Journal of Information Technology Education: Innovations in Practice

An Official Publication of the Informing Science Institute
InformingScience.org

JITEiip.org

Volume 25, 2026

RETAINING ARTIFICIAL INTELLIGENCE LITERACY: EVIDENCE FROM A PROJECT-BASED UNDERGRADUATE COURSE

Senhao Xiong	Khon Kaen University, Khon Kaen, Thailand	senhao.x@kkumail.com
Yushuang Wu*	Khon Kaen University, Khon Kaen, Thailand	yushuang.w@kkumail.com
HaoHong Nie	Khon Kaen University, Khon Kaen, Thailand	haohong.n@kkumail.com
KuiHua Li	Chongqing Three Gorges Medical College, Chongqing, China	kuihua.l@kkumail.com

* Corresponding author

ABSTRACT

Aim/Purpose	This paper examines whether a short project-based Artificial Intelligence (AI) literacy course is associated with larger gains in applying AI concepts, regulating AI-supported problem solving, and reasoning ethically among first-year undergraduates, and whether any advantages remain detectable at a four-week short-delay follow-up.
Background	Artificial Intelligence literacy provision is expanding in higher education, but many short interventions are evaluated using single-group designs and only immediate posttests. Comparative evidence on short-delay retention in non-Computer Science (non-CS) or non-specialist cohorts remains limited, and existing evidence has not sufficiently separated applied concepts, strategy regulation, and ethical awareness as distinct literacy strands.
Methodology	A quasi-experimental mixed-methods design compared an intervention group (n = 180) with a contemporaneous comparison group (n = 180) from the same university and study stage. Both groups completed measures at pretest, immediate posttest, and approximately four weeks after course completion. The quantitative strand was integrated with 155 reflective essays and focus-group interviews with 19 students from the intervention cohort.

Accepting Editor Maria Joao Ferreira | Received: May 19, 2026 | Revised: July 4, 2026 | Accepted: August 7, 2026.
Cite as: Xiong, S., Wu, Y., Nie, H., & Li, K. (2026). Retaining artificial intelligence literacy: evidence from a project-based undergraduate course. *Journal of Information Technology Education: Innovations in Practice, 25,* Article 39. https://doi.org/10.28945/5872

Contribution	The study contributes strand-specific evidence on the short-delay retention profile of a compact project-based AI literacy course. Its contribution is not simply that an AI literacy course was associated with improvement, but that applied concept use, strategy regulation, and ethical awareness exhibited distinct patterns of persistence following the same intervention. This distinction extends prior AI literacy intervention work by separating immediate gains from short-delay retention across literacy dimensions.
Findings	The intervention group showed larger differential gains in applied AI concepts and strategy regulation at both posttest and four-week follow-up. Ethical awareness showed a statistically significant immediate comparative advantage, but the follow-up contrast was smaller, included zero in its confidence interval, and was not statistically reliable.
Recommendations for Practitioners	Short AI literacy courses should embed concepts into project decisions and provide students with repeated opportunities to plan, monitor, revise, and justify AI-supported work. Ethical reasoning should be revisited after course completion through subsequent cases, critique tasks, or discipline-specific applications, rather than being treated as a one-time element.
Recommendations for Researchers	Future studies should use comparison groups, delayed follow-ups, artifact-based evidence, and behavioral measures to test which course components sustain different strands of AI literacy across time.
Impact on Society	The findings indicate that compact, project-based AI literacy courses are associated with applied conceptual gains and greater self-reported strategic awareness among non-specialist undergraduates, while also showing that responsible AI use is less likely to consolidate without continued reinforcement.
Future Research	Further research should test longer follow-up intervals, randomized or matched designs, component-level reinforcement models, behavioral measures of regulation, and transfer across institutions, disciplines, and student populations.

Keywords	artificial intelligence literacy, project-based learning, non-computer science students, first-year undergraduates, quasi-experimental design, mixed-methods study, retention effects

INTRODUCTION

Artificial Intelligence (AI) systems have moved from specialist settings into ordinary study, work, and civic life. First-year undergraduates now encounter AI tools before many of them have learned how to judge whether a tool fits a problem, how to interpret its output, or how to reason about the consequences of relying on it. Research on technological change suggests that automation restructures rather than removes tasks, making adaptation and reskilling a persistent educational concern (Acemoglu & Restrepo, 2019). Work on human-centered and collaborative AI places similar emphasis on human judgment, oversight, and co-production in AI-supported activity (Dellermann et al., 2019; Duan et al., 2019; Shneiderman, 2020).

In this article, AI literacy is treated as a set of competencies for non-specialists combining conceptual understanding, situated application, strategic regulation, and ethical judgment (Long & Magerko, 2020; UNESCO, 2021). Non-specialists are defined here as undergraduate learners outside Computer Science (CS) and specialized AI programs who need to use, evaluate, and justify AI-supported work without being trained as model developers. Reviews of AI literacy provision identify persistent gaps in opportunities for non-CS students to engage with AI concepts, authentic use, and responsible

practice beyond isolated exposure (Almatrafi et al., 2024; Laupichler et al., 2022; Zawacki-Richter et al., 2019).

The present study addresses a more specific gap in that literature. It examines not only whether a compact project-based AI literacy course is associated with comparative gains relative to a contemporaneous comparison group, but also whether those gains persist selectively across different strands of AI literacy after a short delay. Three strands are treated as primary outcomes: applied AI concepts, strategy regulation during AI-supported problem solving, and ethical awareness in AI use. The study, therefore, moves beyond a general question of whether an AI literacy course improves post-course scores and asks which gains remain detectable four weeks later and which appear more fragile.

This distinction matters for curriculum design. Short AI literacy courses are increasingly adopted as a scalable provision for non-specialist students (Laupichler et al., 2022; Long & Magerko, 2020), yet most evaluation evidence concerns the instructional period itself. If different AI literacy strands respond differently to a compact intervention, with some persisting through repeated project action and others requiring later reinforcement, then curriculum design must move beyond a blanket conclusion that short-form project-based learning works. The practical issue is how to sequence concepts, tool use, strategy regulation, and ethical reasoning across courses so that short interventions become a starting point rather than a complete solution. The study uses a comparison-group, three-wave, mixed-methods design to examine that issue.

The quantitative strand drew on a dual-group, three-wave dataset (N = 360; 180 intervention, 180 comparison). Because allocation was not randomized, the design is quasi-experimental; observed group differences are interpreted as associations between course participation and outcome trajectories rather than as isolated causal effects. The qualitative strand drew on 155 reflective essays and semi-structured focus-group interviews with 19 students from the intervention cohort. It was used to explain quantitative patterns rather than to establish comparative effectiveness independently. Taken together, these sources allow the study to ask whether any differential advantage remained visible after the course had ended and what course features students associated with those changes.

BACKGROUND OF THE STUDY

AI LITERACY FOR NON-SPECIALIST UNDERGRADUATES

AI literacy for non-specialists does not require every learner to become a programmer or model developer. In this study, the term refers to students outside CS and specialized AI programs who still need to make informed decisions about AI-supported study, work, and civic activity. Such literacy requires a working grasp of how AI systems are used, what kinds of problems they can and cannot address, how data and model choices shape outputs, and why ethical decisions remain part of practical AI use. Across policy and research frameworks, conceptual understanding remains a recurring foundation because it underpins informed evaluation, meaningful engagement, and responsible decision-making about AI (Long & Magerko, 2020; OECD, 2022; Touretzky et al., 2019; UNESCO, 2021). Recent reviews make the same point in practical terms, stressing that non-specialists need a working grasp of principles and workflows even when they are not expected to build novel algorithms from scratch (Almatrafi et al., 2024; Laupichler et al., 2022).

The conceptual strand in this study concerns understanding widely used machine learning (ML) approaches, such as supervised learning, unsupervised learning, classification, regression, and neural networks, and recognizing when these approaches map onto academic and workplace problems. The instructional emphasis is on using and evaluating models as data-driven tools rather than on advanced programming or mathematical derivations. This emphasis reduces prerequisite demands and supports delivery to large non-CS cohorts, which is critical for broad adoption in higher education (Laupichler et al., 2022; Long & Magerko, 2020).

Conceptual knowledge alone is insufficient. Students also need to solve authentic problems in ways that involve planning, monitoring, and revision. The present study treats this capacity as strategy regulation in AI-supported problem solving. Drawing on self-regulated and metacognitive learning perspectives, the process can be described as a sequence of problem framing, selection of relevant AI ideas and candidate routes, planning and execution of an AI-based approach, examination of results, and revision when outputs do not fit the problem (Schraw et al., 2006; Winne & Azevedo, 2014). Effective performance requires awareness of what one knows, strategic selection of methods, and ongoing monitoring to ensure that data, tools, and outputs remain aligned with the intended task.

Ethical awareness forms the third focal strand. Responsible AI practice requires students to consider benefits, harms, fairness, human autonomy, privacy, and accountability when AI-supported solutions are proposed or deployed. Ethical guidance converges around principles such as fairness, beneficence, and human autonomy. However, these principles matter most when students learn to apply them to concrete decisions about data collection, tool choice, output interpretation, and deployment boundaries (European Commission, 2019; Floridi et al., 2018; Hagendorff, 2020; Jobin et al., 2019; OSTP, 2022; UNESCO, 2021). In this study, those principles were embedded in project briefs and reflection prompts so that ethical reasoning accompanied the work rather than appearing only as a detached topic.

Project-Based Learning as a Design Logic for AI Literacy

Project-based learning (PBL) is a pedagogical approach in which students develop a tangible product or solution through extended inquiry (Prince & Felder, 2006; J. W. Thomas, 2000). It draws on situated learning principles such as collaboration, work on problems grounded in real contexts, and cycles of reflection and revision (Herrington & Oliver, 2000). In AI education, these features are especially relevant because conceptual understanding, data judgment, model evaluation, and ethical justification have to be coordinated rather than studied in isolation.

A typical PBL sequence begins with a driving problem, moves through goal setting and inquiry, and requires learners to translate abstract ideas into testable artifacts (Bell, 2010; Kokotsaki et al., 2016; Krajcik & Shin, 2014). In the context of AI literacy, students who design and implement AI-supported solutions to self-identified problems must make decisions about whether AI is needed, which data are suitable, which method fits the problem, how outputs should be evaluated, and what safeguards are necessary. These decisions create repeated opportunities for concept application and strategy regulation. They also make ethical reasoning visible because students must justify not only whether a tool works, but whether its use is appropriate.

The broader course design also drew on empowerment-related ideas such as self-efficacy, meaningfulness, and perceived impact (Bandura, 1997; Ryan & Deci, 2000; Spreitzer, 1995; K. W. Thomas & Velthouse, 1990; Tierney & Farmer, 2002). These constructs informed the low-barrier and student-centered design of the course. They are treated in the present article as part of the instructional context rather than as primary comparative outcomes because the strongest matched data across groups were available for applied concepts, strategy regulation, and ethical awareness.

A further design issue concerns retention. Short AI literacy interventions can raise immediate awareness, but immediate posttests do not show whether students retain applied concepts, regulatory strategies, or ethical judgment after the instructional setting ends. Retention is especially relevant to ethical awareness because responsible AI use depends on repeated contextual judgment, critique, and transfer rather than on a single exposure to principles. This makes reinforcement across later coursework and disciplinary contexts a central curriculum question, not only an evaluation issue.

Evidence from short AI literacy interventions underscores the need for delayed, strand-specific evaluation. Recent reviews report rapid growth in AI literacy curricula, workshops, and course activities, but much of the evaluation evidence remains concentrated on immediate post-course perceptions, self-reported confidence, or single-group learning gains (Almatrafi et al., 2024; Laupichler et al.,

2022). A recent evaluation of a project-based AI literacy course further supports the value of authentic projects for AI learning (Kong et al., 2024). However, the literature still offers limited comparative evidence on whether the use of applied concepts, strategy regulation, and ethical awareness is retained in the same way after instruction ends. This gap justifies examining retention by strand rather than treating AI literacy as a single undifferentiated outcome.

Research Gap and Research Questions

Existing AI literacy frameworks prioritize different combinations of knowledge, use, confidence, and responsible practice (Kong et al., 2024; Laupichler et al., 2022; Long & Magerko, 2020; UNESCO, 2021). Empirical intervention studies have increasingly examined AI literacy courses and workshops, but the evidence base remains uneven for short-course interventions with non-CS learners, comparison groups, and delayed follow-up measurement. Less is known about whether short non-CS interventions are associated with comparative gains in AI-supported problem solving as strategy regulation in action, or whether such gains remain visible after a short delay. Many course evaluations report immediate gains, but immediate posttests cannot show whether students retain applied competencies beyond the instructional period or whether different literacy strands persist in the same way.

The study's novelty lies in its focus on strand-specific retention. Rather than treating AI literacy as a single aggregate outcome, the study separates applied concept use, strategy regulation, and ethical awareness, then examines each at pretest, posttest, and four-week follow-up against a contemporaneous comparison group. This structure shows not only whether students improved, but also which dimensions of AI literacy showed short-term persistence and which required further instructional reinforcement.

The present study was conducted as part of a university-wide AI literacy initiative at a comprehensive university in mainland China. The intervention consisted of a stand-alone 18-hour course delivered across six three-hour sessions. A contemporaneous comparison group did not receive the course during the study period but completed the same outcome measures on the same schedule. The design addresses the following research questions:

(1) Is the intervention group associated with larger gains than the comparison group in applying AI concepts to authentic problems at posttest and follow-up?

(2) Is the intervention group associated with larger gains than the comparison group in AI-supported problem-solving strategies, particularly planning, monitoring, and revision, at posttest and follow-up?

(3) Is the intervention group associated with larger immediate gains in ethical awareness, and does any comparative advantage remain at the four-week follow-up?

(4) What features of the project-based course do students identify as supporting these changes?

METHODOLOGY

Study Context and Participants

The study was part of a university-wide AI literacy initiative at a comprehensive university in mainland China. The intervention was a non-credit-bearing, stand-alone, 18-hour, project-based AI literacy course delivered over six three-hour sessions. The target population was first-year undergraduate learners outside Computer Science and specialized AI programs. Prior programming experience was recorded but was not used as an exclusion criterion, because the course was designed for non-specialist learners with varied starting points. Participation was voluntary, and students could withdraw at any time without penalty.

For the controlled quantitative analysis, 360 first-year undergraduates were included: 180 in the intervention group and 180 in a contemporaneous comparison group, both drawn from the same

university and the same stage of study. Students were invited through university course announcements and class-level notices. No course grade, credit, financial payment, or other academic reward was attached to research participation. The sampling strategy was convenience-based, with the comparison cohort selected from existing first-year student groups who had not enrolled in the project-based AI literacy course during the study period but agreed to complete the same survey and test schedule. Inclusion criteria were first-year undergraduate status, consent to participate, and completion of the baseline measure. Students were excluded from the analytic sample if they did not provide consent, were not in the target first-year cohort, or had no usable baseline outcome data. Both groups completed the same outcome measures at pretest, immediate posttest, and approximately four-week follow-up. The qualitative strand drew on the intervention cohort only: 155 students submitted self-reflective essays, and 19 students participated in focus-group interviews after the course. Table 1 summarizes the group composition, demographic characteristics, and prior programming and AI experience of the quantitative analytic sample.

Table 1. Characteristics of the quantitative analytic sample (N = 360)

Characteristic	n	%
Intervention group	180	50.0
Comparison group	180	50.0
Female	209	58.1
Male	151	41.9
Prior programming course(s)	79	21.9
No prior programming	281	78.1
Used AI tools before this study	312	86.7
No prior AI exposure	48	13.3

Note. Group-level baseline outcome values are reported in Table 3.

Comparison Condition and Timing

The comparison cohort should be understood as a contemporaneous comparison condition rather than a no-exposure control group. These students were drawn from the same university and study stage as the intervention cohort. They did not participate in the 18-hour project-based AI literacy course, and no structured project-based AI literacy instruction was provided to them by the course team during the study period. However, because AI tools were already widely available and the majority of participants reported prior AI tool use at baseline (86.7%), incidental AI engagement in regular coursework or everyday study could not be fully excluded. The baseline AI exposure item asked whether students had used AI tools before the study; it was used descriptively to characterize the sample rather than as a basis for excluding participants. The estimated group-by-time contrasts therefore represent the added value associated with participation in the structured project-based course relative to the comparison cohort's ordinary learning context, not relative to a condition of zero AI exposure.

The follow-up was scheduled approximately four weeks after course completion. Minor natural variation in the actual interval was retained as part of the implementation context, so the follow-up is interpreted as a short-delay retention point rather than an exact fixed-day measurement.

Course Design and Instructional Materials

The instructional design combined short conceptual briefings with guided tool use, project planning, data work, peer discussion, iterative revision, and structured reflection. Students learned core ideas in machine learning (ML) and deep learning, explored no-code and low-code tools, and worked in

teams on a self-defined project. The course aimed to move students beyond general awareness of AI toward situated use, judgment, and justification.

Ethical principles for AI use, especially beneficence, fairness, and human autonomy, were embedded in the project cycle rather than taught only as a detached add-on. Students were expected to justify tool choices, data decisions, and likely consequences as part of the work itself. One team, for example, developed an AI posture coach for squats. The project required students to consider whether an image-classification mode or a pose-estimation mode better matched the task, how training data should be collected, and what limitations should be disclosed before deployment.

Across the course, students engaged in approximately eight hours of collaborative project work. They produced AI-supported solutions to self-defined problems, collected data, trained or configured models, presented their artifacts, and reflected on challenges, feasible improvements, and ethical considerations. The ethical discussion addressed not only the final artifact but also upstream decisions regarding data collection, processing, and interpretation. Table 2 maps the course dimensions to instructional components and corresponding evidence sources.

Table 2. Course design mapped to focal strands, activities, and evidence sources

Dimension	Meaning	Course components	Evidence and measures
Applied AI concepts	Understanding how AI concepts and workflows apply to practical problems	Mini-lectures and tool demos; five-step machine learning (ML) workflow embedded in tasks and assignments: problem, data, preprocessing, training, and inference	AI concepts for problem-solving test: pretest, posttest, follow-up
Strategy regulation	Planning, monitoring, evaluating, and revising while applying AI	PBL project scaffolds; design brief and planning sheet; mid-project checkpoints; model evaluation and iteration; end-of-project rationale	Metacognitive strategies in PBL scale: pretest, posttest, follow-up; reflective essays
Empowerment context	Confidence, meaningfulness, and perceived value in using AI	Low-barrier tools; instructor coaching; peer critique; staged showcases; emphasis on transferability to study and work	Intervention-only qualitative evidence retained for course interpretation
Ethical awareness	Reasoning about benefits, risks, fairness, and human autonomy in AI use	Three-principle ethics introduction; ethics checklist and data card; "Is AI needed?" prompt in proposals and presentations	AI ethics survey; pretest, posttest, follow-up; ethics sections in reports and presentations

Note. The course framework informed course design. The comparative analyses focused on three outcomes measured in both groups: applied AI concepts, strategy regulation, and ethical awareness.

MEASURES

Applied AI concepts were assessed using a 7-item AI Concepts for Problem Solving test, developed to align with the course objectives and the AI literacy framework used in the study. Each item presented a short, authentic scenario, and students selected the most appropriate AI approach from four options. Items were scored dichotomously as correct or incorrect, so total scores represented the number of correct responses and ranged from 0 to 7. The items sampled the use of machine-learning and deep-learning ideas in realistic problem contexts. The instrument targeted application rather than rote recall; students were asked, for example, to select an appropriate method for clustering student behavior patterns, to identify suitable training data for pronunciation detection, and to diagnose

overfitting from training and validation accuracy curves. The seven test items are provided in Appendix A.

Strategy regulation was measured using a 4-item Likert-type scale targeting four observable aspects of AI-supported problem-solving: defining the problem before tool use, monitoring whether outputs matched task goals, revising prompts/data/methods when results were weak, and comparing alternative AI approaches before deciding which result to use. Ethical awareness was measured with a 6-item Likert-type scale covering beneficence, fairness, and human autonomy in the use of AI for practical tasks. Both scales used a five-point response format from 1 (strongly disagree) to 5 (strongly agree), yielding mean scores in that range. The scales were constructed to match the focal course strands and were reviewed for content alignment before administration. The strategy regulation and broader empowerment items are provided in Table B1 (Appendix B), and the ethics items are provided in Table C1 (Appendix C). All three instruments were administered to both groups at pretest, posttest, and follow-up.

Internal consistency was examined in the primary analytic sample (N = 360) at each measurement wave. For the objective AI Concepts for Problem Solving test, reliability was assessed using the Kuder-Richardson Formula 20 (KR-20), an internal consistency index for dichotomously scored items, and item functioning was examined using item difficulty and corrected item-total correlations. The seven-item test deliberately sampled different AI task families, including clustering, classification, regression, training-data selection, overfitting diagnosis, and multimodal detection, rather than repeating items from a single narrow content domain. KR-20 estimates for this instrument, therefore, need to be interpreted in light of construct breadth and test brevity, and the test is treated as a brief, scenario-based indicator of applied concept use rather than as a broad, standardized achievement scale. For the two Likert-type scales, internal consistency was assessed using Cronbach's alpha and McDonald's omega, with mean inter-item correlations reported as an additional index of scale coherence. These indices are reported in Appendix E. The two survey scales are treated as focused indicators of perceived strategy regulation and ethical awareness.

To supplement the self-report measure of ethical awareness, a random sample of 40 intervention-cohort project reports from the full set of 180 reports (22.2%) was double-rated with a three-dimensional rubric covering autonomy, beneficence, and fairness. Two trained raters independently scored each report on a 0-3 scale for each dimension, and the averaged rubric scores were used for the convergent analysis. Inter-rater reliability was high, as indicated by the intraclass correlation coefficient, ICC(2,1) = 0.86, 95% CI [0.79, 0.91], and averaged rubric scores showed a moderate positive association with ethics survey gains, $r = .43$, $p < .01$. These artifact ratings were used as convergent evidence rather than as a separate primary outcome. The scoring rubric is provided in Table F1 (Appendix F).

Baseline Equivalence, Data Quality, and Missing Data

Baseline equivalence was examined before outcome modeling and is treated as part of the design argument rather than as a passing descriptive check. The two cohorts were drawn from the same university and study stage, and the observed baseline differences on the three focal outcomes were small. Baseline standardized mean differences ranged from -0.05 to -0.09, which reduces the plausibility that the observed comparative patterns are a simple artifact of large initial outcome imbalance. Nonetheless, because allocation was not randomized, small unmeasured baseline differences in motivation, prior AI interest, or concurrent learning activity cannot be excluded; the baseline equivalence check addresses only the observed outcome levels.

Data completeness was screened before model estimation. Participants with incomplete waves were retained in the mixed-model analysis so that all available observations contributed to estimation. Sensitivity checks using covariate-adjusted models and complete-case re-estimation did not alter the direction of the core findings. These checks do not eliminate the limitations of a non-randomized design, but they reduce the likelihood that the reported pattern is driven solely by baseline outcome differences or by the treatment of missing waves.

Qualitative Materials and Analysis

The qualitative strand drew on 155 post-course reflective essays and semi-structured focus-group interviews with 19 students from the intervention cohort. Essay submissions accounted for 86.1% of the intervention cohort. Focus-group participants were recruited from the same intervention cohort after course completion; the available records do not establish an exact overlap between essay writers and interview participants. The qualitative materials were used to explain the quantitative patterns rather than to make direct claims about comparative effectiveness across groups.

The interview guide asked students to describe what they learned about applying AI to real problems, how the project helped them plan and monitor their work, how they made data-driven decisions, how they considered ethical principles, and where they expected to apply AI after the course. The guide is provided in Appendix D. The qualitative materials were originally written or conducted in Chinese. The illustrative quotations reported in the Results section were translated into English by the authors, and identifying details were removed before reporting. The qualitative analysis followed a thematic approach informed by Braun and Clarke (2006). Coding combined deductive categories from the course framework and research questions with inductive refinement from the reflective essays and interview transcripts. One researcher conducted initial coding across the full qualitative corpus, and a second researcher reviewed the codebook, theme boundaries, and a sample of coded materials to check whether the data supported the final themes. Disagreements about code placement and theme naming were resolved through discussion and revision of theme definitions. Because the qualitative strand was used for explanatory integration rather than for stand-alone theory building, the analysis did not use a formal kappa coefficient as an independent reliability statistic. The procedure moved through familiarization, initial coding, code consolidation, theme review, and theme naming. The final themes were used in a joint display to connect the quantitative trajectories with students' descriptions of the course processes that supported or limited change. Theme frequencies in Table 6 are reported as indicators of thematic salience, not as measures of effect size or comparative effectiveness.

Statistical Analysis

For each primary outcome, group-by-time patterns were estimated using linear mixed-effects models with fixed effects for group, time point, and their interaction, together with a participant-level random intercept to account for repeated measurements. Time point was treated as a three-level factor: pretest, posttest, and follow-up, with pretest as the reference point. The model can be summarized as outcome = group + time + group x time + participant random intercept. The interaction terms captured whether the change over time differed between the intervention and comparison groups.

Model-based estimates are reported as unstandardized coefficients (b) with standard errors, 95% confidence intervals, and p-values. The reported b values are group-by-time interaction estimates: they represent the intervention group's differential gain relative to the comparison group at each time point, not raw within-group change scores. The analysis focused on two planned contrasts for each outcome: the difference in pre-to-post change between groups and the difference in pre-to-follow-up change between groups. Standard errors and p-values were taken from the same mixed-model contrast output used to estimate the coefficients, and the analysis scripts were prepared to preserve the exact model specification and contrast definitions. To aid practical interpretation, standardized differential effects (Hedges' g) were computed from observed group differences in change scores using the pooled baseline standard deviation and a small-sample correction. Statistical significance was evaluated at $\alpha = .05$.

All analyses were conducted in R (Version 4.3.2; R Core Team, 2023). Missing outcome data were handled within the mixed-model framework under a missing-at-random assumption, so participants with incomplete later waves were not removed from the primary repeated-measures analysis after baseline inclusion. Complete-case re-estimation was used as a sensitivity check, and covariate-adjusted models were used to check whether the direction of the main findings changed after

accounting for observed baseline information. Model diagnostics included inspection of residual distributions and fitted values for major departures from normality or heteroscedasticity. The qualitative strand was analyzed separately and then integrated with the quantitative findings through a joint display. The purpose of this integration was explanatory: the qualitative materials clarified why concept application, strategy regulation, and ethical awareness changed in the quantitative patterns observed. Additional reporting details on the statistical models, planned contrasts, missing-data treatment, sensitivity checks, and qualitative coding procedure are summarized in Appendix G.

RESULTS

Baseline means were closely aligned across groups on all three outcomes. The analysis examined whether the intervention group showed greater change than the comparison group and whether any advantage remained visible at follow-up. Figure 1 provides a visual summary of the three outcome trajectories and shows the common posttest-to-follow-up drop-off in the intervention group, especially for ethical awareness. Concept application, corresponding to Research Question 1, showed the largest and most durable comparative gains. Strategy regulation, corresponding to Research Question 2, also improved clearly. Ethical awareness, corresponding to Research Question 3, moved in the same direction immediately after the course, but the follow-up contrast was smaller and not statistically reliable at $\alpha = .05$. Descriptive statistics are shown in Table 3, and planned mixed-model contrasts with standardized differential effects are shown in Table 4.

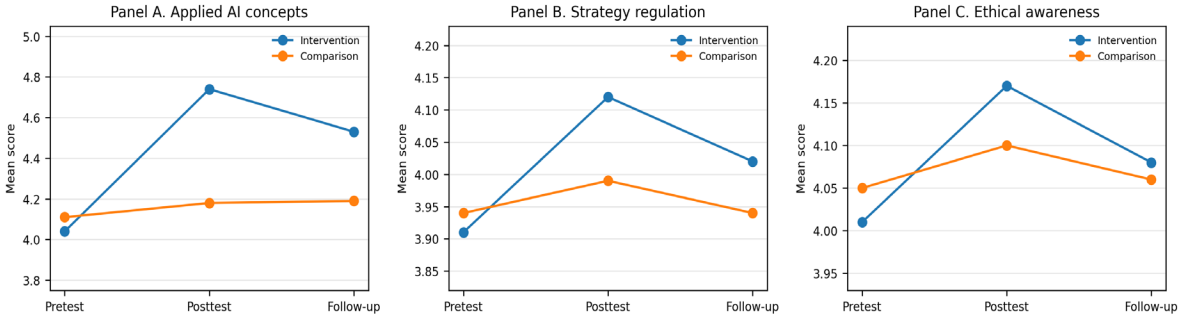

Figure 1. Descriptive mean scores for the three focal outcomes across pretest, posttest, and follow-up by group

Note. Values are descriptive group means at each measurement point. Inferential group-by-time estimates, including confidence intervals and p-values, are reported in Table 4.

Table 3. Descriptive means and standard deviations by group and time point

Outcome	Group	Pretest M (SD)	Posttest M (SD)	Follow-up M (SD)
Applied AI concepts	Intervention	4.04 (1.40)	4.74 (1.50)	4.53 (1.57)
Applied AI concepts	Comparison	4.11 (1.40)	4.18 (1.53)	4.19 (1.54)
Strategy regulation	Intervention	3.91 (0.37)	4.12 (0.38)	4.02 (0.41)
Strategy regulation	Comparison	3.94 (0.37)	3.99 (0.38)	3.94 (0.41)
Ethical awareness	Intervention	4.01 (0.47)	4.17 (0.50)	4.08 (0.57)
Ethical awareness	Comparison	4.05 (0.43)	4.10 (0.44)	4.06 (0.49)

Note. Applied AI concepts scores range from 0 to 7 (number of correct responses). Strategy regulation and ethical awareness scores range from 1 to 5 (mean of Likert-type items). Inferential estimates are reported in Table 4.

Table 4. Planned group-by-time contrasts and standardized differential effects

Outcome	Pre-to-post b [95% CI]	Posttest Hedges' g	p	Pre-to-follow-up b [95% CI]	Follow-up Hedges' g	p
Applied AI concepts	0.54 [0.41, 0.68]; SE = 0.07	0.45	< .001	0.41 [0.28, 0.55]; SE = 0.07	0.29	< .001
Strategy regulation	0.19 [0.15, 0.22]; SE = 0.02	0.41	< .001	0.12 [0.09, 0.16]; SE = 0.02	0.27	< .001
Ethical awareness	0.11 [0.05, 0.16]; SE = 0.03	0.25	< .001	0.05 [-0.01, 0.10]; SE = 0.03	0.15	.090

Note. Coefficients are unstandardized group-by-time interaction estimates from linear mixed-effects models. SE = standard error. Hedges' g values are standardized differential effects calculated from observed change-score contrasts using the pooled baseline standard deviation and a small-sample correction. Positive values indicate larger gains in the intervention group than in the comparison group. The follow-up contrast for ethical awareness (p = .090) did not reach the prespecified significance threshold of $\alpha = .05$.

Changes in Applied AI Concepts

In relation to Research Question 1, the concept test showed the clearest advantage for the intervention. At baseline, the comparison group (M = 4.11, SD = 1.40) and the intervention group (M = 4.04, SD = 1.40) were closely aligned. By posttest, the intervention group mean had risen to 4.74 (SD = 1.50), whereas the comparison group mean was 4.18 (SD = 1.53). At follow-up, the intervention group remained higher (M = 4.53, SD = 1.57) than the comparison group (M = 4.19, SD = 1.54). The mixed model confirmed a larger differential gain for the intervention group from pretest to posttest, b = 0.54, SE = 0.07, 95% CI [0.41, 0.68], p < .001, and from pretest to follow-up, b = 0.41, SE = 0.07, 95% CI [0.28, 0.55], p < .001. These coefficients represent the intervention group's advantage in gain over the comparison group at each time point rather than the intervention group's raw pre-to-post improvement.

This pattern is consistent with the structure of the course. Students were introduced to concepts such as classification, regression, data quality, and overfitting, and they also had to apply those ideas when selecting tools, shaping datasets, and justifying design decisions in their projects. The gain is therefore best interpreted as applied conceptual performance rather than short-term recall.

Development of Strategy Regulation in AI-Supported Problem Solving

In relation to Research Question 2, strategy regulation also favored the intervention group. Baseline means were close, with the comparison group at M = 3.94 (SD = 0.37) and the intervention group at M = 3.91 (SD = 0.37). At posttest, the intervention group reached M = 4.12 (SD = 0.38), compared with M = 3.99 (SD = 0.38) in the comparison group. At follow-up, the intervention group remained above baseline (M = 4.02, SD = 0.41), whereas the comparison group returned nearly to its starting point (M = 3.94, SD = 0.41). The interaction terms supported a larger differential gain for the intervention group at posttest, b = 0.19, SE = 0.02, 95% CI [0.15, 0.22], p < .001, and at follow-up, b = 0.12, SE = 0.02, 95% CI [0.09, 0.16], p < .001.

Students described strategy regulation in operational terms. They reported planning a route before using a tool, checking whether outputs matched the task, comparing alternatives, and revising data or input strategies when early results were weak. This matters because metacognitive language in survey studies can remain abstract. In the qualitative materials, planning, monitoring, checking, and revision were anchored in project decisions rather than treated as detached study skills.

Ethical Awareness in AI Use

In relation to Research Question 3, ethical awareness improved more modestly than the other outcomes. Baseline means were similar, with the comparison group at M = 4.05 (SD = 0.43) and the intervention group at M = 4.01 (SD = 0.47). At posttest, the intervention group rose to M = 4.17 (SD = 0.50), compared with M = 4.10 (SD = 0.44) in the comparison group. At follow-up, the intervention mean was M = 4.08 (SD = 0.57), only slightly above the comparison mean of M = 4.06 (SD = 0.49). The mixed model showed a statistically significant intervention advantage from pretest to posttest (b = 0.11, SE = 0.03, 95% CI [0.05, 0.16], $p < .001$), but the pretest-to-follow-up contrast was smaller. It did not reach the prespecified significance threshold (b = 0.05, SE = 0.03, 95% CI [-0.01, 0.10], p = .090). The confidence interval for the follow-up contrast includes zero, and the result should not be interpreted as evidence of a retained comparative advantage.

The artifact evidence points in the same general direction as the survey data. Many teams were beginning to connect ethical principles to concrete design choices, and the independently rated project reports showed reliable scoring and a moderate association with gains on the ethics survey. The weaker retention pattern remains important: concept application and strategy regulation were repeatedly practiced through explicit task demands, whereas ethical reflection relied more on deliberate instructional cues. The present evidence supports the claim that ethical awareness can be raised in a short, project-based course, but it does not support the claim that the comparative advantage was retained at the four-week follow-up.

Integrated Qualitative Findings

In relation to Research Question 4, four themes were identified in the intervention cohort: conceptual understanding of AI, using AI for problem solving, empowerment to use AI, and ethical considerations in AI use. Across all 288 coded instances from the 155 reflective essays and focus-group interviews with 19 students, ethical considerations accounted for the largest share (92 instances, 31.94%), followed by conceptual understanding (78 instances, 27.08%), using AI for problem solving (72 instances, 25.00%), and empowerment to use AI (46 instances, 15.97%). This frequency distribution reflects thematic salience in students' reflective essays and focus-group accounts; it indicates what students chose to discuss and elaborate on, not the relative magnitude of comparative quantitative gains or their retention at follow-up.

Students most often described learning as movement through an action sequence rather than as the accumulation of isolated facts. They returned to concepts when choosing tools, argued about what data were suitable, checked whether outputs matched the problem, revised their approach when results were weak, and debated what would count as responsible deployment. In that sense, the project served as a setting that brought conceptual, strategic, and ethical considerations into a single workflow.

These accounts help explain the quantitative findings. The strongest and most durable gains appeared in the dimensions exercised most repeatedly through project action: using concepts in context and regulating one's strategy while working with AI tools. Ethical awareness also appeared prominently in students' accounts, but often as a reflective layer added during justification and discussion rather than as a continuously enacted practice. This pattern is consistent with the quantitative results: ethical awareness showed an immediate comparative advantage, but that advantage did not hold as reliably at the four-week follow-up. The qualitative evidence, therefore, supports the view that ethics requires more deliberate reinforcement across later learning activity to produce durable comparative gains.

Table 5 integrates the quantitative findings across the three focal outcomes with the corresponding qualitative themes. Table 6 summarizes the themes and codes identified in the qualitative materials, and Table 7 provides selected translated quotations illustrating those themes. The quotations support the explanatory strand and are not treated as independent comparative outcome data.

Table 5. Joint display of quantitative comparisons and qualitative themes

Strand	Quantitative pattern	Course-linked qualitative theme
Applied AI concepts	Larger differential gains at posttest and follow-up in the intervention group	Repeated use of concepts while making data, tool, and model decisions
Strategy regulation	Larger differential gains at posttest and follow-up in the intervention group	Planning, checking, revision, and peer discussion during project work
Ethical awareness	Significant immediate intervention advantage; follow-up contrast smaller and not statistically reliable at $\alpha = .05$	Ethics prompts supported the identification of risks and safeguards, but the pattern indicates a need for sustained reinforcement.

Table 6. Themes and codes from thematic analysis

Theme	Codes	Total coded instances
Conceptual understanding of AI	AI algorithms (56, 19.44%); applications of AI (22, 7.64%)	78 (27.08%)
Using AI for problem solving	Data selection (21, 7.29%); review of concepts (15, 5.21%); model evaluation (15, 5.21%); algorithm selection (11, 3.82%); model improvement (10, 3.47%)	72 (25.00%)
Empowerment to use AI	AI application project (16, 5.56%); interest in AI (11, 3.82%); peer discussion (8, 2.78%); hands-on application of concepts (6, 2.08%); sense of capability (5, 1.74%)	46 (15.97%)
Ethical considerations in AI use	Responsibility (35, 12.15%); privacy (33, 11.46%); discrimination and bias (16, 5.56%); intellectual property (8, 2.78%)	92 (31.94%)

Note. Percentages are based on all coded instances across focus-group transcripts and self-reflective essays (N = 288 coded instances in total). Theme totals sum to 288. Frequencies reflect thematic salience in student discourse and are not measures of comparative outcome effectiveness. Selected translated quotations are reported in Table 7 to show how the qualitative materials support the themes.

Table 7. Selected translated quotations illustrating qualitative themes

Theme	Illustrative translated quotation
Conceptual understanding of AI	Before this course, I thought AI was just entering a question and getting an answer. I later realized that the quality of the result depends on how data are collected, how the model is selected, and how the trained model is evaluated.
Conceptual understanding of AI	This was the first time I connected the steps of machine learning. It was not only about knowing words such as classification and regression, but about defining the problem, preparing data, and then training and testing the model.
Using AI for problem solving	We did not use the tool immediately at the beginning. We first discussed which AI method was suitable for the task and then divided the work to collect data and test the results.
Using AI for problem solving	The first trained model worked reasonably well on our own samples, but it became inaccurate when we used another set of pictures. We therefore checked the data again and added pictures from different angles and lighting conditions.

Theme	Illustrative translated quotation
Using AI for problem solving	I realized that using AI is not something finished in one attempt. When the result is not good, we need to go back and check whether the data are too limited or whether the samples given to the model are not clear enough.
Ethical considerations in AI use	When collecting pictures or voices, we discussed that we could not casually use other people's personal information, especially face and voice data, without obtaining consent.
Ethical considerations in AI use	If the training data come from only one type of person, the model may not be accurate for others. Therefore, we need to consider whether the data are fair, instead of only looking at overall accuracy.
Empowerment to use AI	I used to think that I could not do an AI project without programming experience. After using low-code tools in this course, I felt that I could also apply AI to learning or professional problems in the future.

Note. Quotations were translated from Chinese into English by the authors. Identifying details were removed to protect participant confidentiality.

Table 8 presents six illustrative student projects from the rated project-report sample, together with their implementation challenges and proposed improvements.

Table 8. Illustrative student projects selected from the rated project-report sample

Project title	Observed challenges	Improvement strategies
Distinguishing coins and banknotes with a no-code classifier	Coins with similar appearances were difficult to distinguish; folded banknotes reduced accuracy.	Collect more varied photos for model training.
Dog breed recognition with a no-code classifier	The target scope included too many dog breeds for the available training data.	Narrow the scope to selected breeds and expand the dataset through mirrored and cropped images.
Age classifier using a no-code classifier	Portrait background and brightness affected age recognition.	Use higher-quality or purpose-made portraits for training and validation.
Squatting posture detection with a no-code classifier	Class imbalance appeared across the three target groups.	Balance training data across groups and consider pose-based or three-dimensional data for the task.
COVID-19 information chatbot with Microsoft Azure QnA Maker	A single source provided insufficient and quickly outdated training data.	Use multiple sources and update the knowledge base regularly.
Sign language recognition with a no-code classifier	Letters J and Z involve motion rather than static hand shapes.	Use an AI-enabled platform more suitable for motion gestures.

Note. The intervention cohort produced 180 project reports. A random sample of 40 reports (22.2%) was rated for evidence of ethical reasoning. The six projects listed in Table 8 were selected from those rated reports to illustrate common project types, implementation challenges, and student-proposed improvements. They are included as illustrative examples rather than as separate outcome cases.

DISCUSSION

This study examined an 18-hour project-based AI literacy course using a quasi-experimental mixed-methods design with a contemporaneous comparison group and a short, delayed follow-up. The findings were not uniform across strands. For Research Question 1, the intervention group showed

larger differential gains in applied AI concepts at both posttest and follow-up. For Research Question 2, the intervention group also showed larger gains in self-reported strategy regulation at both time points. For Research Question 3, ethical awareness showed a statistically significant immediate advantage, but the four-week follow-up contrast was smaller and not statistically reliable. For Research Question 4, qualitative evidence indicated that students linked these changes to repeated project decisions, tool use, data evaluation, peer discussion, and ethics prompts. The findings therefore support a claim about selective short-delay retention, not a blanket claim that every strand of AI literacy improves in the same way or endures equally well after a brief course.

These findings extend prior AI literacy work in three ways. First, they support framework-based arguments that non-specialist learners need applied conceptual understanding, strategic use, and responsible judgment rather than exposure to AI terminology alone (Long & Magerko, 2020; UNESCO, 2021). Second, they add to retention evidence in project-based AI literacy research by showing that strands repeatedly practiced through project-based action were more likely to remain detectable after the course than ethical awareness, which depended more heavily on explicit prompts and subsequent reinforcement. Third, they clarify a boundary in short-course AI literacy provision: a compact intervention can serve as an entry point, but it should not be treated as sufficient for durable, responsible AI reasoning in later learning contexts. This pattern is consistent with the project-based learning literature, which emphasizes authentic inquiry, artifact production, and iterative revision as mechanisms for durable learning (Bell, 2010; Kokotsaki et al., 2016; Krajcik & Shin, 2014).

The magnitude of the observed effects should be interpreted in both practical and statistical terms. In raw-score units, the intervention advantages were modest: the largest posttest contrast was 0.54 points on a 0-7 applied-concepts test, equivalent to approximately 7.7% of the scale range. The strategy-regulation and ethical-awareness contrasts were smaller on 1-5 scales, and the ethical-awareness advantage was not statistically reliable at follow-up. Hedges' g values calculated from the observed change-score contrasts were 0.45 for applied AI concepts, 0.41 for strategy regulation, and 0.25 for ethical awareness at posttest; the corresponding follow-up estimates were 0.29, 0.27, and 0.15. These values place the observed advantages in the small-to-moderate range.

The main contribution is not a claim of large effects from a brief course. Rather, the study characterizes a selective retention profile: applied concept use and self-reported strategy regulation showed detectable comparative advantages beyond course completion, whereas ethical awareness showed a more fragile pattern. The posttest-to-follow-up decline also indicates that immediate post-course gains should not be treated as equivalent to retained learning. This distinction matters for course design because it indicates which strands of AI literacy are more likely to be sustained by repeated project action and which require deliberate later reinforcement; this is a practical implication that a single-group or immediate-posttest design could not have produced.

Because the design was quasi-experimental and allocation was not randomized, the observed group differences should be interpreted as associations between course participation and outcome trajectories rather than as evidence of isolated causal effects. The pattern is consistent with a course-related explanation, and the baseline equivalence check, sensitivity analyses, and convergent qualitative evidence collectively strengthen that interpretation. Nevertheless, the possibility that unmeasured pre-existing differences - in motivation, prior AI engagement, or concurrent learning activity - contributed to the observed pattern cannot be fully excluded.

Within this interpretive boundary, the pattern identifies where a short-form, project-based AI literacy course appears most promising for instruction. The intervention did not merely expose students to AI terminology. It repeatedly required them to use concepts from ML and deep learning inside concrete project decisions, and to plan, check, and revise their work as tools, data, and outputs changed. Those repeated task demands are consistent with the stronger, more durable differential advantages observed in applied concepts and strategy regulation. The qualitative record supports that

interpretation: students described planning a route before using tools, monitoring intermediate outputs, and revising weak or implausible results while the work was underway.

The ethical awareness finding should be interpreted more cautiously. The course produced a modest but statistically significant immediate comparative advantage, and both reflections and project artifacts indicate that students were beginning to articulate fairness, privacy, autonomy, bias, and safeguards in relation to concrete design choices. What did not hold as reliably over the short delay was the comparative advantage itself. The confidence interval for the follow-up contrast includes zero, and the result should not be read as evidence of retained comparative effectiveness. This does not mean that ethics cannot be addressed in a brief intervention; rather, ethical awareness appears less likely to remain comparatively stable when it is introduced early and only lightly reinforced. Relative to concepts and strategy regulation, ethical reasoning appears to require more deliberate return, reuse, debate, case analysis, and critique across later learning activities.

Implications for Short-Form AI Literacy Course Design and Evaluation

For course design, the retention profile suggests that short AI literacy interventions show the most promise when conceptual content and project decision making are tightly interwoven. Students showed stronger, more durable differential gains when they had to apply concepts to make repeated decisions about problems, data, tools, and outputs. Ethical content calls for a different design response. Brief exposure can raise awareness, but sustained comparative advantage is more likely to require structured reinforcement through new cases, critique tasks, discipline-specific reuse, and assessment prompts that require students to justify responsible AI decisions beyond the initial course. In curriculum terms, an introductory AI literacy course should be followed by discipline-specific assignments in which students must revisit tool choice, data quality, bias, privacy, human oversight, and accountability in later coursework.

For evaluation, the combination of a comparison group and a delayed follow-up strengthens the evidential value of the study relative to single-group pre-post designs. It allows the argument to move from simple course-period improvement to comparative differential gain and short-delay retention. That level of evidence is more useful for higher education curriculum decisions than an immediate post-course survey alone, especially when institutions are considering whether compact AI literacy courses can serve large non-specialist cohorts. At the same time, the quasi-experimental design means that these findings are best treated as informative rather than definitive, and as a basis for further controlled inquiry rather than as grounds for strong causal conclusions. Future evaluations should compare different reinforcement models, include behavioral or artifact-based measures of strategy regulation, and examine whether AI literacy transfers into later coursework, internships, or discipline-specific tasks.

Limitations and Boundary Conditions

Several limits define the boundary of the claims made in this article. The design was quasi-experimental rather than randomized. Students self-selected into the intervention through voluntary enrollment, and the comparison group comprised first-year undergraduates who did not enroll during the study period. This means that students who chose to participate in the project-based course may have differed systematically from those who did not - in motivation to engage with AI, prior enthusiasm for technology, or readiness to invest discretionary time. The observed baseline standardized mean differences for the three focal outcomes ranged from -0.05 to -0.09, indicating that the groups were closely aligned on measured outcome levels before the intervention. However, baseline equivalence checks address only observed outcome levels; they do not rule out unmeasured differences in motivation, prior AI interest, or concurrent learning experiences that could have contributed to the differential trajectories. The sensitivity analyses and convergent qualitative evidence reduce the

plausibility of a purely artifactual result, but residual selection bias remains a boundary condition for interpretation.

A related consideration is the high rate of prior AI tool use in the sample. Across both groups, 86.7% of participants reported having used AI tools before the study. The generalizability of the findings to cohorts with substantially lower prior AI exposure is uncertain, and future studies should examine whether the observed pattern holds in samples with more varied starting points.

The follow-up was short, centered around four weeks, which means the findings concern early retention rather than long-term transfer or durable consolidation. Whether the observed differential advantages persist beyond this window, diminish further, or are maintained with additional reinforcement is an open question that the present design cannot address.

The qualitative strand came only from the intervention cohort, making it suitable for explaining the processes associated with outcome change within that group but not for direct cross-group comparison of experiences. The qualitative evidence, therefore, informs interpretation of the quantitative patterns rather than providing independent evidence of comparative effectiveness.

Strategy regulation was measured with a brief 4-item self-report Likert-type scale. Scores on this instrument reflect students' perceived use of planning, monitoring, and revision strategies rather than direct behavioral observation of regulatory activity during task performance. The observed differential gains on this scale should be interpreted as evidence of greater self-reported strategic awareness in the intervention group, not as a direct measure of strategic behavior. Behavioral or artifact-based measures of regulation would be needed to strengthen that inference.

All three instruments were intentionally short to support repeated administration to large first-year cohorts. Reliability evidence from the primary analytic sample was mixed. The 7-item concepts test showed limited internal consistency, especially at pretest, consistent with its role as a brief, scenario-based applied-performance indicator rather than a broad, standardized scale. The two Likert-type scales showed more coherent internal structure, particularly when omega estimates were considered alongside alpha. The outcome measures should therefore be interpreted as focused indicators of applied concept use, perceived strategy regulation, and ethical awareness rather than as exhaustive latent-trait measures.

The study was conducted at a single university among first-year undergraduates who followed a single course design. The claims should remain within the domain of short-form higher-education AI literacy interventions until replicated in other institutional, disciplinary, and cultural settings.

CONCLUSION

This study contributes comparative evidence on where a compact project-based AI literacy intervention is most and least likely to produce short-delay retention effects. In response to the first research question, the intervention group showed greater differential gains in applied AI concepts at posttest and at the four-week follow-up. In response to the second research question, the intervention group also showed larger gains in self-reported strategy regulation at both time points. In response to the third research question, ethical awareness showed a modest but statistically significant immediate comparative advantage; however, the four-week follow-up contrast was smaller and was not statistically reliable, indicating that the comparative advantage did not hold over the short delay. In response to the fourth research question, students associated learning with repeated project decisions, tool use, data evaluation, revision, peer discussion, and ethics prompts. Because the design was quasi-experimental, these patterns should be interpreted as consistent with a course-related explanation rather than as evidence of isolated causal effects.

For higher education, the findings suggest that short, project-based AI literacy courses are associated with applied conceptual gains and greater self-reported strategic awareness that persist beyond the

instructional period. The study extends AI literacy intervention research by showing that different literacy strands exhibited varying levels of persistence following the same brief intervention, which is the article's main original contribution. Ethical reasoning showed a more fragile pattern, consistent with the interpretation that responsible AI use requires continued reinforcement rather than a single brief treatment. Institutions considering compact AI literacy provision for non-specialist undergraduates should account for this asymmetry in how different strands of AI literacy respond to short-form instruction and should build in structured opportunities to revisit ethical reasoning in subsequent coursework or disciplinary contexts. The main boundaries of this conclusion are the quasi-experimental design, the single-university first-year sample, the brief self-report strategy scale, the intervention-only qualitative strand, and the four-week follow-up window. Future research should examine longer-term retention, transfer into authentic coursework, behavioral evidence of regulation, and reinforcement models across disciplines and institutions.

Data Availability

De-identified individual-level quantitative data, the codebook or data dictionary, anonymized survey instruments, and statistical scripts will be made available as supplementary materials or through a recognized repository in accordance with the journal's data availability policy. Because the reflective essays, focus-group transcripts, and project artifacts contain contextual details that could compromise participant confidentiality, the raw qualitative materials are not publicly shared. Selected translated and de-identified quotations are reported in the manuscript for illustrative purposes. Additional de-identified qualitative excerpts, the interview guide, the coding framework, and rubric scoring documentation can be provided where this is consistent with institutional ethics approval and participant consent.

Ethical Approval and Consent

This study received prior ethics approval from the relevant institutional review board before data collection. The classroom-based evaluation involved adults only. Participation was voluntary and did not affect course grades or standing. Written or electronic informed consent was obtained before data collection, and students could withdraw without penalty. Student-level records were de-identified before analysis and stored securely. The comparison group completed the same measurement schedule but did not lose access to any required curriculum, because the intervention was a voluntary non-credit course rather than a compulsory program requirement. Consent to use de-identified quotations and artifacts was obtained.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The authors declare no competing interests.

References

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, *33*(2), 3-30. https://doi.org/10.1257/jep.33.2.3

Almatrafi, O., Johri, A., & Lee, H. (2024). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019-2023). *Computers & Education Open*, *6*, Article 100173. https://doi.org/10.1016/j.caeo.2024.100173

Bandura, A. (1997). *Self-efficacy: The exercise of control.* W. H. Freeman.

Bell, S. (2010). Project-based learning for the 21st century: Skills for the future. *The Clearing House*, *83*(2), 39-43. https://doi.org/10.1080/00098650903505415

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

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

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data: Evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63-71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

European Commission. (2019). *Ethics guidelines for trustworthy AI.* High-Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People: An ethical framework for a good AI society. *Minds and Machines*, *28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*, 99-120. https://doi.org/10.1007/s11023-020-09517-8

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational Technology Research and Development*, *48*(3), 23-48. https://doi.org/10.1007/BF02319856

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kokotsaki, D., Menzies, V., & Wiggins, A. (2016). Project-based learning: A review of the literature. *Improving Schools*, *19*(3), 267-277. https://doi.org/10.1177/1365480216659733

Kong, S.-C., Cheung, M.-Y. W., & Tsang, O. (2024). Developing an artificial intelligence literacy framework: Evaluation of a literacy course for senior secondary students using a project-based learning approach. *Computers and Education: Artificial Intelligence*, *6*, Article 100214. https://doi.org/10.1016/j.caeai.2024.100214

Krajcik, J., & Shin, N. (2014). Project-based learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 275-297). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.018

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, *3*, Article 100101. https://doi.org/10.1016/j.caeai.2022.100101

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

OECD. (2022). *OECD framework for the classification of AI systems.* OECD Publishing. https://doi.org/10.1787/cb6d9eca-en

OSTP. (2022). *Blueprint for an AI Bill of Rights: Making automated systems work for the American people.* The White House Office of Science and Technology Policy. https://data.aclum.org/storage/2025/01/OSTP_www_whitehouse_gov_ostp_ai-bill-of-rights.pdf

Prince, M. J., & Felder, R. M. (2006). Inductive teaching and learning methods: Definitions, comparisons, and research bases. *Journal of Engineering Education*, *95*(2), 123-138. https://doi.org/10.1002/j.2168-9830.2006.tb00884.x

R Core Team. (2023). *R: A language and environment for statistical computing (Version 4.3.2).* R Foundation for Statistical Computing. https://www.R-project.org/

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68-78. https://doi.org/10.1037/0003-066X.55.1.68

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, *36*(1-2), 111-139. https://doi.org/10.1007/s11165-005-3917-8

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Spreitzer, G. M. (1995). Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal*, *38*(5), 1442-1465. https://doi.org/10.2307/256865

Thomas, J. W. (2000). *A review of research on project-based learning.* Autodesk Foundation.

Thomas, K. W., & Velthouse, B. A. (1990). Cognitive elements of empowerment: An interpretive model of intrinsic task motivation. *Academy of Management Review*, *15*(4), 666-681. https://doi.org/10.2307/258687

Tierney, P., & Farmer, S. M. (2002). Creative self-efficacy: Its potential antecedents and relationship to creative performance. *Academy of Management Journal*, *45*(6), 1137-1148. https://doi.org/10.5465/3069429

Touretzky, D. S., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K-12: What should every child know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 9795-9799. https://doi.org/10.1609/aaai.v33i01.33019795

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence.* https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

Winne, P. H., & Azevedo, R. (2014). Metacognition. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 63-87). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.006

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education*, *16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0

APPENDIX A. EXCERPT FROM THE AI CONCEPTS FOR PROBLEM SOLVING TEST

1. A university learning analytics team has click-stream logs from 8,000 first-year students but no labels about study styles. They want to group students by similar behavior patterns to tailor support. Which approach is most appropriate? A. Clustering algorithm; B. Classification algorithm; C. Regression algorithm; D. Recommender system based on user ratings.
2. The campus language lab is building a tool to detect Mandarin tone mispronunciations on a fixed word list. Which audio data would be most suitable to collect for training? A. Only correct pronunciations of each word; B. Each word pronounced with all four tones, whether correct or incorrect; C. Long random paragraphs read in Mandarin; D. Pairs of words from different languages mixed in one recording.
3. A biology instructor wants an AI that auto-generates captions for photos of campus flora and fauna to streamline teaching preparation. Which training data are most appropriate? A. Photos of target-domain plants or animals with concise species names as captions; B. Photos of any plants or animals with long descriptive paragraphs as captions; C. Photos of local plants or animals without captions; D. Photos of household objects with their names as captions.
4. An NBA fan wants to predict the number of games a team will win this season from past match statistics. Which type of algorithm should be used? A. Classification; B. Regression; C. Clustering; D. Object detection.

5. The IT helpdesk triages student tickets into urgent and non-urgent categories. Historical tickets are labeled with the final triage outcome. Which approach is most appropriate? A. Unsupervised clustering on ticket texts; B. Rule-based keyword matching only; C. Supervised text classification using labeled tickets; D. Image segmentation.
6. A student team trains an image model that gets 98% training accuracy but only 71% validation accuracy on the same task. What is the most appropriate next step? A. Train for more epochs to push training accuracy to 100%; B. Increase data diversity and/or regularization and tune using a proper validation split; C. Remove the validation set; D. Report training accuracy only.
7. A media literacy project aims to flag likely deepfakes by analyzing video and associated transcripts. Which is the best overall strategy for a first practical solution in a short course? A. Train a new multimodal network from scratch on a tiny class dataset; B. Use pretrained models and fine-tune on a curated labeled subset; C. Use k-means clustering on raw pixels and raw words; D. Rely on a single hand-crafted rule.

APPENDIX B. EXCERPT FROM THE STRATEGY REGULATION AND BROADER EMPOWERMENT SURVEY

Table B1 presents the strategy regulation items and the broader empowerment-context items.

Table B1. Strategy regulation and broader empowerment survey items

Component	Item ID	Item statement
Strategy regulation	SR1	Before using an AI tool, I define the problem and decide what evidence or data is needed.
Strategy regulation	SR2	I monitor whether AI outputs match the task goals, constraints, and available evidence.
Strategy regulation	SR3	When AI outputs are weak, biased, or unclear, I revise prompts, data, methods, or tool choices.
Strategy regulation	SR4	I compare alternative AI approaches or outputs before deciding which result to use.
Impact	I1	I want to use AI to address problems that matter in my community or field.
Impact	I2	Applying AI in my coursework or future work can create positive outcomes for others.
Creative self-efficacy	C1	When my first idea does not work, I can generate alternative AI-based solutions.
Creative self-efficacy	C2	I can combine features from different AI tools to create new approaches.
AI self-efficacy	S1	I can select and apply an appropriate AI method to a task I have not seen before.
AI self-efficacy	S2	I can troubleshoot and resolve issues when an AI approach does not work as expected.
Meaningfulness	M1	Learning to solve problems with AI is personally valuable to me.
Meaningfulness	M2	Becoming proficient in AI problem-solving will help me reach my academic and career goals.

Note. The four SR items formed the strategy regulation scale used as a primary quantitative outcome. The remaining items were retained as broader empowerment-context items and were not treated as primary comparative outcomes in the present article.

APPENDIX C. EXCERPT FROM THE AI ETHICS SURVEY

Table C1 presents the AI ethics survey items.

Please rate each statement from 1 (strongly disagree) to 5 (strongly agree).

Table C1. AI ethics survey items

Principle	Item ID	Item statement
Beneficence	B1	When applying AI to solve a problem, benefits should clearly outweigh potential risks to individuals and society.
Beneficence	B2	I consider possible harms and plan safeguards before deploying an AI-supported solution.
Fairness	F1	Access to AI systems, data, and outcomes should be distributed fairly across different user groups.
Fairness	F2	I check for and address bias in datasets, labels, and model performance across subgroups before accepting results.
Human autonomy	H1	Humans should retain final decision authority for consequential outcomes produced by AI systems.
Human autonomy	H2	AI tools should be designed to support human choice, including providing explanations and the ability to opt out.

APPENDIX D. EXCERPT FROM THE SEMI-STRUCTURED FOCUS GROUP INTERVIEW GUIDE

1. *Using AI to solve problems:* What did you learn about applying AI to real problems in this course? Probe: Give an example where an AI tool changed how you approached the task.
2. *PBL process:* How did the project help you plan your approach, monitor progress, and evaluate results? Probe: Describe a decision point where you compared alternatives and chose one.
3. *Data work:* How did you decide which data to collect and how to process them? Probe: What steps did you take to handle noisy or biased data?
4. *Conceptual understanding:* Which AI concepts or workflows do you feel you can now apply, and in what contexts? Probe: What concept was hardest to grasp, and how did you overcome it?
5. *Empowerment:* Did the project change your confidence in using AI or your sense that it is meaningful to you? Probe: Describe a moment when you had to be creative after a solution failed.
6. *Ethics and responsible use:* How did you consider human autonomy, beneficence, and fairness in your project? Probe: Would you deploy your solution in a real setting? What safeguards would you add?
7. *Transfer and future use:* Where do you expect to use AI next, in coursework, internships, or everyday tasks? Probe: What additional skills or supports would help you apply AI responsibly?
8. *Course design feedback:* What aspects of the course or project structure helped most? What should be changed and why?

APPENDIX E. RELIABILITY AND ITEM DIAGNOSTICS FOR THE PRIMARY ANALYTIC SAMPLE (N = 360)

Internal consistency and item diagnostics were examined using the de-identified individual-level quantitative dataset from the primary analytic sample. The objective AI Concepts for Problem Solving test was evaluated using KR-20, item difficulty, and corrected item-total correlations (Table E1). The strategy regulation scale and AI ethics survey were evaluated using Cronbach's alpha, McDonald's omega, and mean inter-item correlations at each wave, as reported in Tables E2 and E3, respectively.

Because the concepts test contains only seven scenario-based items that deliberately sample different AI task families, its reliability estimates should be interpreted as evidence for a brief applied performance indicator rather than for a broad standardized achievement scale. The low KR-20 on the pretest reflects the limited homogeneity expected from a short cross-task test before instruction; it does not, by itself, indicate that the items were unsuitable for sampling applied concept use. Posttest and follow-up estimates were higher but still modest. The self-report scales showed more coherent internal structure, particularly when omega was considered alongside alpha.

Table E1. AI Concepts for Problem Solving test reliability and item diagnostics

Wave	KR-20	Item difficulty range (p)	Corrected item-total r range
Pretest	.171	0.511-0.700	-0.016-0.142
Posttest	.381	0.528-0.714	0.027-0.262
Follow-up	.397	0.528-0.717	0.107-0.213

Note. Item difficulty p-values reflect the proportion of correct responses, with higher values indicating easier items. Corrected item-total correlations are based on the full sample at each wave.

Table E2. Strategy regulation scale reliability

Wave	Cronbach's alpha	McDonald's omega	Mean inter-item r
Pretest	.659	.797	.327
Posttest	.710	.821	.379
Follow-up	.677	.806	.345

Note. McDonald's omega is reported alongside alpha because four-item scales can show alpha values that are sensitive to the number of items and the item covariance structure. The four items correspond to problem definition, monitoring, revision, and comparison of alternative AI approaches.

Table E3. AI ethics survey reliability

Wave	Cronbach's alpha	McDonald's omega	Mean inter-item r
Pretest	.645	.772	.232
Posttest	.738	.821	.320
Follow-up	.742	.823	.324

Note. The AI Ethics Survey includes items addressing beneficence, fairness, and human autonomy. Reliability estimates are reported for the full six-item scale at each wave.

APPENDIX F. RUBRIC FOR EVALUATING ETHICAL REASONING IN STUDENT PROJECT REPORTS

Table F1 presents the rubric used to evaluate ethical reasoning in student project reports.

Table F1. Rubric for evaluating ethical reasoning in student project reports

Dimension	Definition	0	1	2	3
Autonomy	Respect for human agency in AI use and decision making	Not mentioned	Mentioned superficially	Linked to user choice or transparency	Integrated into system design or justification
Beneficence	Promotion of well-being and harm reduction through AI design	Absent	Stated intent only	Concrete steps described	Demonstrated mitigation strategies in decision flow
Fairness	Attention to bias, equity, and representativeness	Absent	General mention of fairness	Identified specific bias risks	Proposed or implemented corrective measures

APPENDIX G. ADDITIONAL REPORTING DETAILS FOR STATISTICAL AND QUALITATIVE ANALYSIS

The primary quantitative models used fixed effects for group, time, and group-by-time interaction, with a participant-level random intercept. The planned contrasts were pretest-to-posttest and pretest-to-follow-up differential gains. Missing later-wave outcome observations were retained within the mixed-model framework after baseline inclusion, rather than removed via listwise deletion. Covariate-adjusted and complete-case sensitivity checks were used to examine whether the direction of the three main findings changed under alternative specifications.

For qualitative analysis, the coding process combined deductive categories derived from the research questions with inductive refinement informed by the essays and interviews. One researcher completed the initial coding, and a second researcher reviewed the codebook, theme boundaries, and a sample of coded materials. The final themes were used to explain the quantitative patterns rather than to claim independent comparative effectiveness.

AUTHORS

Senhao Xiong, Khon Kaen University, Khon Kaen, Thailand, Email: senhao.x@kkumail.com

Senhao Xiong is a doctoral student in Innovation, Technology, and Learning Sciences at Khon Kaen University, Thailand. His research interests include artificial intelligence literacy, educational technology, project-based learning, learning sciences, and higher education curriculum design. His current work examines how emerging technologies support problem-solving, strategic learning, and responsible AI use in higher education.

Yushuang Wu (Corresponding author), Khon Kaen University, Khon Kaen, Thailand, Email: yushuang.w@kkumail.com

Yushuang Wu is a doctoral student in Innovation, Technology, and Learning Sciences at Khon Kaen University, Thailand. Her research interests include artificial intelligence in education, educational technology, learning sciences, project-based learning, and cultural and creative education. She is especially interested in the use of digital technologies to support student engagement, creative practice, and interdisciplinary learning.

HaoHong Nie, Khon Kaen University, Khon Kaen, Thailand, Email: haohong.n@kkumail.com

HaoHong Nie is a doctoral student in Educational Administration at Khon Kaen University, Thailand. He also serves as Secretary of the Youth League General Branch of the School of Electronics and Internet of Things at Chongqing Vocational University of Electronic Science and Technology, China. His research interests include practical education, vocational education, educational administration, student development, and educational technology in applied talent cultivation.

KuiHua Li, Chongqing Three Gorges Medical College, Chongqing, China, Email: kuihua.l@kkumail.com

KuiHua Li is a doctoral student in Educational Administration at Khon Kaen University, Thailand. He is also the Director of the Cooperation and Exchange Center at Chongqing Three Gorges Medical College in China. His research interests include educational administration, higher vocational education, institutional cooperation, international exchange, and the management of applied education programs. His work focuses on strengthening educational cooperation and improving the quality of vocational and professional education.