement Levels

The Shapiro-Wilk test indicated that the learner engagement data for both learning objectives L1, $p < .001$ and L2, $p < .001$ exhibited a non-normal distribution. Therefore, we conducted a Mann-Whitney U test but found no significant differences in student engagement for L1, $p = .281$ or L2, $p = .860$.

When inquiring about the elements of the ActiveAI modules that engaged them versus those that led to disengagement, students overwhelmingly favored the hands-on experience of constructing AI solutions. This preference was particularly directed towards the intelligent agent component, a feature present in both the treatment and control conditions. For example, P3 mentioned, *"I found the hands-on classification task particularly memorable, especially when*

I encountered a little failure. It made a lasting impression and was the most interesting part for me."

P2 elaborated on this preference: *"I enjoy practical tasks more than theoretical ones. For instance, physically demonstrating a concept is far more appealing than just learning about it theoretically. Theoretical content tends to be less concrete and more abstract, making it somewhat dull. For example, the second part of the module, categorizing images of dogs playing or sleeping, was engaging because it was hands-on. However, the third part, which focused on machine learning bias in society, felt more tedious as it was largely theoretical. Such theoretical aspects are often basic and can be deduced using common sense or everyday experiences."*

6 Discussion and Limitations

The ActiveAI modules demonstrated significant post-test performance in ActiveAI, with L1 showing a notable advantage over the control condition, while L2 did not exhibit a significant difference. These results suggest that the tutoring system has the potential to enhance AI literacy among learners, though the differential impact between L1 and L2 raises intriguing questions about the relative effectiveness of these modules.

Qualitative analysis of students' interactions suggests that the heuristic of "AI thinks just like humans" might have influenced the outcomes, particularly in L2. This heuristic allowed students to answer some image classification questions correctly without fully grasping the underlying AI concepts. While ActiveAI's feedback and hints feature has shown potential in aiding learners to identify and rectify such misconceptions, as evidenced by think-aloud sessions, the current design of the L2 assessment may not adequately reflect this learning process. Future research should integrate these findings into the instructional and assessment design processes to more effectively capture and enhance the learning experience. Future work can also investigate the use of generative AI to provide more personalized feedback and hints [25].

In addition, this observation has significant implications for K-12 AI education, particularly in the design of AI Literacy assessments. To counteract the reliance on such heuristics, we recommend incorporating distractors that challenge this assumption. This approach can help identify students' misconceptions and assess their understanding of AI principles more accurately, encouraging deeper engagement with AI concepts. For example, the fruit image classification task could be redesigned to more effectively test understanding of AI capabilities. A revised question might ask:

Redesigned Question: When training an image classification model to identify poisonous plants from photos, which features should it prioritize in training data?

a. The presence of specific poisonous ingredients.
b. The color and shape of the plant.
c. The plant's common habitat.

Applying the “AI thinks just like humans” heuristic would lead students to choose option a, which, while logical from a human perspective, is incorrect for an AI trained solely on visual data. This type of question, coupled with targeted feedback for incorrect choices, could effectively address misconceptions and promote a deeper understanding of AI principles. Furthermore, educators could intentionally incorporate this heuristic in instructional design, embedding it within the experiential learning cycle to provoke cognitive conflict [15], leading to enriched learning outcomes.

Our findings indicate that students reported a statistically significant increase in AI competency after using ActiveAI compared to the control condition. This improvement may be attributed to the system’s on-demand hints and targeted feedback, which likely enhanced learners’ self-efficacy in AI literacy. However, as these outcomes are based on self-reported surveys, caution is warranted in interpreting these results due to potential biases such as social desirability bias and the lack of objective measures. Further research should explore the transferability of this increased competency to real-world interactions with AI technologies, potentially through project-based assessments [34].

Additionally, the study highlighted that both the treatment and control conditions maintained relatively high levels of student engagement, with hands-on activities involving intelligent agents identified as key engaging factors. This insight suggests that the inclusion of intelligent agents in AI literacy education could be a valuable strategy for enhancing student engagement in digital learning environments. However, the engagement measures were also self-reported, which calls for caution in interpreting these findings. Future studies should incorporate more objective measures of engagement, such as behavioral observations or interaction log data, to provide a more comprehensive understanding of student engagement.

Also, conducting the study with a limited sample of students from one grade in East Asia may limit the generalizability of the findings. Cultural and educational system differences could influence how AI literacy is perceived and learned. Replicating the study in diverse educational settings and with students from different cultural backgrounds would help validate the findings and enhance their applicability.

7 Conclusion

A controlled classroom experiment involving 171 middle school learners demonstrates that ActiveAI effectively aids middle school students in acquiring AI literacy and developing AI technology competency. This was achieved through the interactive tutoring system’s active learning experiences, consisting of experiential learning and hands-on interactions with intelligent agents. The system’s ability to provide immediate, targeted feedback and on-demand hints has addressed the need for an effective AI literacy learning experience without imposing extensive demands on teacher resources. This study contributes insights into the design of a scalable and pedagogically sound tutoring system, enhancing the accessibility of K-12 AI literacy education.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Ali, S., DiPaola, D., Lee, I., Hong, J., Breazeal, C.: Exploring generative models with middle school students. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2021)
2. Bell, T., Alexander, J., Freeman, I., Grimley, M.: Computer science unplugged: school students doing real computing without computers. New Zealand J. Appl. Comput. Inf. Technol. **13**(1), 20–29 (2009)
3. Bonwell, C.C., Eison, J.A.: Active learning: creating excitement in the classroom. 1991 ASHE-ERIC higher education reports. ERIC (1991)
4. Chiu, T.K., Meng, H., Chai, C.S., King, I., Wong, S., Yam, Y.: Creation and evaluation of a pretertiary artificial intelligence (AI) curriculum. IEEE Trans. Educ. **65**(1), 30–39 (2021)
5. Crow, T., Luxton-Reilly, A., Wuensche, B.: Intelligent tutoring systems for programming education: a systematic review. In: Proceedings of the 20th Australasian Computing Education Conference, pp. 53–62 (2018)
6. Druga, S.: Growing up with AI: Cognimates: from coding to teaching machines. Ph.D. thesis, Massachusetts Institute of Technology (2018)
7. Eguchi, A., Okada, H., Muto, Y.: Contextualizing AI education for k-12 students to enhance their learning of AI literacy through culturally responsive approaches. KI-Künstliche Intelligenz **35**(2), 153–161 (2021)
8. Felder, R.M., Brent, R.: Active learning: an introduction. ASQ Higher Educ. Brief **2**(4), 1–5 (2009)
9. Gennari, R., Melonio, A., Pellegrino, M.A., D'Angelo, M.: How to playfully teach AI to young learners: a systematic literature review. In: Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter, pp. 1–9 (2023)
10. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2007)
11. Henry, J., Hernalesteen, A., Collard, A.S.: Teaching artificial intelligence to k-12 through a role-playing game questioning the intelligence concept. KI-Künstliche Intelligenz **35**(2), 171–179 (2021)
12. Kaspersen, M.H., Bilstrup, K.E.K., Van Mechelen, M., Hjort, A., Bouvin, N.O., Petersen, M.G.: High school students exploring machine learning and its societal implications: opportunities and challenges. Int. J. Child-Comput. Interact., 100539 (2022)
13. Kluger, A.N., DeNisi, A.: The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol. Bull. **119**(2), 254 (1996)
14. Kolb, D.A.: Experiential learning: experience as the source of learning and development. FT Press (2014)
15. Limón, M.: On the cognitive conflict as an instructional strategy for conceptual change: a critical appraisal. Learn. Instr. **11**(4–5), 357–380 (2001)
16. Ma, R., Sanusi, I.T., Mahipal, V., Gonzales, J.E., Martin, F.G.: Developing machine learning algorithm literacy with novel plugged and unplugged approaches. In: Proceedings of the 54th ACM Technical Symposium on Computer Science Education, vol. 1, pp. 298–304 (2023)

17. Mariescu-Istodor, R., Jormanainen, I.: Machine learning for high school students. In: Proceedings of the 19th Koli Calling International Conference on Computing Education Research, pp. 1–9 (2019)
18. Mayer, R.E.: Cognitive Theory of Multimedia Learning. Cambridge Handbook of Multimedia Learning, vol. 41, pp. 31–48 (2005)
19. Nagashima, T., et al.: Using anticipatory diagrammatic self-explanation to support learning and performance in early algebra. Grantee Submission (2021)
20. Ng, D.T.K., Leung, J.K.L., Su, M.J., Yim, I.H.Y., Qiao, M.S., Chu, S.K.W.: AI literacy in K-16 classrooms. Springer International Publishing AG (2023). https://doi.org/10.1007/978-3-031-18880-0
21. Ogan, A., Aleven, V., Jones, C.: Culture in the classroom: challenges for assessment in ill-defined domains. In: Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at Intelligent Tutoring Systems, pp. 92–100 (2006)
22. Piaget, J.: The Construction of Reality in the Child, vol. 82. Routledge (2013)
23. Prince, M.: Does active learning work? A review of the research. J. Eng. Educ. **93**(3), 223–231 (2004)
24. Song, J., Yu, J., Yan, L., Zhang, L., Liu, B., Zhang, Y., Lu, Y.: Develop AI teaching and learning resources for compulsory education in China. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 16033–16039 (2023)
25. Stamper, J., Xiao, R., Hou, X.: Enhancing LLM-based feedback: insights from intelligent tutoring systems and the learning sciences. arXiv preprint arXiv:2405.04645 (2024)
26. Su, J., Ng, D.T.K., Chu, S.K.W.: Artificial intelligence (AI) literacy in early childhood education: the challenges and opportunities. Comput. Educ.: Arti. Intell. **4**, 100124 (2023)
27. Theophilou, E., Lomonaco, F., Donabauer, G., Ognibene, D., Sánchez-Reina, R.J., Hernàndez-Leo, D.: AI and narrative scripts to educate adolescents about social media algorithms: insights about AI overdependence, trust and awareness. In: Viberg, O., Jivet, I., Muñoz-Merino, P.J., Perifanou, M., Papathoma, T. (eds.) Responsive and Sustainable Educational Futures: 18th European Conference on Technology Enhanced Learning, EC-TEL 2023, Aveiro, Portugal, September 4–8, 2023, Proceedings, pp. 415–429. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-42682-7_28
28. Toivonen, T., Jormanainen, I., Kahila, J., Tedre, M., Valtonen, T., Vartiainen, H.: Co-designing machine learning apps in K–12 with primary school children. In: 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), pp. 308–310. IEEE (2020)
29. Touretzky, D., Gardner-McCune, C., Cox, B., Uchidiuno, J., Kolodner, J., Stapleton, P.: Lessons learned from teaching artificial intelligence to middle school students. In: Proceedings of the 54th ACM Technical Symposium on Computer Science Education, vol. 2, pp. 1371–1371 (2022)
30. Touretzky, D., Gardner-McCune, C., Seehorn, D.: Machine learning and the five big ideas in AI. Int. J. Artif. Intell. Educ., 1–34 (2022)
31. Tseng, Y.J., Xiao, R., Bogart, C., Savelka, J., Sakr, M.: Assessing the efficacy of goal-based scenarios in scaling AI literacy for non-technical learners. In: Proceedings of the 55th ACM Technical Symposium on Computer Science Education, vol. 2, pp. 1842–1843 (2024)
32. Voulgari, I., Zammit, M., Stouraitis, E., Liapis, A., Yannakakis, G.: Learn to machine learn: designing a game based approach for teaching machine learning to primary and secondary education students. In: Interaction Design and Children, pp. 593–598 (2021)

33. Wang, B., Rau, P.L.P., Yuan, T.: Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. Behav. Inf. Technol. **42**(9), 1324–1337 (2023)
34. Williams, R., et al.: AI+ ethics curricula for middle school youth: lessons learned from three project-based curricula. Int. J. Artif. Intell. Educ. **33**(2), 325–383 (2023)
35. Wood, D., Bruner, J.S., Ross, G.: The role of tutoring in problem solving. J. Child Psychol. Psychiatry **17**(2), 89–100 (1976)
36. Zimmerman, B.J.: Becoming a self-regulated learner: an overview. Theory Pract. **41**(2), 64–70 (2002)