Journal of General Education and Humanities
Vol. 5, No. 5, October 2026, pp. 6987 – 7003, https://doi.org/10.58421/gehu.v5i5.1993
ISSN 2963-7147

Tangible AI in the Elementary Classroom: Effects on AI Literacy, Computational Thinking, and Learner Dispositions

Harsa Wara Prabawa[1], Jajang Kusnendar[1], Yuli Sopianti[1], Andhin Dyas Fitriani[1], Rohati[2]
[1]Universitas Pendidikan Indonesia, Jawa Barat, Indonesia
[2]Universitas Jambi, Jambi, Indonesia

Article Info

Article history:

Received 2026-08-09
Revised 2026-10-08
Accepted 2026-10-30

Keywords:

Tangible AI
AI Literacy
Embodied Cognition
AI-PCK
Elementary Education

ABSTRACT

Artificial intelligence (AI) is increasingly ubiquitous, yet AI literacy remains underdeveloped in elementary education, where young learners often view AI in anthropomorphic or magical terms. This study examined changes in computational thinking (CT), AI concept mastery, learning motivation, and programming self-efficacy following a tangible AI intervention, and explored the teacher's implementation experience. Guided by a framework integrating the AI Literacy Framework and the SEAME Model, a four-week intervention (4 × 60 minutes) was implemented with 36 fifth-grade students and one teacher in a public elementary school, using a single-group pretest–posttest design within a mixed-methods approach. Quantitative data were analyzed using repeated-measures ANCOVA, with prior computing experience as a covariate; qualitative data from semi-structured teacher interviews (during and after the intervention) and observations of all four sessions were analyzed thematically. Significant pre–post gains were found for all four composite outcomes, largest for AI concept mastery (M = 8.4 to 15.1; η^2p = .70) and total CT score (η^2p = .56), although the CT abstraction subdimension did not improve (M = 2.3 to 2.1). Students able to explain how machine learning models learn from data rose from 28% to 83%. Anthropomorphic descriptions of AI decreased from 89% to 25% of students, with post-intervention explanations increasingly referring to sensors, data, and training. Three qualitative themes emerged: strong pedagogical confidence but limited AI content knowledge, persistent infrastructural constraints, and reduced technical anxiety through gradual mastery. Tangible AI may offer a promising bridge to early AI literacy; however, without a control group, changes cannot be causally attributed to the intervention, and these single-school findings are preliminary.

This is an open-access article under the CC BY-SA license.

Corresponding Author:

Harsa Wara Prabawa
Computer Science Education Study Program, Faculty of Mathematics and Science Education, Universitas Pendidikan Indonesia, Bandung, Indonesia

1. INTRODUCTION

Artificial intelligence (AI) now permeates children's everyday lives, from voice assistants and recommendation engines to intelligent tutoring systems, making AI literacy a foundational competence that should be developed from the elementary years onward [1] [2]. However, the elementary level remains underserved. While AI coursework and resources have proliferated in secondary and higher education, developmentally appropriate tools and curricula for young learners remain scarce [3], leaving teachers without the scaffolding needed to introduce AI concepts in a meaningful and accessible way.

This gap matters because young children tend to conceptualize AI through anthropomorphic or magical lenses, attributing human-like emotions, intentions, and even sentience to systems that are, in fact, programmable artifacts [4], [5]. The "black box" character of modern AI, in which data processing, pattern recognition, and algorithmic decision-making remain hidden from users, reinforces these misconceptions. Physical, observable systems that visibly respond to input may help to open this black box, supporting a transition from anthropomorphic toward mechanistic understandings of AI [4], [6].

The empirical evidence on how to achieve this remains fragmented in four respects. First, most elementary interventions target computational thinking (CT) or general computing through unplugged or block-based programming [7], [8], whereas studies of tangible tools for teaching AI concepts specifically remain scarce [6], [9]. Second, cognitive outcomes, such as conceptual understanding and CT, and affective outcomes, such as motivation and self-efficacy, are typically examined in isolation rather than as an integrated set. Third, although anthropomorphic misconceptions are well documented [4], [5], there is little evidence on whether tangible interventions are associated with a measurable shift toward mechanistic understanding. Fourth, teacher preparedness, particularly the technological, pedagogical, and content knowledge needed to facilitate hands-on AI activities, has received less attention than student outcomes [10], [11], [12], [13]. The present study addresses all four gaps within a single intervention and evaluation design.

Several related terms are used in this literature and are distinguished here. Physical computing refers to interactive systems that sense and respond to the physical world through sensors, actuators, and microcontrollers [14]. Unplugged activities exclude computational hardware altogether and teach computing concepts through games, cards, and role-play [7] [15]. Embodied learning describes learning that arises from the integration of physical action, perception, and cognition, with or without digital hardware [16]. In this study, tangible AI denotes the use of physically manipulable or sensor-based artifacts, namely the AI for Oceans simulation, sensor cards, decision-tree and action tokens, and a biased data deck, to make otherwise invisible AI processes (sensing, rule-based decision-making, and data bias) observable and manipulable by learners. The intervention therefore draws on both the physical computing and unplugged traditions within an embodied learning approach.

Because the study draws on several frameworks, each is assigned a distinct role (Figure 1). Embodied cognition and constructionism are the learning theories that explain why a tangible approach should support learning. The AI Literacy Framework and the SEAME Model specify what is taught and assessed. The TPACK framework provides the lens for

examining how the teacher enacts the intervention. Figure 1 shows how these layers connect the intervention to the expected learning mechanisms, student outcomes, and teacher implementation factors.

Embodied cognition theory holds that abstract knowledge is grounded in sensorimotor experience rather than acquired through disembodied symbol manipulation alone [17], and constructionism adds that learning is deepened when learners build and test tangible, shareable artifacts [18]. Together, these theories yield four expectations, labelled (a) to (d) in Figure 1. (a) Physically manipulating representations of AI processes, such as arranging sensor cards to simulate perception, provides concrete referents that support more accurate mental models of AI than verbal or symbolic instruction alone [16], [19]. (b) Externalizing implicit computational steps, such as breaking a classification task into IF-THEN-ELSE rules laid out on the floor, makes those steps ordered and inspectable, supporting decomposition, algorithmic thinking, and debugging [20]. (c) Tracing a visible causal chain from input, through an inspectable rule or model, to output should replace implicit attributions of intentionality with mechanistic explanations of system behavior [21]. (d) Repeatedly building, testing, and correcting working artifacts provides mastery experiences that are expected to strengthen students' motivation and self-efficacy toward AI and programming [18].

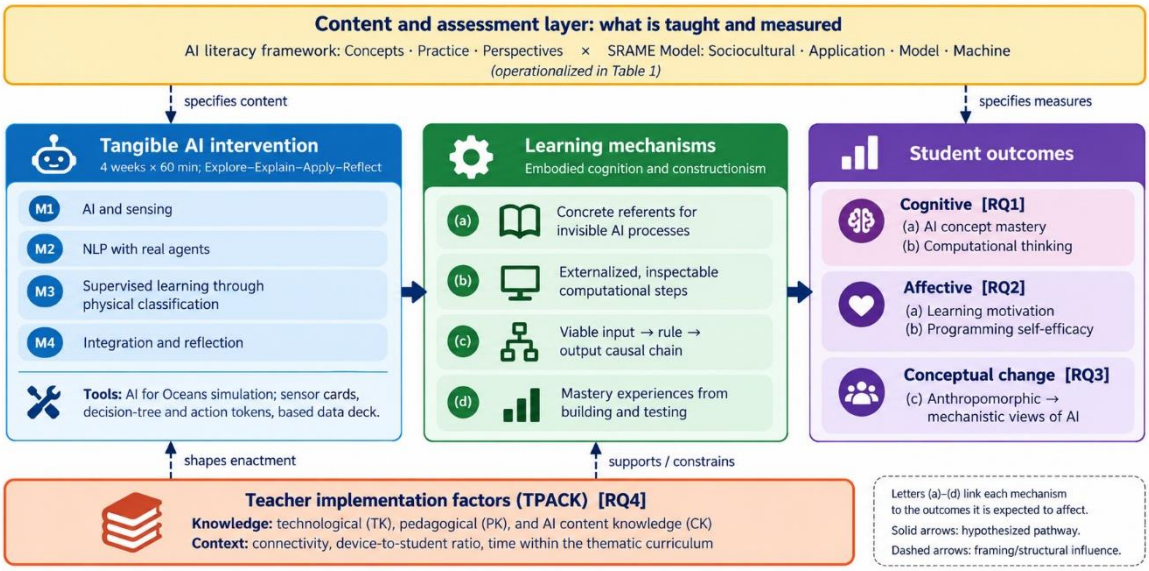


Figure 1. Conceptual Framework

To specify learning content and outcomes, the study integrates two AI-specific frameworks. The AI Literacy Framework [22], [23], organizes outcomes into three dimensions: Concepts (technical knowledge of how AI systems work, including machine learning and computer vision), Practice (applying AI techniques such as data collection, model training, and iterative testing), and Perspectives (ethical and sociotechnical issues such as fairness, accountability, and transparency). The SEAME Model [2], developed within the AI4K12 initiative, complements it with four levels: Social/Ethical impact, Application of AI in everyday tools, Model representations such as decision trees, and the

Machine level of sensors and hardware that enable AI. The two frameworks are complementary rather than redundant: the AI Literacy Framework defines the competence domains, whereas SEAME distinguishes the technical layers, from hardware to societal impact, at which those competences are exercised. This distinction is particularly relevant to a tangible intervention, in which the Machine level is physically present. Table 1 maps each framework dimension to its learning objectives, intervention activities, measurement instruments, and corresponding research question.

Table 1. Integration of the AI Literacy Framework and the SEAME Model with intervention activities, measurement instruments, and research questions

Framework Dimension	Learning Objective	Intervention Activity	Measurement Instrument	RQ
Technical Knowledge / Concepts (AI Literacy Framework); Model and Machine (SEAME)	Understand how sensors perceive data and how rule-based logic (IF-THEN-ELSE) transforms input into output.	Module 1: Introduction to AI and Sensing (AI for Ocean simulation, sensor cards); Module 2: Natural Language Processing with Real Agents (decision tree and action tokens)	AI Concept Mastery test (technical knowledge items); classroom observation of conceptual change indicators	RQ1, RQ3
Practice / Application (AI Literacy Framework); Application (SEAME)	Apply data collection and iterative model training to classify physical objects	Module 3: Supervised Learning Through Physical Classification (labeling data, training simple models with sensor cards and action tokens)	AI Concept Mastery test (application items); CT Skills test (algorithmic thinking and pattern recognition subscales)	RQ1
Perspectives / Ethics (AI Literacy Framework); Social/Ethical (SEAME)	Recognize how biased or unrepresentative training data can produce unfair or inaccurate decisions	Biased Data Deck activity (embedded in Modules 3 and 4); Module 4: Integration and Reflection	AI Concept Mastery test (ethical perspective items); teacher interview (TPACK protocol); classroom observation	RQ1, RQ3
Computational thinking (cross-cutting; supports all three AI Literacy Framework dimensions)	Develop abstraction, decomposition, algorithmic thinking, pattern recognition, and debugging.	All four modules, with particular emphasis on the Decision Tree and Action Tokens activity (Module 2) and iterative testing (Module 3)	CT Skills test [24]	RQ1
Affective and motivational dispositions (supports engagement with Application and Practice dimensions)	Increase self-efficacy and intrinsic motivation toward AI and programming tasks.	Collaborative, project-based challenge concluding each module; gradual scaffolding of task difficulty across four weeks	Learning Motivation Survey (adapted from Glynn et al., 2011 [25]); Self-Efficacy Survey (adapted from Tsai et al., 2018 [26])	RQ2
Teacher preparedness (cross-cutting; informs delivery of all dimensions above)	Examine how teachers integrate technological, pedagogical, and content knowledge while facilitating the intervention.	Teacher facilitation across all four modules	Semi-structured interview (TPACK protocol [10]; classroom observation	RQ4

Finally, because hands-on AI activities depend heavily on teacher facilitation, the study adopts the TPACK framework [10] to examine implementation. TPACK conceptualizes effective technology integration as the interplay of technological, pedagogical, and content knowledge within a specific context. This lens is particularly relevant because elementary teachers frequently report limited AI content knowledge and uncertainty about integrating AI into their teaching [11], [13].

Accordingly, this study examines a four-week tangible AI intervention with fifth-grade students and its implementation by the classroom teacher. Because the study used a single-group design, the research questions address pre-to-post change and conceptual development rather than causal effects. They are grouped into student-level and teacher-level questions: 1) To what extent do students' AI concept mastery and computational thinking skills change from pretest to posttest following the tangible AI intervention, after controlling for prior computing experience? 2) To what extent do students' learning motivation and programming self-efficacy change from pretest to posttest following the intervention, after controlling for prior computing experience? 3) How do students' conceptions of AI change over the course of the intervention, particularly from anthropomorphic toward mechanistic explanations? and 4) How does the classroom teacher draw on technological, pedagogical, and AI content knowledge when implementing the intervention, and what pedagogical and infrastructural challenges arise?

2. METHOD

2.1 Research Design

This study used a single-group pretest–posttest design embedded within a mixed-methods approach. Because there was no control or comparison group, the design is best described as pre-experimental rather than quasi-experimental [27]. It was chosen deliberately for this exploratory first implementation, whose aims were to examine whether a tangible AI intervention is feasible in a regular elementary classroom, to estimate the size of pre-to-post change, and to document implementation conditions for future controlled studies. The design cannot rule out alternative explanations such as maturation, testing effects, or other classroom experiences during the same period. Pre-to-post differences are therefore interpreted as change associated with the intervention, not as causal effects. Quantitative pretest and posttest data addressed student outcomes (RQ1–RQ2). Qualitative data from classroom observations and teacher interviews were collected during the intervention to explain these outcomes and to address students' conceptions of AI (RQ3) and teacher implementation (RQ4).

The study took place in one public elementary school in Bandung, Indonesia, that was purposively selected [28] because it had basic technology facilities and school support for technology-integrated learning. Purposive selection applied to the school and the class, not to individual students: one fifth-grade class was chosen because its teacher was willing to participate, and all students in that class were invited. The final sample comprised 36 students (aged 11–12 years; 19 girls and 17 boys) and their homeroom teacher (8 years of teaching experience). None of the students had previously received AI or structured programming instruction, and the teacher had no formal training in AI education. Because

the sample consists of one class, one teacher, and one school, it is context-specific, and the findings are not intended to be generalized statistically to other elementary students or teachers.

This study has been approved and authorized by the school principal. Parents or guardians provided written consent, students gave verbal assent after the study was explained in simple language, and teachers provided written consent. Participation was voluntary and did not affect students' grades. All data were anonymized using participant codes, stored securely, and accessible only to the research team.

2.2 Intervention Design

The intervention was a four-week, activity-based program grounded in constructivist and constructionist learning theories [18] and in embodied cognition theory [16], [17]. It consisted of four weekly 60-minute sessions (four hours in total) integrated into the existing thematic curriculum. The four modules were: (1) Introduction to AI and Sensing, in which students used AI for Oceans and Sensor Cards to explore how sensors collect data (available at https://studio.code.org/courses/oceans/units/1/lessons/1/levels/1); (2) Natural Language Processing with Real Agents, in which students used Decision Tree and Action Tokens to "unplugged programming"; (3) Supervised Learning Through Physical Classification, in which students labeled data, trained simple models, and used the Biased Data Deck to observe how unbalanced data makes a model fail; and (4) Integration and Reflection, in which groups designed a combined AI system and explained how it works.

Each session followed the same Explore–Explain–Apply–Reflect cycle: students first explored the tools (about 15 minutes); the teacher then explained the underlying AI concept (10 minutes); groups of 3–4 students completed a hands-on challenge (25 minutes); and the session closed with a reflection in which students explained what the system did and why (10 minutes). The tools are described in detail in Table 1, and they were selected for their accessibility, affordability, and suitability for elementary students [4], [6].

The intervention was delivered by the classroom teacher, with a researcher present as an observer and technical assistant. Before implementation, the teacher attended two preparation sessions (total 120 minutes) led by the researchers, covering the AI concepts in each module, hands-on practice with the tools, and the lesson plans. A short planning meeting was held before each session.

2.3 Data Collection Tools and Instruments

All student instruments were administered in Indonesian before and after the intervention and were read aloud by the teacher to reduce the reading load. Instruments adapted from research involving older learners were simplified in language to suit fifth-grade students, reviewed by two experts for age-appropriateness, and pilot-tested on a small group of six fifth-graders from a different class prior to use. Instrument reliability was assessed using the study's sample.

AI concept mastery was measured using a 20-item multiple-choice test developed by the researchers. The items were designed to cover three dimensions of the AI Literacy Framework [22]: technical knowledge (7 items), application (8 items), and ethical

perspectives (5 items). Two experts evaluated the relevance of each item, and content validity was calculated using Aiken's V formula (V = 0.75–0.77); items with low ratings were revised. Each item was scored 1 (correct) or 0 (incorrect), resulting in a total score range of 0–20 (KR-20 = 0.72). Computational thinking (CT) was measured using a 25-item multiple-choice test covering abstraction, decomposition, algorithmic thinking, pattern recognition, and debugging, adapted from Korkmaz et al. [24] (score range 0–25; KR-20 = 0.78).

Learning motivation was measured with 25 items adapted from the Science Motivation Questionnaire II [25], and programming self-efficacy with 15 items adapted from the Computer Programming Self-Efficacy Scale [26]. Both were rated on a 5-point scale. Because the original instruments were developed for university students, the items were rewritten in simpler language, and references to "science" or "programming courses" were replaced with "learning about AI and computers." Internal consistency was acceptable (Cronbach's α = 0,74 and 0,72, respectively).

Prior computing experience, used as a covariate, was measured at pretest with a short self-report question asking how many hours per week students usually spent on programming activities or on writing prompts for AI tools (answer options: Never, less than 1 hour per day, more than 1 hour per day). Students' conceptions of AI (RQ3) were collected before and after the intervention through an open-ended question / short interviews. Each response was coded as anthropomorphic (describing AI as having feelings, intentions, or its own thinking), mechanistic (referring to sensors, rules, data, or training), or unclear.

Qualitative data were collected through semi-structured interviews with the teacher at three points (before, during, and after the intervention). The interviews were audio-recorded, transcribed, and guided by a TPACK-based protocol [10] that asked about the teacher's use of technological, pedagogical, and content knowledge, the challenges faced, and the student learning observed. All four sessions were also observed by one researcher, who took field notes on students' interaction with the tools, the teacher's instructional steps, signs of conceptual change, and technical problems.

2.4 Data Analysis

Each outcome (CT, AI concept mastery, learning motivation, and self-efficacy) was analyzed with a repeated-measures ANCOVA, with time (pretest, posttest) as the within-subjects factor and prior computing experience (mean-centered) as the covariate. Although repeated-measures ANCOVA is feasible with 36 participants, because each student serves as their own comparison, the small sample requires caution. We therefore checked the assumptions (normality of residuals with Shapiro–Wilk tests; homogeneity of regression slopes through the time × covariate interaction), applied a Holm–Bonferroni correction [29] across the four outcomes, and report partial eta squared (η^2p) as the effect size. A sensitivity analysis in G*Power [30]showed that, with N = 36 and power = .80, the design could reliably detect only medium-to-large effects ($dz \geq 0.48$); smaller effects may have been missed. All 36 students completed both tests, so there were no missing data.

The qualitative data were analyzed using thematic analysis following Braun and Clarke [31]: reading the data repeatedly, generating initial codes, grouping codes into

themes, reviewing and naming the themes, and reporting them. Two researchers coded the data independently, compared their codes, and resolved disagreements through discussion. Trustworthiness was supported by triangulating interviews with observation notes, by asking the teacher to confirm a summary of the findings (member checking), and by keeping a record of coding decisions. Because only one teacher participated, the themes are presented as findings from a single illustrative case, not as general patterns among teachers.

The two strands were integrated at three stages. During data collection, the observations focused on the same constructs measured by the tests, such as signs of anthropomorphic or mechanistic reasoning. During analysis, the quantitative results were compared with the observation and interview evidence to see whether they confirmed or added to each other. During interpretation, the qualitative findings were used to explain the quantitative results, for example, why the abstraction dimension of CT did not improve.

3. RESULTS AND DISCUSSION

This study produced mutually supportive evidence from quantitative and qualitative data sources, indicating a significant improvement in students' cognitive understanding of artificial intelligence (AI) concepts as well as positive changes in their affective attitudes toward AI technology. Data analysis of 36 upper-grade elementary school students and one classroom teacher revealed meaningful learning gains, while also highlighting key challenges in implementation and the need for teacher professional development.

3.1. AI Concept Mastery and Computational Thinking

After controlling for factors related to experience interacting with AI and computer programming (measured based on self-reported hours of exposure to programming or simply writing prompts), an Analysis of Covariance (ANCOVA) revealed statistically significant increases in all measured outcomes (see Table 1).

Table 1 presents the ANCOVA results for all four outcomes, with prior computing experience (mean-centered) as the covariate. AI concept mastery increased from $M = 8.4$ ($SD = 2.3$) to $M = 15.1$ ($SD = 2.6$), a mean gain of 6.7 points (95% CI [5.16, 8.24]), $F(1, 34) = 78.52$, $p < .001$, $\eta^2p = .70$. This was the largest change among the four outcomes. The time × covariate interaction was not significant, indicating that gains did not depend on students' prior computing experience.

The proportion of students who could accurately explain how machine learning models learn from data rose from 28% (n = 10) at pretest to 83% (n = 30) at posttest. This proportion was obtained from open-ended item on the AI Concept Mastery Test, which asked students to "explain how a computer learns to recognize objects". Responses were scored as accurate when they mentioned all three of the following elements: (a) the system is given labeled examples or data, (b) it is trained to find patterns in those examples, and (c) it uses the learned patterns to classify new input. Responses mentioning only one or two of these elements, or none, were scored as not accurate. Two raters scored all responses independently, and disagreements were resolved through discussion. For example, a typical pretest response was "AI uses programmed code," while a typical posttest response was "input is provided, and sensors look for patterns they have learned from training data."

Table 1. Pre–post comparison of student outcomes, with prior computing experience as covariate (N = 36)

Measure	Pretest M (SD)	Posttest M (SD)	Mean difference [95% CI]	F(1, 34)	p (Holm)	η^2p
AI concept mastery (0–20)	8.4 (2.3)	15.1 (2.6)	6.7 [5.16, 8.24]	78.52	< .001	.70
CT total score (0–25)	14.2 (3.1)	19.7 (2.8)	5.5 [3.78, 7.22]	42.18	< .001	.56
Learning motivation (1–5)	3.2 (0.8)	4.1 (0.6)	0.9 [0.48, 1.32]	18.95	< .001	.37
Programming self-efficacy (1–5)	2.8 (0.9)	3.9 (0.7)	1.1 [0.65, 1.55]	24.31	< .001	.42

Table 2. Pre–post comparison of the five computational thinking dimensions (N = 36; 5 items each, range 0–5)

CT dimension	Pretest M (SD)	Posttest M (SD)	Mean difference	t(35)	p (Holm)	Cohen's d
Abstraction	2.3 (1.0)	2.1 (1.1)	−0.2	−1.00	.324	−0.17
Decomposition	3.2 (1.1)	4.5 (0.7)	1.3	6.50	< .001	1.08
Algorithmic thinking	2.4 (1.1)	4.3 (0.8)	1.9	8.77	< .001	1.46
Pattern recognition	3.3 (1.0)	4.6 (0.6)	1.3	6.00	< .001	1.00
Debugging	3.0 (1.2)	4.2 (0.8)	1.2	5.54	< .001	0.92
Total (0–25)	14.2 (3.1)	19.7 (2.8)	5.5	–	–	–

The total CT score increased from M = 14.2 (SD = 3.1) to M = 19.7 (SD = 2.8), a mean gain of 5.5 points (95% CI [3.78, 7.22]), $F(1, 34) = 42.18$, $p < .001$, $\eta^2p = .56$. Table 2 reports the five CT dimensions separately. Because each dimension contains only five items, the dimensions were compared with paired-sample t-tests, with a Holm–Bonferroni correction for the five comparisons, and these analyses are treated as exploratory. Algorithmic thinking showed the largest gain (M = 2.4 to 4.3, $t(35) = 8.77$, $p < .001$, $d = 1.08$). Decomposition, pattern recognition, and debugging also increased significantly. Abstraction was the only dimension that did not improve: the mean decreased slightly from 2.3 to 2.1, and this decrease was not statistically significant, $t(35) = -1.00$, $p = .324$, $d = -0.17$. Item-level inspection showed that most errors occurred on items requiring students to identify which details of a problem could be ignored.

3.2. Learning Motivation and Programming Self-Efficacy

Learning motivation increased from M = 3.2 (SD = 0.8) to M = 4.1 (SD = 0.6), mean gain = 0.9 (95% CI [0.48, 1.32]), $F(1, 34) = 18.95$, $p < .001$, $\eta^2p = .37$. At posttest, 94% of students (n = 34) rated the activities as "interesting" or "very interesting" on item of the motivation survey.

Programming self-efficacy increased from M = 2.8 (SD = 0.9) to M = 3.9 (SD = 0.7), mean gain = 1.1 (95% CI [0.65, 1.55]), $F(1, 34) = 24.31$, $p < .001$, $\eta^2p = .42$. The proportion of students who rated themselves "confident" or "very confident" in programming tangible

AI to perform simple tasks rose from 19% (n = 7) to 72% (n = 26). All four outcomes remained significant after the Holm–Bonferroni correction.

3.3. Students' Conceptions of AI

Students' conceptions of AI were collected before and after the intervention through a written open-ended question, "How does a robot or AI know what to do?", followed by short group interviews in which students explained their answers. These conception data came from students; they were separate from the teacher interviews. Each student's response was assigned to one category using the following criteria. Anthropomorphic: the response attributes feelings, intentions, wishes, or independent thinking to AI (e.g., "the robot wants to help," "AI thinks like a person"). Mechanistic: the response explains AI behavior through input, sensors, rules, data, or training, without attributing mental states (e.g., "it follows the rules it was given"). Mixed or unclear: the response contains both types of explanation, or neither. Two researchers coded all responses independently, and disagreements were resolved through discussion.

At pretest, 89% of students (n = 32) gave anthropomorphic explanations, 5.5% (n = 2) gave mechanistic explanations, and 5.5% (n = 2) gave mixed or unclear explanations. At posttest, anthropomorphic explanations decreased to 25% (n = 9), mechanistic explanations increased to 72% (n = 26), and 3% (n = 1) were mixed or unclear. Typical posttest mechanistic responses described AI as "a tool that can be programmed and follows instructions" and noted that "real AI can make mistakes if not trained with good data, but is still useful for helping humans."

The ethical scenario items showed smaller changes. When asked whether it is fair for a facial recognition system to perform better for people with lighter skin, the proportion of students who showed awareness of algorithmic bias increased from 33% (n = 12) to 44% (n = 16). At posttest, 42% of students (n = 15) showed deep reflection on ethical issues. A response was classified as deep reflection when it [(a) identified a fairness, bias, or privacy problem, (b) explained its cause, such as unbalanced training data, and (c) suggested a consequence or solution]. Responses meeting only criterion (a) were classified as awareness without deep reflection.

3.4. Teacher Implementation

Because only one teacher participated, the following themes are reported as findings from a single illustrative case. The quotations presented were selected because they expressed ideas that appeared in at least two of the three interviews and were consistent with the observation field notes, so they represent recurring patterns in the data rather than isolated remarks. Each quotation is labeled with the interview it came from.

Theme 1: High pedagogical confidence, low AI content knowledge. The teacher reported confidence in managing the class and adapting activities, but limited knowledge of how AI works. (Interview 3, post-intervention): "I feel very comfortable managing the classroom and adapting activities to different learning styles—that's what I do every day. However, when students ask deeper questions about how AI actually works, or why AI responses sometimes don't match expectations, I realize that I'm learning alongside them. I

have to say more often than usual, 'I don't know—let's figure it out together.'" The teacher said that this gap initially caused anxiety but later encouraged a collaborative classroom atmosphere in which students explored and experimented.

Theme 2: Technical challenges and infrastructure needs. Connectivity problems disrupted [three of the four] sessions, according to the observation notes. The teacher noted (Interview 2, mid-intervention): "When the 'AI for Oceans' app couldn't connect to the internet, it disrupted the flow of learning. Students became frustrated, and I had to switch to offline activities that required extra effort to explain." Device sharing was a second challenge. In Weeks 1 and 2, groups of three to four students shared one device, and observations showed that students who were waiting became disengaged while the student holding the device dominated. From Week 3, the teacher reduced the groups to one or two students per device. The teacher described the difference (Interview 3): "Magic happened when the children became interested in exploring the devices and how they worked. When they took turns, the students waiting became disengaged, and the one holding the device dominated the learning process."

Theme 3: Reduced technical anxiety through gradual mastery. The teacher's concerns about technical complexity decreased as they gained experience with AI for Oceans and the Sensor Cards. In the first interview, the teacher said: "I was worried I'd break something or that the technology would be too complicated." In the final interview, the teacher said: "AI technology is actually quite intuitive. This technology works in a way that mirrors how our brains process information. And it's very possible to represent this in a tangible form. After taking the time to learn it, I finally realized that this is something I can definitely explore further with other learning materials."

3.5. Integrating the Quantitative and Qualitative Findings

This subsection brings together the test results and the observation, student-response, and interview data. For each finding, it asks whether the qualitative evidence confirms, elaborates, or contradicts the quantitative result.

AI concept mastery: confirmed and elaborated. The large pre-to-post gain in AI concept mastery ($\eta^2p = .70$) and the rise in accurate explanations of machine learning (28% to 83%) were confirmed by the classroom observations. Students' talk changed during the sessions: in Sessions 1 and 2, students mostly described the system as "knowing" or "wanting" things; from Session 3 onward, the field notes record more references to sensors, rules, and training data. The observations also elaborated the test result by showing when understanding shifted. The clearest changes occurred when the system made a visible error, for example, when the decision tree trained with the Biased Data Deck failed, and students had to trace the error back to the data. At such moments, students described AI as "a tool that does what it's told, even if it sometimes makes mistakes." At the same time, the qualitative data set a limit on the test result. Students could explain what the system did, but the observation notes contain few instances of students explaining why a particular rule or feature was chosen. This suggests that the test gains reflect foundational rather than deep algorithmic understanding.

Computational thinking: partly confirmed, partly discordant. The gains in algorithmic thinking and decomposition matched what was observed in Module 2, where students had to place IF-THEN-ELSE tokens in a physical sequence and correct them when the agent acted wrongly. The lack of change in abstraction (Table 3) is consistent with the observation data: no activity required students to decide which features to leave out, and the field notes record few instances of such reasoning. In this case, the qualitative data help explain a non-significant quantitative result rather than contradict it.

Motivation and self-efficacy: confirmed, with a qualification. The increases in motivation and programming self-efficacy agree with the high proportion of students who found the activities interesting (94%) and with the teacher's reports of student engagement. However, the observations show that engagement was not uniform. When three to four students shared one device, the students who were waiting disengaged (Theme 2). The average gains may therefore conceal differences between students who handled the tools often and those who handled them rarely. The motivational and self-efficacy gains therefore occurred under conditions in which the unplugged materials carried much of the instruction.

Ethical understanding: confirmed. The small increase in bias awareness (33% to 44%) is consistent with the observation data, which show that discussions of bias were brief and usually took place at the end of a session. Taken together, the two strands indicate that the technical outcomes were supported by both the tests and the observations, whereas ethical understanding developed less in both.

3.6. The Role of Embodiment

Students showed a large gain in AI concept mastery and a shift from anthropomorphic toward mechanistic explanations of AI. The observations suggest that this shift occurred mainly when students could trace a visible chain from input, through a rule or model, to output.

These findings are consistent with embodied cognition theory, which holds that abstract knowledge is grounded in sensorimotor experience [17], and with embodied learning, in which physical action, perception, and cognition jointly build understanding [16]. One plausible explanation is that the tangible tools served as concrete referents [19], making normally invisible processes observable and manipulable. The observations further suggest that observable cause and effect, rather than physical novelty alone, may be what replaces intentional explanations of AI with mechanistic ones. Because the study had no comparison group, however, we cannot determine how much of the change is due to the physical tools themselves and how much to the AI content, the structured Explore–Explain–Apply–Reflect cycle, or the teacher's guidance. The pattern is also consistent with distributed cognition accounts, in which learning emerges from the interaction of learners, tools, and teaching practice rather than from a tool in isolation [32].

These findings add to calls to "open the black box" of AI for young learners [2, 4]. Future implementations should design tangible activities that make cause and effect visible, including moments when the system fails, and should pair them with explicit teacher questioning that links the physical activity to the underlying AI concept.

3.7. Scope and Limitations of the Tangible AI Approach

Students' explanations became more mechanistic, but they mostly described what the system did rather than why. Abstraction did not improve.

The pre-to-post improvements are consistent with the proposition that tangible AI can help elementary students build a foundational, mechanistic understanding of AI. They do not show that tangible AI is more effective than software-only instruction, because this study included no software-only comparison. Our results also resemble those of Zimmermann-Niefield et al. [6], whose students developed an intuitive understanding of machine learning but a limited grasp of the reasoning behind specific model operations. The instruments used here were not designed to assess such deeper reasoning.

For abstraction, three explanations are plausible. First, the concreteness of the materials may have kept students' attention on specific, visible features, whereas abstraction requires deciding which details to ignore. Research on concreteness fading suggests that concrete materials support initial understanding but need to be gradually replaced by more symbolic representations if learners are to generalize [33]. Second, no module taught abstraction explicitly. Third, the abstraction subscale has only five items, so the small change of 0.2 points is within the range that measurement error could produce.

Tangible AI may be most useful as an entry point to AI literacy and may need to be followed by more symbolic activities to support deeper technical understanding. Future curricula should include explicit abstraction tasks, for example, asking students to remove unnecessary features from a sensor-card set before training a model. Comparative studies are needed to establish whether tangible approaches add value over software-only instruction.

3.8. Social and Ethical Understanding

Section 3.3 showed a gap between technical and ethical outcomes. While 83% of students could accurately explain how machine learning works, bias awareness increased only from 33% to 44%, and 42% of students showed deep ethical reflection at posttest.

Several explanations are possible, and this design cannot distinguish between them. The first concerns instructional time. Explicit discussion of bias and privacy took about 5 to 10 minutes per session, roughly 8 to 17 percent of instructional time, which may not have been enough to develop ethical reasoning to the same degree as technical understanding. The second concerns sequencing: the Biased Data Deck was introduced only in Modules 3 and 4, giving students less time to revisit ethical ideas. The third concerns development: reasoning about fairness in abstract scenarios, such as facial recognition, may be more demanding for 10- to 11-year-old students than explaining a visible mechanism. The fourth concerns measurement: ethical understanding was assessed with a small number of scenario items, which may not have captured all relevant learning. The time explanation is the most plausible one in our view, but it should be read as a hypothesis for future studies, not as a demonstrated cause.

This gap is consistent with the concern raised by Selbst and Barocas [21] that technical understanding of AI, developed without sociotechnical perspectives, can leave learners able to use AI vocabulary without the critical awareness needed to recognize

harmful applications. It also supports Kong et al.'s [23] argument that ethics is better integrated into technical activities than added as a separate unit.

Future implementations could integrate ethics into the model-training activities themselves, for example by asking students to check a training dataset for group representation before training a model, or to predict the consequences of specific system errors. Studies that vary the time spent on ethics would show whether instructional time explains the gap.

3.9. Teacher Development and TPACK

The teacher showed strong pedagogical knowledge and enough technological knowledge to manage the tools, but limited AI content knowledge, and reported moving from anxiety to confidence over the four weeks. Because only one teacher participated, this profile is an illustrative case, not evidence about elementary teachers in general. It shows that this combination of strengths and gaps can occur and describes one teacher's experience of the intervention, but it cannot show how common this pattern is. The case is broadly consistent with studies reporting that many teachers have limited training in AI content despite general pedagogical confidence [11, 13]. The TPACK framework [10] helps explain why this matters: effective teaching requires an integrated understanding of technology, pedagogy, and content rather than separate competence in each, and this integrated knowledge develops through sustained work on authentic tasks. The teacher managed the content gap by adopting an exploratory, co-learning stance. This worked in this case, but we cannot conclude that it would work equally well for teachers with different dispositions, class sizes, or levels of support.

A co-learning stance is one strategy worth testing at a larger scale, alongside, not instead of, professional development that directly targets AI content knowledge. Because the teacher reported pressure to cover mathematics, language, and science standards, positioning tangible AI as a way to deepen existing subject content, rather than as an extra subject, may support more sustainable adoption [2], [22]. Schools should also plan for reliable connectivity and a device ratio of no more than one or two students per device, since both conditions affected engagement in this case.

4. CONCLUSION

This study examined a four-week tangible AI intervention with 36 fifth-grade students and one teacher in a single elementary school. The intervention was associated with substantial pre-to-post gains in AI concept mastery and computational thinking, and with smaller but meaningful gains in learning motivation and programming self-efficacy. The proportion of students who could accurately explain how machine learning works rose from 28% to 83%, and anthropomorphic descriptions of AI fell from 89% to 25%. Two outcomes improved less. The abstraction dimension of computational thinking did not improve, and ethical understanding developed more slowly than technical understanding: bias awareness rose only from 33% to 44%. The teacher case showed strong pedagogical confidence but limited AI content knowledge, and connectivity and device-sharing problems constrained implementation.

The main contribution of this study is preliminary evidence that tangible, physically manipulable activities can help elementary students move from anthropomorphic toward more mechanistic understandings of AI. The study also examined cognitive, affective, conceptual-change, and teacher-implementation outcomes within a single design. For practice, the findings suggest that tangible AI is a feasible entry point to AI literacy in regular classrooms when it is combined with a structured lesson cycle, explicit ethics content, reliable connectivity, and a low student-to-device ratio.

These findings should be read as preliminary, not as confirmatory evidence of effectiveness. The single-group design without a control condition cannot rule out maturation, testing, or novelty effects. Because the tangible materials and the Explore–Explain–Apply–Reflect cycle were implemented together, their separate contributions cannot be determined. The sample was small (N = 36) and came from one school with one teacher, which limits statistical power and generalizability, and the teacher findings represent a single illustrative case. The intervention was short (four hours), the AI Concept Mastery Test was researcher-developed, and no delayed follow-up assessment was conducted, so the durability of the observed changes is unknown.

Future research should use controlled designs with random assignment or matched comparison groups, larger samples across multiple schools and teachers, and delayed follow-up assessments. Studies comparing tangible, software-based, and hybrid AI instruction would show whether the observed changes are due to the physical nature of the tools, to the AI content, or to novelty. Dismantling designs that vary the tangible materials and the lesson cycle independently would clarify the role of each. Further work should also test whether ethics integrated throughout technical activities and explicit abstraction tasks strengthen the outcomes that improved least, and how sustained professional development can build teachers' AI content knowledge**.**

Acknowledgment

This research is funded by Universitas Pendidikan Indonesia through the research scheme/program of Riset Afirmasi Dosen with Contract No. 222 /U[0N40.F7/PT.01.01/2026).

REFERENCES

[1] D. Long and B. Magerko, "What is AI Literacy? Competencies and Design Considerations," in *International Conference on Human Factors in Computing Systems (CHI)*, 2020.

[2] D. Touretzky, C. Gardner-McCune, F. Martin and D. Seehorn, "Envisioning AI for K-12: What Should Every Child Know about AI?," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[3] R. Williams, S. Ali, N. Devasia, D. DiPaola, J. Hong, S. P. Kaputsos, B. Jordan and C. Breazeal, "AI + Ethics Curricula for Middle School Youth: Lessons Learned from Three Project-Based Curricula," *International Journal of Artificial Intelligence in Education,* vol. 33, 2022.

[4] S. Druga and A. J. Ko, "How do children's perceptions of machine intelligence change when training and coding smart programs?," in *IDC '21: Proceedings of the 20th Annual ACM Interaction Design and Children Conference*, 2021.

[5] P. H. Kahn, T. Kanda, H. Ishiguro, N. G. Freier, R. L. Severson, B. T. Gill, J. H. Ruckert and S. Shen, ""Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot," *Developmental Psychology,* vol. 48, no. 2, pp. 303-314, 2012.

[6] A. Zimmermann-Niefield, M. Turner, B. Murphy, S. K. Kane and R. B. Shapiro, "Youth Learning Machine Learning through Building Models of Athletic Moves," in *IDC '19: Proceedings of the 18th ACM International Conference on Interaction Design and Children*, 2019.

[7] T. C. Bell, J. Alexander, I. Freeman and M. Grimley, "Computer Science Unplugged: school students doing real computing without computers," *New Zealand Journal of Applied Computing and Information Technology,* vol. 13, no. 1, pp. 20-29, 2009.

[8] S. Y. Lye and J. H. L. Koh, "Review on teaching and learning of computational thinking through programming: What is next for K-12?," *Computers in Human Behavior,* vol. 41, 2014.

[9] S. Druga, S. T. Vu, E. Likhith and T. Qiu, "Inclusive AI literacy for kids around the world," in *FL2019: Proceedings of FabLearn 2019,* 2019.

[10] D. Petko, P. Mishra and M. J. Koehler, "TPACK in context: An updated model," *Computers and Education Open,* vol. 8, 2025.

[11] A. Castro, B. Díaz, C. Aguilera, M. Prat and D. Chávez-Herting, "Identifying Rural Elementary Teachers' Perception Challenges and Opportunities in Integrating Artificial Intelligence in Teaching Practices," *Sustainability,* 2025.

[12] I. H. Y. Yim, "A critical review of teaching and learning artificial intelligence (AI) literacy: Developing an intelligence-based AI literacy framework for primary school education," *Computers and Education: Artificial Intelligence,* vol. 7, 2024.

[13] M. Yue, M. S.-Y. Jong and D. T. K. Ng, "Understanding K–12 teachers' technological pedagogical content knowledge readiness and attitudes toward artificial intelligence education," *Education and Information Technologies,* vol. 29, 2024.

[14] M. Przybylla and R. Romeike, "Physical Computing and its Scope - Towards a Constructionist Computer Science Curriculum with Physical Computing," *Informatics in Education,* vol. 13, no. 2, 2014.

[15] H. Darabipourshiraz, M. Bhat and D. Long, "Introducing AI without computers: Hands-on literacy and ethical sense-making for young learners.," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing System*, 2025.

[16] A. N. Antle and A. F. Wise, "Getting down to details: Using theories of cognition and learning to inform tangible user interface design," *Interacting with Computers,* vol. 25, no. 1, 2013.

[17] M. Wilson, "Six views of embodied cognition," *Psychonomic Bulletin & Review,* vol. 9, pp. 625-636, 2002.

[18] I. Levin, A. L. Semenov and M. Gorsky, "Smart Learning in the 21st Century: Advancing Constructionism Across Three Digital Epochs," *Education Sciences,* vol. 15, no. 1, 2025.

[19] K. M. Clark, "Embodied Imagination: Lakoff and Johnson's Experientialist View of Conceptual Understanding," *Review of General Psychology,* vol. 28, no. 2, 2024.

[20] O. Parnafes and A. Disessa, "Relations between Types of Reasoning and Computational Representations," *International Journal of Computers for Mathematical Learning,* vol. 9, 2004.

[21] A. D. Selbst and S. Barocas, "The Intuitive Appeal of Explainable Machines," *Fordham Law Review,* vol. 87, no. 3, 2018.

[22] D. T. K. Ng, J. K. L. Leung, S. K. W. Chhu and M. S. Qiao, "Conceptualizing AI literacy: An exploratory review," *Computers and Education: Artificial Intelligence,* vol. 2, 2021.

[23] S.-C. Kong, W. M.-Y. Cheung and G. Zhang, "Evaluating an Artificial Intelligence Literacy Programme for Developing University Students' Conceptual Understanding, Literacy, Empowerment and Ethical Awareness," *Educational Technology & Society,* 2022.

[24] Ö. Korkmaz, R. Çakir and M. Y. Özden, "A validity and reliability study of the computational thinking scales (CTS)," *Computers in Human Behavior,* vol. 72, pp. 558-569, 2017.

[25] S. M. Glynn, P. Brickman, N. Armstrong and G. Taasoobshirazi, "Science motivation questionnaire II: Validation with science majors and non-science majors," *Journal of Research in Science Teaching,* vol. 48, no. 10, 2011.

[26] M.-J. Tsai, C.-Y. Wang and P.-F. Hsu, "Developing the computer programming self-efficacy scale for higher education students," *Journal of Educational Computing Research,* vol. 56, no. 8, 2018.

[27] W. R. Shadish, T. D. Cook and D. T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Boston: Houghton Mifflin, 2002.

[28] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan and K. Hoagwood, "Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research," *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 42, 2015.

[29] S. Lee and D. K. Lee, "What is the proper way to apply the multiple comparison test?," *Korean Journal of Anesthesiology*, vol. 71, no. 5, 2018.

[30] H. Kang, "Sample Size Determination and Power Analysis Using the G*Power Software," *Journal of Educational Evaluation for Health Professions*, vol. 18, 2021.

[31] V. Braun and V. Clarke, "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, 2008.

[32] M. Sidji, W. Smith and M. J. Rogerson, "Adopting the Theory of Distributed Cognition for Human-AI Cooperation," in *OzCHI '24: Proceedings of the 36th Australasian Conference on Human-Computer Interaction*, 2024.

[33] S. Suh, "Using Concreteness Fading to Model & Design Learning Process," in *ICER '19: Proceedings of the 2019 ACM Conference on International Computing Education Research*, 2019.

[34] A. Lindner and M. Berges, "Can you explain AI to me? Teachers' pre-concepts about artificial intelligence," in *2020 IEEE Frontiers in Education Conference (FIE)*, Sweden, 2020.