Enhancing Accounting Information System Students' AI Literacy and Internal Control Evaluation Skills Through a Generative AI Assisted Digital Portfolio

Zaifudin Zainol

Accounting & Finance Department, Universiti Tenaga Nasional, Malaysia

DOI: https://doi.org/10.47772/IJRISS.2026.101000389

Received: 03 September 2026; Accepted: 24 October 2026; Published: 03 November 2026

ABSTRACT

This mixed-method, one-group pre-test/post-test study examines changes in accounting information systems (AIS) performance, artificial intelligence (AI) literacy and learning engagement during a six-week generative AI (GenAI)-assisted digital-portfolio intervention involving 40 Bachelor of Accounting students. Students analysed an e-commerce revenue cycle, assessed internal-control risks, critiqued AI-generated recommendations, revised a risk–control matrix and documented ethical considerations. Reported paired-sample results indicate significant increases in AIS performance (mean change = 12.05 points, $t(39) = 10.09$, $p < .001$), AI literacy (0.80 points, $t(39) = 9.28$, $p < .001$) and engagement (0.48 points, $t(39) = 6.23$, $p < .001$). Portfolio performance was positively associated with AIS gains ($r = .450$, $p = .004$) and AI-literacy gains ($r = .314$, $p = .049$). The findings are consistent with the value of assessment designs that require students to interrogate AI output and justify their professional judgement.

Keywords: accounting information systems; generative artificial intelligence; AI literacy; internal control

INTRODUCTION

GenAI has quickly transformed higher education because it can do a variety of things like explain concepts, give examples, and summarize information. It can also respond to student questions. Because of these capabilities, schools are worried that students will use GenAI for coursework, answer exam questions, and even plagiarize course material. While this technology has many uses in education, students and professors are concerned with accuracy, bias, whether it promotes independent thinking, and the ethics of authorship. UNESCO (Miao & Holmes, 2023) recommends using GenAI in a human-centered approach. Here students learn about GenAI's capabilities and helps students learn to verify and validate the outputs. AACSB International (2024) states that students should learn about the opportunities and threats that artificial intelligence (AI) poses to society and the business world in general. In addition, students need to have the knowledge to shape the future of AI ethically and responsibly.

Due to the rapid change in technologies, methodologies, practices and procedures in the accounting information system (AIS), education on this area is becoming increasingly difficult. This is because AIS integrates several areas including accounting, transactions, data, and processes. AIS education helps students develop the skills necessary to identify vulnerabilities in business processes and risks that affect the quality of information. These students should be able to recommend controls that address vulnerabilities in business process reliability. Accounting education research has identified the importance of combining accounting and information systems education and preparing students to identify vulnerabilities, risks and controls in a firm (Al-Htaybat et al., 2018; Apostolou et al. 2014; Sledgianowski et al., 2017). With GenAI, it is even more important to educate students to identify the possible vulnerabilities in the systems and recommend solutions.

The main issue isn't if students use AI. The main issue is whether an assessment design can make AI use visible, accountable and intellectually demanding. If students don't have to retain the prompt, or otherwise assess and resolve the weaknesses of the AI response (e.g. missing risks, threats, or opportunities), they may hide their weak understanding by using AI to generate analyses and reports. On the other hand, if students need to retain the

prompts, audit the AI response for failings, identify and analyze the missing risks, threats and opportunities, access and integrate reliable evidence, edit and adjust the AI response to better align with professional judgement and standards, and explain their assessment and adjustment of the AI response, they can engage critically with the AI response and the underlying AI tool. This kind of integration is consistent with the recommendations from a recent accounting education scholarship (Ballantine et al., 2024; Dong et al., 2024). Further, empirical studies indicate that students recognize the value of AI tools (e.g., ChatGPT) for simplifying and improving the learning process in accounting, but there is variability in students' perceptions and outcomes with the tool (Blondeel et al., 2025; Sundkvist & Kulset, 2024; Zhou & Luo, 2025).

The research examines the integration of a GenAI-assisted AIS internal-control portfolio with pre-test and post-test measures, a criterion-referenced portfolio rubric, reflective responses and classroom-observation evidence. Attention is given to whether students verify AI outputs against AIS evidence rather than treating machine-generated explanations as authoritative.

1.1 Problem statement

The educational difficulty is therefore both disciplinary and methodological. In AIS, students must move beyond memorising definitions of controls to explaining how a particular weakness compromises an information-processing objective and how an appropriately positioned control would reduce the exposure. GenAI can provide a readily available starting explanation, but its responses may be generic, omit important elements of the transaction cycle or recommend measures that do not match the facts of a case. The central instructional problem is to design activities that develop students' independent ability to detect and correct those weaknesses rather than merely improve the presentation of AI-generated text.

This study addresses that problem through an assessed digital portfolio requiring a documented sequence of initial AI assistance, critical interrogation, revision and professional justification. The resulting empirical question concerns changes in student outcomes during a structured six-week AIS learning experience, together with the relationship between the quality of students' portfolios and their measured learning gains. The study does not seek to isolate the independent contribution of GenAI from the remaining instructional elements.

Access to GenAI does not necessarily mean accounting students will improve their accounting judgement and automate processes. Students may accept GenAI responses that they deem "fluent" as being "correct" and may miss unsupported conclusions and unverified assumptions. Moreover, students may believe the system recommendations are suitable for their organisation and may fail to exercise judgement and consider the appropriateness of the system recommendations in the organisational context. Current assessments may evaluate the student's knowledge of AIS, but there are very few ways to evaluate how students integrate GenAI systems and assess judgement in context to Automated System Recommendations. There is, therefore, a need for an educational intervention to guide students on how to automate AIS-based tasks, while retaining responsibility for human judgement and evidential reasoning, and academic integrity.

1.2 Research questions, objectives and hypotheses

Table 1. Alignment of the research questions, objectives and hypotheses

Reference	Research question and objective	Hypothesis
RQ1 / Objective 1	Does the AI-assisted AIS case portfolio improve students' internal-control evaluation performance?	H1: Post-intervention AIS performance is significantly higher than pre-intervention performance.
RQ2 / Objective 2	Does the intervention improve students' AI literacy?	H2: Post-intervention AI-literacy scores are significantly higher than pre-intervention scores.
RQ3 / Objective 3	Does the intervention improve students' learning engagement?	H3: Post-intervention engagement scores are significantly higher than pre-intervention scores.

RQ4 / Objective 4	How is portfolio quality associated with learning improvement?	H4: Portfolio performance is positively associated with AIS and AI-literacy gains.
RQ5 / Objective 5	How do students describe the usefulness, limitations and ethical implications of GenAI in AIS learning?	Exploratory qualitative question; no directional hypothesis.

Table 1 establishes the alignment between the five research questions, their corresponding objectives and the hypotheses or analytical approaches applied in this study. The first three questions focus on within-student changes between the initial and final assessments. H1 addresses the AIS internal-control test, H2 addresses AI literacy and H3 addresses learning engagement. These hypotheses can be examined with the paired observations obtained from the same 40 students, provided that pre-test and post-test records are correctly matched. The wording 'improve' in the research questions describes the intended educational outcome; the statistical comparisons themselves test change over time rather than isolating the independent effect of GenAI from other instructional activities.

The fourth question adopts an associational perspective by examining whether students who submitted higher-quality digital portfolios also achieved larger gains in AIS performance and AI literacy. H4 therefore concerns positive correlations, not a causal pathway from portfolio marks to learning. The fifth question is qualitative and exploratory, addressing student perceptions of usefulness, limitations and ethical responsibilities without imposing a directional hypothesis. The combination provides a coherent mixed-method framework: quantitative comparisons establish the pattern and magnitude of measured changes, whereas the qualitative component explores how students experienced the structured use of GenAI.

1.3 Contribution of the study

The empirical contribution is the joint reporting of three pre/post learning outcomes, portfolio-based disciplinary performance and complementary student-reflection and classroom-observation evidence within the same AIS intervention. This combination permits comparison between what students reported about using GenAI, what was observed during learning activities and what they demonstrated on structured assessments. As these data originate from a single cohort without a comparison group, the contribution is best characterised as evidence of change and association in an authentic instructional setting, rather than definitive evidence that GenAI itself produced the change.

This paper contributes to the scholarship of teaching and learning (SoTL) in three respects. First, it operationalises GenAI as a fallible draft that students must interrogate and correct, shifting assessment from fluency to verified professional judgement. Second, unlike approaches that teach AI literacy as a separate digital competence, it embeds prompt construction, evidence verification and ethical disclosure directly into the disciplinary AIS tasks of process mapping, risk recognition and control design. Third, it offers a six-week portfolio, five-criterion analytic rubric and triangulated evaluation protocol that lecturers can adopt and test empirically. These are contributions of instructional design, not yet demonstrated causal effects.

LITERATURE REVIEW AND CONCEPTUAL FOUNDATION

2.1 Technology and the changing AIS curriculum

In accounting education, attention has generally been paid to the integration of accounting processes and information technology. Common elements include transactions and documentation; various database types and concepts; and internal controls and systems. Apostolou et al. (2014) noted that AIS education research has addressed the integration of technologies and systems, curricular design, case studies and student learning. Thus, AIS is suitable for class-based innovations. Sledgianowski et al. (2017) suggested integrating technology and information systems competencies into accounting education, with a view to preparing accounting graduates to work in data-intensive environments. Al-Htaybat et al. (2018) noted that accounting educators have recognised the requirement to prepare accounting students for a digitally-oriented profession.

In AIS, internal control analysis is particularly relevant to authentic assessment. A control is not correct simply because it sounds reasonable. Students have to identify the objective being safeguarded; understand the process and the risk; consider the control's feasibility; and understand the interaction of preventive, detective and corrective controls. COSO (2013) states that internal control is a process that is directed toward achieving the objectives of the organization. Therefore, a strong student response will define the risk in question, and state how the control being recommended mitigates that risk.

2.2 Generative AI in accounting education

The accounting education literature has pivoted quickly from speculating about GenAI to publishing empirical studies and exemplifying pedagogical applications. Ballantine et al. (2024) consider the implications of AI across the accounting education continuum and identify potential benefits, threats and challenges. Dong et al. (2024) examine the growing incorporation of LLMs in accounting and finance research and practice, and identify a number of open questions, including dependability and integrity.

There is considerable potential for GenAI to improve account education, including teaching. The extent to which it can be used to facilitate personalised instruction at scale remains to be seen. Alkan (2025) considers the education challenges GenAI may help address, including irregular access to high quality teaching, lack of instruction at the right level, excessive conditionality on learner readiness, and market failures. Student-centered evidence suggests that perceived usefulness is high, but that in itself does not indicate it is used for learning. Sundkvist and Kulset (2024) indicate that most students anticipate a positive and responsible integration of GenAI into their learning. Zhou and Luo (2025) indicate students generally anticipate GenAI will improve their learning experience.

Blondeel et al. (2025) explore how students use ChatGPT to assist them with accounting data analytics. They show that there is a more complex relationship between a student's knowledge and experience with AI and their performance in the course than simply using ChatGPT as a reference tool. They recommend assignments where students integrate the use of AI to assist them, RNWE, and identify critical errors. This is the same logic that is employed in the AIS Portfolio.

The studies support a more guided integration than outright banning or completely unbounded integration. The AI Assessment Scale shows how the role of AI within an assessment can be communicated to students. The role of AI within an assessment can be designed with the learning objectives in mind. Rewriting tasks and aligning the work of students with learning objectives is appropriate (Furze et al., 2024). In the current study, students can use GenAI to produce an initial analysis, or suggest prompts to improve analysis, but must retain an audit trail, explain and substantiate their answer and be accountable for the final version. This supports the integrated, high-transparency, human-in-the-loop approach.

2.3 AI literacy, critical evaluation and ethical use

The conceptual rationale for the intervention is that AI literacy and AIS judgement can reinforce one another when students are required to justify the use of technology within a domain-specific task. Prompt formulation requires a sufficiently accurate representation of the transaction process; evaluation requires the student to compare the generated explanation with established accounting and internal-control concepts; and revision requires articulation of a defensible control response. This sequence provides observable evidence of evaluation rather than treating frequency of AI use as a proxy for competence (Long & Magerko, 2020; Ng et al., 2021).

Constructive alignment further suggests that the learning activity, assessment rubric and intended outcomes should reward the same forms of reasoning (Biggs, 1996). Accordingly, the intervention assesses the accuracy of process analysis, the fit between risks and controls, the identification of AI-output deficiencies and the quality of ethical reflection. The portfolio is also an authentic assessment: it resembles a set of working papers through which an accounting practitioner explains a problem, evaluates alternatives and documents a justified recommendation (Gulikers et al., 2004; Villarroel et al., 2018).

There is a difference between just being able to use a chatbot and being literate in AI. Ng et al. (2021) describe AI literacy as consisting of understanding, application, evaluation and ethical considerations. Long and Magerko (2020) describe AI literacy in a similar way. AI literacy in the context of AIS refers to creating prompts to reflect a specific business process, recognizing when a model provides unsupported or generic information, and verifying the recommendations a model generates. AI literacy also entails protecting sensitive organizational information.

The evaluation component is important in understanding AI systems, especially large language models, since they often create plausible language, but this does not mean the information is accurate. Kasneci et al. (2023) highlight the risks of bias, privacy and over-reliance and the educational benefits of individualized learning. Similarly, Lo (2023) notes that pedagogy and safeguards are important in preventing the overreliance on AI in education. Academic integrity is also important, and the authors argue that AI is still powerful enough to invade the academic process and student judgement. To assess authentic student work, processes and precautions are important to make authorship and judgment evident, and to ascertain students are not simply relying on AI.

Table 2. Alignment with learning outcomes and MQF 2.0 domains

Domain	Intended learning evidence	Assessment evidence
Cognitive skills	Analyse a transaction cycle, identify control weaknesses and justify improvements	Scenario test, process flowchart and risk-control matrix
Digital skills	Use GenAI purposefully while maintaining an audit trail of prompts and responses	Prompt log and AI-output evaluation
Ethics and professionalism	Address confidentiality, authorship, bias and human accountability	Ethical reflection and AI-use declaration
Communication	Present system risks and controls clearly for decision makers	Portfolio narrative and short presentation
Personal and interpersonal skills	Contribute to group inquiry and provide structured peer feedback	Workshop observation and peer-review record

Table 2 explains how the portfolio assessment was designed to correspond with relevant Malaysian Qualifications Framework (MQF 2.0) learning domains. Cognitive skills are operationalised through transaction-cycle analysis, risk identification and justified internal-control recommendations; the associated evidence consists of the scenario test, process flowchart and risk–control matrix. Digital skills are represented by purposeful GenAI use and the preservation and examination of prompts and outputs. The link between the skill and its evidence is important because it assesses students' capacity to evaluate AI-generated information rather than merely their access to, or frequency of use of, AI tools.

Ethics and professionalism are addressed through reflection on confidentiality, bias, authorship and continuing human accountability for final recommendations. Communication is reflected in the clarity of the portfolio narrative and presentation, while personal and interpersonal skills are addressed through collaborative inquiry, peer feedback and workshop participation. This alignment is consistent with constructive alignment as discussed by Biggs (1996): intended outcomes should be visible in the activities and assessment criteria. However, the table is a curriculum-alignment matrix, not a statistical validation of MQF attainment. Neither individual-level achievement by MQF domain nor external accreditation judgements can be inferred from this mapping alone.

METHODOLOGY

3.1 Research design

The unit of quantitative analysis was the individual student. The primary comparisons involved each participant’s pre-intervention and post-intervention scores on identical outcome constructs, allowing within-person changes to be examined. Portfolio ratings were obtained after the instructional sequence and were treated as measures of performance within the intervention, not as an independent experimental manipulation.

Qualitative reflections and observations served a complementary explanatory purpose by identifying the aspects of GenAI-assisted learning that students and classroom observers considered salient.

The mixed-method interpretation follows a convergent logic at the discussion stage: numerical learning changes are read alongside patterns of reflection and observed participation. However, convergence should not be equated with proof of causation, particularly because both self-reported scales and classroom behaviours may be influenced by the same instructional setting.

The study uses a single-group, pre-test/post-test mixed-method design. The quantitative strand evaluates within-student differences in AIS achievement, AI literacy and learning engagement and examines their associations with portfolio quality. The qualitative strand examines students' reflections and classroom-observation records.

3.2 Participants and setting

The cohort included students at two academic levels, offering some variation in prior exposure to accounting subjects while retaining a common AIS learning setting. Group discussion was used to support peer feedback, but the individually submitted portfolios and individual outcome measures are essential to the interpretation of the reported sample size (N = 40). Consequently, the results characterise the participating class and should not automatically be generalised to other universities, AIS syllabi or student populations.

The participant characteristics reported in Table 5 describe the final analytical cohort. A journal-ready audit trail should additionally identify the sampling approach, eligibility criteria, any invitations or non-participation and whether all 40 students provided complete paired outcome data. These details cannot be reconstructed from the aggregate results alone.

The reported sample comprises 40 Bachelor of Accounting students enrolled in an AIS subject, with 24 in Year 2 and 16 in Year 3, aged 20–23 years. Students worked in groups of five for discussion and peer feedback, while submitting individual portfolios and individual pre/post assessments.

3.3 AIS case and instructional intervention

The revenue-cycle scenario was suitable for assessing the connection between AIS concepts and practical control decisions because the listed weaknesses cut across authorisation, segregation of duties, information accuracy, access management and reconciliation. For example, shared accounts obscure individual accountability; unauthorised price adjustments and refunds expose the entity to revenue leakage or misstatement; and delayed bank reconciliation weakens timely detection of errors or irregularities. A competent response therefore requires more than naming a control: students must specify the control objective, the responsible actor, the point of implementation and the evidence that would demonstrate its operation (COSO, 2013).

The intended pedagogical progression was cumulative. Students first represented the process, then identified risks, interrogated AI-generated proposals, revised their analysis after peer review and finally integrated the work into an individual portfolio. The sequence aims to turn GenAI into a revisable artefact that makes students' reasoning visible. The actual dates, model version, implementation fidelity and any departures from this plan should be reported using delivery records rather than inferred from the planned weekly stages.

The six-week design centres on a fictitious e-commerce entity, Nusa Digital Trading Sdn. Bhd., whose revenue cycle incorporates shared logins, insufficient segregation of duties, price overrides, weak refund authorisation, duplicate invoicing, delayed bank reconciliations and inappropriate access to customer information. No real business data are required. Students are guided to map the process, generate an initial AI analysis, verify every material claim against AIS principles, revise a risk-control matrix and justify their final professional recommendation. Table 3 presents the instructional sequence described in the study:

Table 3. Six-week intervention plan

Week	Phase	Learning activity	Evidence produced
1	Baseline and orientation	Consent process; AIS pre-test; AI-literacy and engagement pre-surveys; explanation of permitted AI use and privacy safeguards	Baseline evidence and AI-use agreement
2	Business-process analysis	Map the revenue cycle using a systems flowchart; identify actors, documents, data stores and decision points	Draft flowchart and process narrative
3	Risk and control inquiry	Identify threats to validity, completeness, accuracy, authorisation, custody and confidentiality	Initial risk register
4	GenAI interrogation	Develop prompts, preserve responses and identify omissions, unsupported claims and inappropriate controls	Prompt log and AI-error review
5	Reperformance and peer review	Construct a risk-control matrix; test whether each control addresses a defined risk; exchange rubric-based feedback	Revised matrix and feedback response
6	Portfolio and reflection	Submit individual portfolio; complete post-test and post-surveys; focus-group reflection; peer/expert observation	Final portfolio and outcome evidence

Table 3 describes the six-stage instructional sequence underlying the GenAI-assisted digital portfolio. In Week 1, students were to establish baseline AIS, AI-literacy and engagement scores and receive orientation on permitted AI use and privacy. In Week 2, the learning task shifted to representing the revenue cycle through a process flowchart, making transaction actors, documents and data flows explicit. Week 3 required identification of system weaknesses and their implications for authorisation, completeness, accuracy, custody and confidentiality. These stages established the accounting and internal-control knowledge needed to assess any suggestions subsequently generated by AI.

The central AI-specific activity occurs in Week 4, where students are required to preserve prompts, examine responses and identify inaccurate, incomplete or insufficiently contextualised recommendations. Week 5 then introduces reperformance and peer review: the student must revise the risk–control matrix and explain why each proposed control addresses a defined weakness. Week 6 concludes with an individual portfolio, follow-up assessments and reflective evidence..

3.4 Alternative assessment task

Equal weighting of the five portfolio criteria gives each component a 20% contribution to the total score. This makes critical evaluation of AI outputs and ethical reflection visible as assessed learning outcomes, rather than optional commentary appended to a technical answer. The risk–control matrix is especially important because it provides a traceable link from a described transaction weakness to a specified response and its rationale. A high-quality portfolio should therefore demonstrate control appropriateness and evidential reasoning as well as technical presentation.

To minimise construct overlap, portfolio scoring should distinguish genuinely independent student evaluation from merely polished AI-derived prose. For reproducibility, the full rubric, marking guidance, anonymisation procedure and any assessor calibration or second marking should accompany the article or supplementary materials.

The portfolio comprises six deliverables: (1) a business-process model, (2) the initial GenAI prompt and response, (3) an annotated audit of errors, omissions and unsupported claims, (4) a risk-control matrix, (5) a revised, evidence-supported management recommendation and (6) an ethical reflection with an AI-use declaration. The draft is intentionally preserved to reveal the learner's verification decisions. Each individual portfolio is assessed using the five criteria in Appendix B, each worth 20 points: process analysis, risk-control

identification, AI-output evaluation, control recommendations and ethical reflection. The rubric bands are excellent (17–20), competent (13–16), developing (9–12) and insufficient (0–8).

The assessment focuses at an AI-assisted level where students are permitted to use GenAI for exploration and initial drafting of their response, but must verify, integrate, and render final judgement on the response. This minimizes incentive to conceal AI use, as the history of the AI prompts is assessed. This also limits outsourcing, because students must identify gaps and omissions of the AI response and explain their revision recommendations with respect to evidence.

3.5 Instruments and measures

The three outcome measures capture related but distinct dimensions of learning. The AIS scenario assessment targets application of technical knowledge to internal-control situations and is reported on a 0–100 scale. The six-item AI-literacy instrument captures perceived capability in understanding, using, critically evaluating and ethically managing AI; the five-item engagement measure captures participation, effort, persistence, application and reflection. Because the latter two instruments use agreement responses, their improvements should be interpreted as changes in self-reported scale scores rather than direct proof of observable AI proficiency or behavioural engagement.

The portfolio rubric provides an additional performance-based measure that is conceptually different from the brief pre/post scales. Reflection and observation records further contribute explanatory information but should not be treated as independent objective measures of learning without appropriate coding and observation procedures. Items, scoring rules, scale anchors and any adapted language should be made available so other researchers can replicate the measurement approach.

Table 4. Data sources and measurement approach

Instrument	Format	Construct/evidence	Basis
AIS internal-control test	20 scenario-based items converted to a 0-100 score	Process understanding, risk recognition, control selection and justification	AIS subject experts; alignment with the selected CLO
AI-literacy scale	6 items, five-point agreement scale	Understanding, prompting, application, critical evaluation and ethical use	Conceptually adapted from Ng et al. (2021) and Long and Magerko (2020)
Learning-engagement scale	5 items, five-point agreement scale	Participation, effort, persistence, application and reflection	Researcher-developed for the intervention
Digital portfolio rubric	Five criteria scored 0-20; total 0-100	Process analysis, risk-control identification, AI-output evaluation, recommendations and ethical reflection	Authentic assessment and AIAS principles
Peer observation	Four dimensions scored 1-5 plus narrative notes	Active participation, collaboration, questioning of AI and use of evidence	CAFEL classroom-observation requirement
Reflective responses	Open-ended survey and focus group	Perceived usefulness, limitations, confidence and ethical concerns	Researcher-developed prompts

Table 4 identifies the multiple measurement instruments used to capture outcomes and the learning process. The AIS internal-control test comprises 20 scenario-based items converted to a score between 0 and 100. It addresses practical process understanding, risk recognition, appropriate control selection and justification, making it the principal discipline-specific outcome for H1. AI literacy is assessed through six five-point items conceptually linked to understanding, prompting, application, critical appraisal and ethics (Long & Magerko, 2020; Ng et al., 2021). Learning engagement is assessed through five five-point items concerning participation, effort, persistence, application and reflection. Thus, H2 and H3 examine scale-based outcomes that are related to, but distinct from, the AIS achievement measure.

The digital portfolio has a different function: it provides an authentic performance measure scored across five equally weighted criteria, resulting in a total of 0–100. Peer observations and open-ended reflections contribute complementary descriptive and qualitative evidence regarding collaborative reasoning, questioning of AI outputs, confidence and ethics.

3.6 Validity, reliability and trustworthiness

Internal consistency is one component of instrument quality, not a substitute for validity. Cronbach's alpha summarises how consistently scale items covary in this cohort; it does not establish that the items fully represent AI literacy or learning engagement. Content alignment with the AIS learning outcomes, review of item wording, comparable pre/post administration conditions and documentation of any scale adaptation are also needed. Reliability of the portfolio scoring similarly depends on transparent decision rules and independent assessor evidence rather than the total mark alone.

For the qualitative component, the trustworthiness of interpretations depends on a documented progression from raw student responses to codes and themes. An appendix with de-identified examples, a codebook, treatment of ambiguous responses and evidence of a second-coder review, if actually undertaken, would increase transparency. Frequencies should describe the number of students whose responses were assigned to a category; because responses can express more than one theme, category percentages need not sum to 100% (Braun & Clarke, 2006).

The AIS case, test, survey and analytic rubric were specified for alignment with subject learning outcomes and established AIS/internal-control concepts. The reported Cronbach's alpha coefficients are .838 and .921 for AI literacy at pre-test and post-test, respectively, and .852 and .876 for engagement.

3.7 Data-collection procedure

The documented intervention framework contains six weekly stages: orientation/baseline assessment; revenue-cycle mapping; identification of AIS risks and controls; GenAI prompt and output evaluation; peer review and reperformance; and final portfolio submission/post-assessment. Exact delivery dates, class timetable, the AI platform and model used, account/access arrangements, deviations from the schedule and the sequence of consent and grading are mandatory empirical-reporting details that remain to be inserted from contemporaneous records. Research participation is voluntary and independent of course grading, with withdrawal available without academic penalty.

3.8 Data analysis

For each quantitative outcome, the primary test compares the two repeated measurements from the same students. The paired-sample t-test evaluates whether the average within-student difference differs from zero, while the Shapiro–Wilk test relates to the distribution of those differences rather than the marginal distribution of pre-test and post-test scores. Confidence intervals express uncertainty around the mean change, and Cohen's dz standardises the average change by the standard deviation of the paired differences (Lakens, 2013).

Pearson correlation was used to examine whether students receiving higher portfolio scores also tended to show larger improvements in AIS performance and AI literacy. Such coefficients describe linear association rather than predictive causality and may be influenced by baseline ability, assessment alignment and measurement error. The significance threshold for statistical inference was $p < .05$; given the small class cohort, substantive interpretation also considers effect magnitude, uncertainty and consistency across measures.

The specified analysis comprises descriptive statistics, internal-consistency reliability, normality assessment of paired differences, paired-sample t-tests, complementary Wilcoxon signed-rank tests, 95% confidence intervals, standardised within-person differences and Pearson correlations between total portfolio score and learning gains.

RESULTS

4.1 Participant profile

Table 5. Reported participant characteristics (N = 40)

Characteristic	Category	n	%
Gender	Female	26	65.0
	Male	14	35.0
Study year	Year 2	24	60.0
	Year 3	16	40.0
Prior AI use	Daily	10	25.0
	Weekly	19	47.5
	Monthly/rarely	11	27.5
Age	20	9	22.5
	21	20	50.0
	22	9	22.5
	23	2	5.0

Table 5 presents the characteristics of the 40 students included in the reported analysis. Female students accounted for 26 participants (65.0%), while 14 (35.0%) were male. Twenty-four students (60.0%) were in Year 2 and 16 (40.0%) were in Year 3. The age distribution was concentrated around 21 years: nine students were aged 20 (22.5%), 20 were aged 21 (50.0%), nine were aged 22 (22.5%) and two were aged 23 (5.0%). The totals for each characteristic sum to the sample size, indicating that the reported demographic categories are internally complete. The relatively narrow age range is consistent with the undergraduate course context but limits direct transfer of findings to working accountants, mature learners or students at very different educational levels.

Prior AI experience is especially relevant to the intervention. Ten students (25.0%) reported daily AI use, 19 (47.5%) reported weekly use, and 11 (27.5%) reported monthly or rare use. In aggregate, 29 students (72.5%) were already using AI at least weekly. Accordingly, the intervention primarily examined structured, accountable use of GenAI within an AIS assessment rather than first-time exposure to AI tools. Existing familiarity may have shaped baseline confidence, willingness to refine prompts and speed of completing activities, but these possibilities were not tested using the available aggregate data. The demographic table should therefore be interpreted as a description of the participating cohort, not as evidence that gender, study year, age or previous AI exposure caused the observed outcomes.

The sample size also informs statistical interpretation. With N = 40, the paired comparisons retain all 40 observations only if each student has both time-point scores. Percentages in this table use the cohort as the denominator, and each respondent corresponds to 2.5 percentage points. Although a single-course cohort can provide useful evidence for a classroom intervention, the non-random and institution-specific nature of the sample restricts generalisability beyond comparable AIS teaching settings.

4.2 Reliability and assumption checks

Table 6. Internal-consistency estimates

Scale	Administration	Items	Cronbach's alpha
AI literacy	Pre	6	0.838
AI literacy	Post	6	0.921
Learning engagement	Pre	5	0.852
Learning engagement	Post	5	0.876

Table 6 reports Cronbach's alpha values for the two multi-item student-response scales at baseline and after the intervention. The six-item AI-literacy measure had alpha values of .838 before the intervention and .921

afterwards. These figures indicate substantial reported internal consistency, with stronger inter-item covariance in the post-test responses. The five-item learning-engagement scale yielded alpha values of .852 and .876 for the two administrations, again indicating consistently related item responses. All four coefficients exceed the commonly used .70 screening level for internal consistency, although alpha alone cannot establish unidimensionality, content adequacy, or the absence of response bias.

The higher post-test alpha for AI literacy should not automatically be interpreted as evidence that the instrument became more valid or that the intervention improved reliability. Alpha can change with item covariance, respondent heterogeneity and score distributions. Because item-level data, item-total correlations and factor-analytic information are not included in the manuscript, the reported estimates cannot be independently reproduced or decomposed. The results support a cautious conclusion that the scales displayed acceptable reported internal consistency at both measurement points, conditional on the source analysis being accurate.

The normality checks reported alongside the table concern the distribution of paired change scores, which is the relevant distribution for the paired-sample t-test. Shapiro–Wilk statistics were W = .971 (p = .391) for AIS performance, W = .968 (p = .308) for AI literacy and W = .977 (p = .565) for learning engagement. Because each p-value exceeds .05, these tests did not identify a statistically significant departure from normality. Such non-significant results do not prove perfect normality; Q–Q plots, distributional summaries and assessment of outlying changes would be useful supplementary checks.

4.3 Pre-post outcome analysis

Table 7. Reported paired sample results

Outcome	Pre M (SD)	Post M (SD)	Mean change	95% CI	t(39)	p	Cohen dz
AIS performance (0-100)	59.18 (7.93)	71.23 (11.69)	12.05	9.63, 14.47	10.09	< .001	1.60
AI literacy (1-5)	2.83 (0.56)	3.64 (0.72)	0.80	0.63, 0.98	9.28	< .001	1.47
Learning engagement (1-5)	3.05 (0.53)	3.53 (0.59)	0.48	0.32, 0.64	6.23	< .001	0.98

Table 7 presents the principal hypothesis-testing results using paired observations from 40 students. For AIS internal-control performance, the pre-test mean was 59.18 (SD = 7.93) on the 0–100 scale, rising to a post-test mean of 71.23 (SD = 11.69). The reported average within-student gain of 12.05 points was statistically significant, t(39) = 10.09, p < .001, with a 95% confidence interval of 9.63 to 14.47 points. The confidence interval contains only positive differences and therefore supports the conclusion that average post-test AIS scores were higher in this cohort under the stated test assumptions. Cohen's dz = 1.60 represents a very large standardised paired difference. The source manuscript also mentions four students with non-positive AIS changes, indicating that improvement was not universal. The larger post-test standard deviation suggests increased variation in students' final scores, although a paired analysis would require individual data to explain that dispersion.

This AIS result is compatible with H1, expressed as a within-cohort comparison. From an educational perspective, a 12.05-point increase may reflect increased familiarity with revenue-cycle risks, more systematic connection of weaknesses to controls and repeated opportunities to revise decisions. Nevertheless, the instructional programme combined ordinary teaching, case analysis, peer feedback and AI-supported work. A one-group pre-test/post-test design cannot separate the incremental contribution of GenAI from exposure to the curriculum, practice with similar questions, maturation or the test–retest process. Statistical significance should not be conflated with causal identification.

AI-literacy scores increased from a reported pre-test mean of 2.83 (SD = 0.56) to 3.64 (SD = 0.72) on a five-point scale. The table reports a mean gain of 0.80, t(39) = 9.28, p < .001, 95% CI [0.63, 0.98] and Cohen's dz = 1.47. This is a large within-person standardised change and is consistent with higher post-intervention ratings of understanding, prompting, evaluation and ethical AI use. The displayed rounded means differ by 0.81, rather than 0.80, so the exact mean difference requires confirmation against unrounded data or the paired difference

scores. This is a rounding or reporting issue to be reconciled, not a basis for substituting a new value without the source records. H2 is supported in terms of the reported within-group score difference, although the self-report scale does not by itself prove an equivalent improvement in objectively observed AI competence.

Learning engagement likewise increased, from M = 3.05 (SD = 0.53) before the intervention to M = 3.53 (SD = 0.59) afterwards. The mean difference was 0.48, $t(39) = 6.23$, $p < .001$, 95% CI [0.32, 0.64], with Cohen's dz = 0.98. Students therefore reported more engagement at the end of the six weeks, providing within-cohort support for H3. The result is compatible with an active-learning environment involving peer review, iterative revision and a tangible portfolio product, but engagement is multidimensional. The reported aggregate scale score cannot reveal which individual engagement items improved most, and it should not be treated as a direct measure of sustained participation outside the observed classes.

Across the three outcomes, the positive confidence intervals, p-values below .001 and sizeable standardised changes provide a consistent pattern of higher post-intervention scores. Cohen's dz is calculated using the standard deviation of paired differences and is not directly interchangeable with the standardised mean difference for two independent groups (Lakens, 2013).

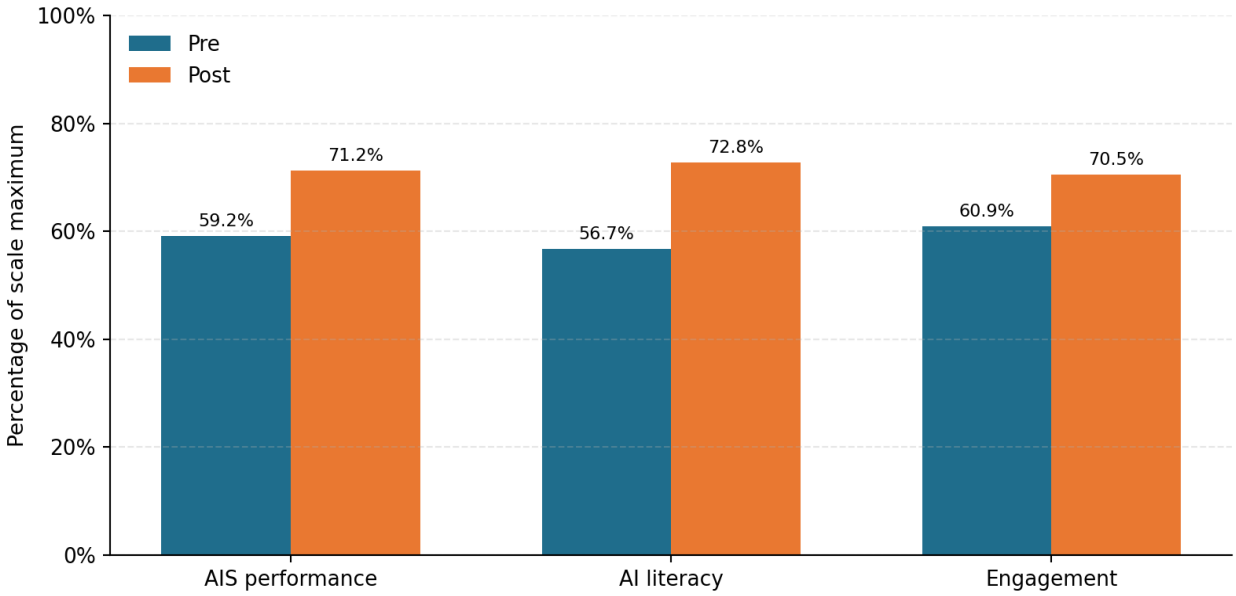

Figure 2. Pre-post outcomes expressed as a percentage of each scale maximum

Figure 2 restates the pre-test and post-test means using the reported maximum of each instrument as the denominator, allowing the three outcomes to appear on a common 0–100 display. AIS performance increases from approximately 59.2% to 71.2% of its 100-point maximum. AI literacy increases from approximately 56.6% to 72.8% of the five-point maximum, while learning engagement increases from approximately 61.0% to 70.6%. Minor differences between these calculations and plotted labels may be due to unrounded source means. The chart conveys the consistent upward direction across all three outcomes.

4.4 Portfolio performance and associations

Table 8. Reported digital-portfolio criterion scores

Portfolio criterion	Mean	SD	% of maximum
Process analysis	15.55	1.57	77.8
Risk-control identification	16.10	1.78	80.5
AI-output evaluation	15.35	1.94	76.8
Control recommendations	15.83	1.62	79.1
Ethical reflection	15.75	1.28	78.8
Total portfolio	78.58	5.99	78.6

Table 8 describes how the students performed on the five components of the individual digital-portfolio assessment. Each criterion had a maximum of 20 points. Risk–control identification recorded the highest mean of 16.10 (SD = 1.78), equivalent to 80.5% of its maximum. This indicates that the strongest average performance was achieved in identifying weaknesses and connecting them to relevant internal controls. Process analysis had a mean of 15.55 (SD = 1.57; 77.8%), while control recommendations averaged 15.83 (SD = 1.62; 79.1%). These results suggest broadly similar performance in process mapping and in translating identified risks into proposed responses.

AI-output evaluation received the lowest reported criterion mean, 15.35 (SD = 1.94; 76.8%). This is substantively important because the educational intervention explicitly aimed to move students beyond uncritical acceptance of fluent automated explanations. A comparatively lower mean on this criterion is consistent with the possibility that recognising omissions, challenging assumptions and explaining why an apparently convincing AI recommendation was inappropriate required more demanding evaluative reasoning. Ethical reflection averaged 15.75 (SD = 1.28; 78.8%), suggesting that students generally demonstrated the ability to articulate at least some considerations concerning disclosure, confidentiality and responsible AI use.

The total portfolio mean was 78.58 out of 100 (SD = 5.99), with a reported observed range of 65–89. This implies generally satisfactory outcomes for the selected marking rubric, though the score distribution and individual criterion bands are not supplied. The five reported criterion means sum to 78.58, which is consistent with the equal-weight scoring system described in the methodology.

Table 9. Associations between portfolio performance and learning gains

Relationship	r	df	t	p	Interpretation
Portfolio total with AIS gain	.450	38	3.109	.004	Significant positive association
Portfolio total with AI-literacy gain	.314	38	2.038	.049	Small-to-moderate positive association

Table 9 examines the relationship between total digital-portfolio performance and changes in independently reported learning outcomes. The association between portfolio score and AIS test improvement was positive and moderate, $r = .450$, $t(38) = 3.109$, $p = .004$. The degrees of freedom of 38 are consistent with a Pearson correlation based on 40 complete paired observations. Substantively, students with higher portfolio scores tended also to show larger AIS score gains. Squaring the correlation gives approximately .203, indicating that portfolio marks and AIS gains shared about 20.3% of their observed linear variation in this bivariate sample.

The correlation between portfolio performance and AI-literacy gain was smaller but positive, $r = .314$, $t(38) = 2.038$, $p = .049$. The corresponding squared correlation is approximately .099, or 9.9% shared linear variation. As its p-value is close to .05, this result is more sensitive to sampling variability and the handling of multiple statistical comparisons than the AIS correlation. Both reported relationships support the direction predicted in H4 at an unadjusted .05 threshold. However, students who perform better on authentic assessments may also differ in motivation, prior proficiency, effort or learning opportunities.

4.5 Qualitative themes and classroom observation

Table 10. Reflective-response theme categories and reported frequencies

Theme	n	%	Analytic interpretation
Verification and critical evaluation of AI outputs	29	72.5	Fluent AI output initially appeared credible, but verification revealed missing or weakly matched controls.
Prompt refinement encouraged deeper questioning	27	67.5	Writing better prompts required students to describe the process and evidence more precisely.
Greater confidence in internal-control analysis	25	62.5	The risk-control matrix helped students connect each weakness to a defensible response.

Efficiency accompanied by risk of over-reliance	22	55.0	AI accelerated initial analysis but could encourage acceptance of confident inaccuracies.
Ethical, privacy and authorship concerns	18	45.0	Students recognised that confidential information should not be uploaded and final judgement remained human responsibility.

Table 10 summarises five reported themes derived from student reflections concerning their experience of GenAI-supported AIS learning. Verification and critical evaluation of AI-generated outputs was the most common category, reported for 29 students (72.5%). This suggests that a substantial portion of the cohort recognised the need to examine apparently plausible machine-generated advice for omitted risks, unsupported assumptions or weak matches between controls and business-process problems. Prompt refinement and deeper questioning appeared in 27 responses (67.5%), indicating that students described the construction of more precise prompts as part of clarifying the process, evidence and requested analysis. Greater confidence in internal-control evaluation was identified for 25 students (62.5%), consistent with the proposition that a structured risk–control matrix can help learners defend their decisions.

A further 22 students (55.0%) were categorised as recognising both the efficiency benefits of GenAI and the danger of over-reliance. This is an important distinction: reducing the time spent generating an initial response does not necessarily mean that the response is accurate, contextually appropriate or professionally defensible. Ethical, privacy and authorship concerns were recorded for 18 students (45.0%), suggesting awareness of confidentiality and the continuing responsibility of students to disclose AI assistance and exercise their own judgement.

Table 11. Reported classroom-observation ratings

Observation dimension	Rating (1-5)	Interpretation
Active participation	4.5	Strong evidence
Collaborative reasoning	4.4	Strong evidence
Questioning of AI outputs	4.3	Strong evidence
Use of evidence and AIS concepts	4.2	Strong evidence
Overall mean	4.35	High observed engagement

Table 11 reports four classroom-observation ratings on a five-point scale. Active participation received the highest score (4.5), followed by collaborative reasoning (4.4), questioning of AI outputs (4.3) and use of evidence and AIS concepts (4.2). Their unweighted arithmetic mean is 4.35, matching the overall value shown in the table. As descriptive observations, the ratings depict a classroom in which participation and peer interaction were positively assessed. They also indicate that questioning AI output and explicitly grounding judgements in AIS evidence were present in the observed learning activities.

The slightly lower rating for evidence use is consistent with the portfolio results in Table 8, where the AI-output evaluation criterion also had the lowest mean. Taken together, these sources suggest a potential teaching priority: supporting students to cite relevant process evidence and explain how a control directly addresses a specified risk.

4.6 Summary of hypothesis findings

The primary within-participant analyses reported statistically significant higher post-intervention scores for AIS performance, AI literacy and learning engagement. The direction of each difference corresponds with H1, H2 and H3, respectively, while the portfolio correlations are positive in the directions specified by H4.

The quantitative, performance-based and reflective evidence sources provide complementary perspectives on a structured GenAI-assisted AIS assessment. Their convergence is informative, but not all sources are equally independent: engagement and AI literacy are self-reported, portfolio quality is a course-specific graded outcome, and classroom observation may involve lecturer judgement. The present evidence therefore supports an account of changes and associated experiences within one cohort, rather than a definitive claim that GenAI alone caused those outcomes.

DISCUSSION

5.1 Interpretation of the AIS performance improvement

One substantive interpretation of the AIS gain is that repeated exposure to the same conceptual chain transaction event, information risk, control objective, proposed response and justification may have improved students' ability to apply internal-control principles to a case. The portfolio required students to revisit their early representations, identify weaknesses in preliminary AI suggestions and articulate defensible recommendations. This sequence supports active engagement with accounting judgement rather than passive reproduction of definitions, consistent with the emphasis on curricular integration in Apostolou et al. (2014) and Sledgianowski et al. (2017).

The educational significance of the 12.05-point difference should be distinguished from the statistical test. The change is substantial on the assessment scale, but interpretation depends on equivalence of pre/post test difficulty, marking consistency and the amount of directly relevant practice students received. Without a non-GenAI comparison group, the study cannot disentangle the effects of instructional time, assessment familiarity, feedback and the AI-assisted verification component.

The reported 12.05-point AIS gain is compatible with the pedagogical rationale of repeated risk assessment, evidence checking and control redesign. However, even after source-data verification, a one-group design does not isolate GenAI from course instruction, repeated testing, maturation or assessment familiarity. The magnitude of $dz = 1.60$ should be assessed against the intervention's duration, baseline knowledge and similarity of training and test tasks.

This design innovates upon existing AIS curriculum recommendations and focuses more on students' evaluative judgment rather than students' operative/technical skills. Sledgianowski et al. (2017) and Apostolou et al. (2014) underscore the importance of integrating case-based teaching of AIS. The present portfolio includes an up-to-date requirement, students must analyze an automated explanation of their recommendation and provide validity evidence as to why their recommendation is more reliable than the AI's recommendation.

5.2 Development of AI literacy

The measured increase supports the argument that AI literacy can be addressed as part of disciplinary reasoning rather than as a standalone lesson in operating a chatbot. Students encountered the practical requirement to specify prompts, preserve a trace of generated recommendations and test those recommendations against a business-process scenario. This corresponds with the evaluative and ethical dimensions of AI literacy described by Long and Magerko (2020) and Ng et al. (2021).

At the same time, a self-reported increase in literacy may reflect increased familiarity or confidence as well as genuine improvement in competence. A stronger future design would include a separate blinded performance test in which students must locate factual errors, explain privacy risks and correct inappropriate AI-generated controls in an unfamiliar scenario. Such a task would assess transfer of learning beyond the specific portfolio case.

The retained AI-literacy increase is compatible with domain-embedded practice in prompting, checking accuracy, identifying unsupported claims and ethical AI use. Its interpretation depends on confirmed instrument administration and the reliability of the underlying scores. A follow-up assessment would clarify whether any change is retained beyond the intervention period.

These conclusions correlate with the accounting education literature. Sundkvist and Kulset (2024) and Zhou and Luo (2025) indicate students believe there is educational worth in utilizing ChatGPT. Blondeel et al. (2025) indicate students use ChatGPT to obtain academic help. Converting the use of AI in the classroom to educational value is the challenge. Wiseman and Foster (2025) describe reperformance and critical error-review. These are useful as they encourage students to identify when and where AI reasoning fails and to construct an adequate professional analysis.

5.3 Student engagement

Engagement may have benefited from the active nature of the tasks, including peer discussion, assessment of competing recommendations and revision of a tangible work product. Unlike an assignment based solely on a final written response, a digital portfolio makes intermediate decisions visible and provides opportunities for formative feedback. This design is consistent with the educational value of active learning and authentic assessment (Freeman et al., 2014; Villarroel et al., 2018).

The observation mean of 4.35/5 offers additional descriptive support for a participatory classroom environment, but it should not be interpreted as an objective validation of the change-score measure. Both may be affected by the lecturer's expectations and the novelty of AI activities. More rigorous studies could combine standardised engagement instruments with attendance, revision histories, time-on-task information and independent observers.

5.4 Portfolio quality and learning gains

The positive associations suggest that portfolio quality may serve as an informative indicator of how successfully students integrate AIS reasoning with the evaluation of AI assistance. Students who could present a coherent process analysis, specify relevant control responses and explain their revisions were also more likely to achieve larger gains on independent outcomes. Such relationships are theoretically consistent with constructive alignment: the assessed process and the intended learning outcomes emphasise common underlying skills (Biggs, 1996).

The moderate AIS association and smaller AI-literacy association should not be compared as evidence that the portfolio improves one competence more than another without a formal test of dependent correlations. Nor can the results rule out the possibility that motivated students performed well on both assessments. Where individual records become available, baseline-adjusted analysis and criterion-level correlations may help clarify these mechanisms, although the sample of 40 would require parsimonious statistical models.

The retained correlations suggest that a well-developed portfolio can be evaluated in conjunction with independently assessed learning gains. Criterion-specific regression could clarify whether AI-output evaluation or risk–control identification is more strongly associated with change, but this would require student-level data, diagnostics, transparent model selection and cautious inference in a small sample.

There is a high level of acceptance because the portfolio reflects an accountable workpaper. It retains the source of the automated suggestions, the evidence used to assess it, the revisions, and the student's conclusion. The design of the portfolio is also aligned to Villarroel et al's (2018) and Gulikers et al's (2004) focus on Evaluative judgement, and addresses some of the integrity concerns relating to academic dishonesty.

5.5 Theoretical implications

First, the findings are compatible with constructive alignment insofar as the instructional activities, rubric criteria and outcome measures centred on practical AIS analysis and evaluated AI use. Second, they are consistent with authentic-assessment principles because students had to document professional judgements in a recognisable workpaper format rather than submit an unexamined AI answer. Third, the reflective evidence points towards AI literacy as a socially and ethically situated capability, integrating critical assessment, source verification and human accountability rather than mere prompt fluency (Ng et al., 2021; Gulikers et al., 2004).

These are plausible pedagogical explanations rather than formal tests of competing learning theories. The study does not measure mediating psychological constructs or test a structural causal mechanism. Its theoretical value lies in illustrating how principles from these literatures can be brought together in a concrete, assessable AIS learning design.

The study integrates constructive alignment, active learning, and authentic assessment with AI literacy and AIS intervention. Constructive alignment justifies the need to align assessment with the intended analytical performance. Active learning justifies the use of discussion and peer assessment and reperformance. Authentic

assessment explains the use of a risk-control portfolio. AI-literacy theory justifies the importance of assessing the student's evaluation of the Ethics and AI. In line with the theories mentioned above, there is a belief that GenAI is most powerful when learning design builds into the student the responsibility to frame the problem, evaluate the evidence, and defend their final judgement.

5.6 Practical implications for AIS lecturers

Implementation should begin with a clearly bounded AIS scenario and an explicit policy describing permitted GenAI assistance. The lecturer can give students a prompt-log template, a checklist for checking generated claims against course evidence and a risk–control matrix in which every recommendation must be linked to a specified weakness. Formative peer review is most useful when students are required to defend or revise their proposed controls rather than simply exchange general comments.

The descriptive weaknesses observed in AI-output evaluation suggest an opportunity to incorporate short calibration exercises. Students might compare a plausible but incomplete AI answer with a stronger evidence-based solution, identify missing authorisation controls or unsupported claims, and explain how the recommendation should change. This would reinforce that analytical judgement and accountability remain with the student, consistent with human-centred AI guidance (Miao & Holmes, 2023).

AIS lecturers can implement this intervention without having to restructure their entire course. One topic or transaction-cycle can be converted into a four to six week portfolio. Lecturers are encouraged to provide an approved case, state the level of AI-use, supply a prompt-recording template, verification checklist and a rubric that rewards discipline knowledge. The lecturer’s role is shifted from providing model answers to providing guidance in determining whether the answer is valid, the evidence is accurate and how the control fits in the process.

There are several safeguards that have to be in place. The case has to be fictitious, preferably set in the future, to ensure no confidential information is provided. Students have to be made aware of the fact that AI outputs may be incorrect or biased. The focus of assessment has to be the reasoning of the student, and the final evaluation, rather than the fluency of the students’ writing. If oral defence is possible, this can ensure that the student is the author of the answer and clarify any information that is not clear to the lecturer. The rubric that is used for final assessment of the portfolio, has to be the same that is used for peer assessment, to enable students to have an understanding of expected quality throughout the intervention.

LIMITATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

An additional consideration is the distinction between data reported in this manuscript and independently reproducible analysis. The current document preserves aggregate tables, reliability coefficients, normality statistics and qualitative category counts, but does not include the student-level paired scores, item responses, portfolio marking sheets, de-identified reflections or original observation logs. As a result, the numerical results can be discussed as reported empirical findings but cannot be independently re-estimated or audited from the manuscript alone. Before publication, the research team should reconcile these tables with the underlying records and document ethics approval or institutional authorisation, consent, data protection, test equivalence, marking procedures and the actual delivery dates.

Future research should also address multiple-comparison risk and the potential for common-method bias when self-reported constructs are measured repeatedly in the same teaching context. A controlled comparison using an otherwise similar non-AI portfolio or a delayed AI intervention, combined with a performance-based assessment of AI-output critique, would improve attribution. A larger multi-institutional study could examine whether prior AI experience, year of study and baseline AIS proficiency moderate learning outcomes, and a delayed test could determine whether students retain the capacity to apply risk–control reasoning to new transaction cycles.

The one-group design additionally remains vulnerable to maturation, repeated testing, concurrent instruction, assessment alignment and lecturer expectancy effects. A delayed-intervention comparison between class

sections or a non-AI case-based comparison condition could help distinguish GenAI-specific mechanisms while maintaining student equity. Delayed transfer and retention tests would further evaluate persistence of learning. The cohort size limits precision and transferability, and large reported dz values (1.60, 1.47 and 0.98) should not be assumed plausible empirical effects without auditing score distributions, assessment difficulty and prior competence.

Implementation may also be affected by uneven paid/free GenAI access, changes in model versions, internet availability, digital-literacy differences, prior AI familiarity, time spent preserving prompt trails and substantial staff workload for rubric marking. Students may use GenAI to generate their critique or ethical reflection, compromising claims about authorship. Mitigations include in-class verification tasks, timestamped draft histories, source-checking demonstrations, a short oral defence, calibrated double marking and documented privacy training. Larger classes may require staged peer review and systematic marker moderation; different AIS topics, including expenditure cycles, payroll and information security, would need contextualised cases and rubrics. Equity, algorithmic bias, privacy and the environmental footprint of computation warrant explicit consideration, following UNESCO's human-centred guidance (Miao & Holmes, 2023).

CONCLUSION

This study examined a six-week GenAI-assisted digital-portfolio activity among 40 Bachelor of Accounting students studying accounting information systems. The reported pre/post results showed improvements in AIS performance (mean change = 12.05 points), AI literacy (0.80 points) and learning engagement (0.48 points), all statistically significant at $p < .001$. Portfolio quality was positively associated with AIS learning gains ($r = .450$, $p = .004$) and AI-literacy gains ($r = .314$, $p = .049$). The portfolio rubric further indicated that risk–control identification was students' strongest criterion on average, while critical evaluation of AI-generated outputs remained comparatively more demanding.

The reported student reflections and classroom observations add context to these quantitative findings. Students described value in revising prompts, verifying generated claims and translating identified risks into justifiable control recommendations, while also recognising the risks of over-reliance, confidentiality breaches and unclear authorship. Together, these results suggest that a carefully structured portfolio can create opportunities for accounting students to practise both AIS judgement and accountable engagement with GenAI.

The central implication is not that unrestricted AI use improves learning, but that AI can be incorporated into assessment in a way that makes human reasoning, evidence checking and ethical accountability visible. Because the research involved a single cohort and lacked a comparison condition, observed changes cannot be attributed uniquely to GenAI. Furthermore, underlying student-level records and qualitative evidence should be checked to substantiate the reported aggregate results before journal submission. Subject to those boundaries, the study offers a transferable pedagogical model for integrating AI literacy with authentic internal-control analysis in undergraduate accounting education.

REFERENCES

1. AACSB International. (2024). Building future-ready business schools with generative AI. https://www.aacsb.edu/insights/reports/building-future-ready-business-schools-with-generative-ai
2. Al-Htaybat, K., von Alberti-Alhtaybat, L., & Alhatabat, Z. (2018). Educating digital natives for the future: Accounting educators' evaluation of the accounting curriculum. Accounting Education, 27(4), 333-357. https://doi.org/10.1080/09639284.2018.1437758
3. Apostolou, B., Dorminey, J. W., Hassell, J. M., & Rebele, J. E. (2014). A summary and analysis of education research in accounting information systems (AIS). Journal of Accounting Education, 32(2), 99-112. https://doi.org/10.1016/j.jaccedu.2014.02.002
4. Ballantine, J. A., Boyce, G., & Stoner, G. (2024). A critical review of AI in accounting education: Threat and opportunity. Critical Perspectives on Accounting, 99, 102711. https://doi.org/10.1016/j.cpa.2024.102711
5. Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher Education, 32, 347-364. https://doi.org/10.1007/BF00138871

6. Blondeel, E., Bullock, T., Gaskin, J., Schuetzler, R., Serre, R., Steffen, J., Wells, T. M., & Wood, D. A. (2025). The effects of generative artificial intelligence (GenAI) on learning in an accounting data analytics course. Journal of Accounting Education, 72, 100987. https://doi.org/10.1016/j.jaccedu.2025.100987
7. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa
8. Centre for Academic Advancement and Flexible Learning. (2025). Guideline for UNITEN academics: Transformative teaching and alternative assessments (Version 1.0). Universiti Tenaga Nasional.
9. Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education, 20, 38. https://doi.org/10.1186/s41239-023-00408-3
10. Committee of Sponsoring Organizations of the Treadway Commission. (2013). Internal control-integrated framework: Executive summary. COSO.
11. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228-239. https://doi.org/10.1080/14703297.2023.2190148
12. Damerji, H., & Salimi, A. (2021). Mediating effect of use perceptions on technology readiness and adoption of artificial intelligence in accounting. Accounting Education, 30(2), 107-130. https://doi.org/10.1080/09639284.2021.1872035
13. Dong, M. M., Stratopoulos, T. C., & Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. International Journal of Accounting Information Systems, 55, 100715. https://doi.org/10.1016/j.accinf.2024.100715
14. Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 111(23), 8410-8415. https://doi.org/10.1073/pnas.1319030111
15. Furze, L., Perkins, M., Roe, J., & MacVaugh, J. (2024). The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI-supported assessment. Australasian Journal of Educational Technology, 40(4), 38-55. https://doi.org/10.14742/ajet.9434
16. Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. Educational Technology Research and Development, 52, 67-86. https://doi.org/10.1007/BF02504676
17. Kasneci, E., Sessler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Gunnemann, S., Hullermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeiffer, F., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274
18. Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863
19. Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), 410. https://doi.org/10.3390/educsci13040410
20. Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727
21. Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693
22. Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041. https://doi.org/10.1016/j.caeai.2021.100041
23. Sledgianowski, D., Gomaa, M., & Tan, C. (2017). Toward integration of Big Data, technology and information systems competencies into the accounting curriculum. Journal of Accounting Education, 38, 81-93. https://doi.org/10.1016/j.jaccedu.2016.12.008

24. Sundkvist, C. H., & Kulset, E. M. (2024). Teaching accounting in the era of ChatGPT: The student perspective. Journal of Accounting Education, 69, 100932. https://doi.org/10.1016/j.jaccedu.2024.100932
25. Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. Assessment & Evaluation in Higher Education, 43(5), 840-854. https://doi.org/10.1080/02602938.2017.1412396
26. Wiseman, D., & Foster, D. (2025). Bridging AI with cost accounting education: A pedagogical framework using reperformance and critical error review. Journal of Accounting Education, 72, 100988. https://doi.org/10.1016/j.jaccedu.2025.100988
27. Zhou, A., & Luo, Y. (2025). Exploring the impact of generative AI on student learning in accounting. Journal of Accounting Education, 72, 100982. https://doi.org/10.1016/j.jaccedu.2025.100982