STEM Education

https://www.aimspress.com/journal/steme

STEM Education, 6(4): 584–605
DOI:10.3934/steme.2026024
Received: 22 February 2026
Revised: 19 April 2026
Accepted: 15 May 2026
Published: 10 June 2026

*Research article*

# Embedding assurance within learning: Empirical evidence from the SAGE framework for repositioning take-home assessment in AI-integrated higher education

**Mahmoud Elkhodr*** **and Ergun Gide**

School of Engineering and Technology, Central Queensland University, Australia

* **Correspondence:** Email: m.elkhodr@cqu.edu.au.

Academic Editor: Shuangji Yao

**Abstract:** This study examined whether a mature, empirically validated generative artificial intelligence (GenAI) intervention framework can produce reliable process evidence when deployed in unsupervised take-home assessments. Twenty-five group submissions from two cybersecurity management cohorts, designed under the Structured AI-Guided Education (SAGE) framework, were audited using the five-check SAGE Audit Protocol. Only 3 of 25 submissions (12%) produced evidence chains that were substantially auditable, and full traceability between documented AI outputs and human evaluation claims was not achieved in any submission. Across both cohorts, the remaining submissions showed mismatches between human-authored tables and AI outputs, generic compliance text, workflow-focused reflections, and process records that were often indistinguishable from reconstructed accounts. The paper identifies a compliance gradient in which conscientious students who follow the process in good faith incur a disproportionate documentation burden, while simulated compliance can produce comparable outputs with less effort. It also highlights a marker's dilemma, where auditing AI-supported process evidence approximately doubles marking time and shifts educators from assessing learning to interpreting logs. The paper argues that conventional unsupervised take-home assessment can no longer function as sufficient standalone assurance in the GenAI era. Rather than abandoning AI-integrated learning or defaulting to low-technology examinations, the findings support assessment architectures in which assurance is structurally embedded. The Defend step subsequently added to the SAGE framework operationalises this principle.

**Keywords:** generative AI, higher education assessment, SAGE framework, academic integrity, take-home assessment, AI orchestration, cybersecurity education

## 1. Introduction

Higher education is perhaps the industry most profoundly disrupted by generative artificial intelligence. The rapid uptake of the technology has created significant pressure on institutions worldwide to respond. GenAI tools have evolved rapidly since ChatGPT first appeared in late 2022. They are no longer limited to producing text. Contemporary models generate multimedia, build software applications, write and debug code, search the web in real time, read and synthesise research papers, and produce references with improving (though not yet fully reliable) accuracy [1]. They are also effective coding tools and increasingly capable of reasoning across domains. Institutional capacity to respond has been outpaced by the rate of model improvement.

Within the next few years, the role of GenAI in higher education is likely to become far less contested. It will become increasingly prominent, normalised, and expected. AI will become what the calculator became to arithmetic. A tool so embedded in professional practice that refusing to use it would be considered a liability rather than a taboo. Today, no one is shamed for being unable to perform long division by hand. The same trajectory is underway for AI-assisted analysis, writing, coding, and problem-solving. The question is not whether students will use these tools. The question is whether institutions will equip them to use the tools responsibly and correctly, or whether we risk leaving students to figure it out on their own.

Institutional responses to GenAI have ranged across a wide spectrum. Some universities initially focused on compliance and monitoring, including the use of AI detection approaches and the reinforcement of academic integrity controls [2]. Confidence in such approaches, however, weakened as concerns grew about whether AI-generated work could be identified reliably in authentic assessment contexts [3, 4]. Others emphasised governance by developing institution-wide guidelines and policy structures for the use of GenAI across teaching, learning, research, and administration [5–7]. A smaller number have attempted to move beyond governance alone by structuring the pedagogical adoption of AI in ways that make its use more explicit, scaffolded, and educationally purposeful rather than merely permitted or prohibited [8–10]. Fewer still have subjected such responses to sustained empirical scrutiny across authentic educational contexts [1, 11–13].

The Structured AI-Guided Education (SAGE) framework was developed through a multi-year program of empirical research. Initial work examined how information and communication technology (ICT) students interacted with ChatGPT under structured conditions, reporting performance improvements of 56 to 87 per cent alongside persistent gaps in student evaluation competency [11]. Subsequent studies extended this evidence across data analytics, cybersecurity management, and systems analysis and design, involving more than 800 students across four Australian campuses [12–17]. More recent work identified distinct typologies of student AI use in permitted assessments and highlighted a competency-confidence inversion, in which students demonstrated increasingly sophisticated interaction strategies while simultaneously expressing regulatory anxiety [15]. This emerging sophistication in students' use of GenAI provided an important rationale for the present study.

SAGE is, at its core, a pedagogy. It builds on the principle of embracing AI rather than rejecting it. The framework aims to empower students by giving them the skills required to verify and validate AI output, to work with AI not as an output machine but as a collaborator, and to exercise professional judgement over when to accept, modify, or reject what the AI produces. In this sense, SAGE

positions human-to-AI collaboration as an emerging professional competency, one that extends the traditional boundaries of human-computer interaction into a domain where the machine generates substantive content that requires human evaluation. The framework operationalises this through a five-step cycle: Generate (use AI to produce an initial output), Evaluate (compare the output against domain standards, regulations, and authoritative sources), Refine (modify the output with evidence-based reasoning, documenting what was changed and why), AI Critic (ask AI to take on a specific persona and re-evaluate your modifications), and Reflect (conduct metacognitive analysis of the AI's strengths, limitations, and failure patterns). This cycle is enacted across two stages: scaffolded tutorial practice (Stage 1) and independent assessment application (Stage 2).

In this paper, we report the findings from our latest implementation of SAGE across two cybersecurity management units, one postgraduate (PG) and one undergraduate (UG), at Central Queensland University. The assessments were designed using the full SAGE protocol. That is, the assessment architecture followed all SAGE stages and incorporated structured templates, AI interaction logs, accept/modify/reject decision tables, and mandatory reflection components. This design produced a rich collection of process artefacts, including the prompts students used, the AI outputs they received, the modifications they made, and the justifications they provided for each decision. These artefacts offer direct evidence of how students interact with AI when working within a structured intervention framework such as SAGE.

We conducted a systematic five-check audit of 25 group submissions across one undergraduate and one postgraduate cohort. The audit examined primary evidence, traceability between tables and appendices, internal data consistency, modification provenance, and reflection specificity. The findings are confronting. Only 3 of 25 submissions (12%) produced evidence chains that were substantially auditable. The remaining submissions exhibited logical checksum failures between human-authored tables and recorded AI outputs, alongside structural indicators consistent with audit-trail simulation as a plausible risk. These patterns were observed across both the PG and UG cohorts.

These findings suggest that students often approached tasks intended to foster critical thinking and AI orchestration as routine take-home assignments that could themselves be completed with AI assistance. The requirement to document AI-human interaction was intended to make the learning process visible, yet in many submissions it functioned primarily as a reporting format rather than as credible evidence of evaluative engagement. Collectively, these patterns point to a structural gap between what open, AI-integrated assessment affords pedagogically and what it can reliably demonstrate for assurance. The issue is not the learning design itself, but the assumption that an unsupervised artefact, however thoughtfully scaffolded, can serve simultaneously as a vehicle for learning and as standalone evidence of authentic student judgement. Addressing that gap requires not the abandonment of open assessment, but its repositioning within assessment designs where assurance is embedded through observable and defensible acts of reasoning. Accordingly, this study is guided by the following research question:

*To what extent do unsupervised take-home assessments produce reliable evidence of authentic student engagement when designed with embedded process logging and reporting requirements?*

To this end, the paper makes the following three contributions.

1. It introduces a replicable five-check audit protocol for verifying AI process artefacts in structured assessments. The protocol operates on internal document consistency rather than probabilistic AI detection, making it transparent, explainable, and independent of any specific detection tool.

2. It provides empirical evidence, drawn from 25 group submissions across two cohorts, that process documentation, alone, under unsupervised take-home conditions does not reliably assure individual attainment, even when the assessment is designed with structured scaffolding, mandated templates, and explicit logging requirements.
3. It identifies a structural incentive problem, termed the compliance gradient, in which conscientious students who follow the process in good faith incur a disproportionate documentation burden, while simulated compliance can produce comparable or superior outputs with less effort. This gradient is not specific to SAGE. It is a property of any assessment that relies on self-reported process evidence under unsupervised conditions in an environment where GenAI tools are available.

The remainder of this paper is organised as follows. Section 2 reviews related work on AI integration frameworks, assessment reform guidance, and the emerging consensus around supervised assurance. Section 3 describes the method, including the assessment context, the SAGE audit protocol, and the analytical approach. Section 4 presents the findings as five empirically grounded themes. Section 5 discusses the findings and their implications, while Section 6 presents the recommendations. Section 7 outlines the study limitations. Directions for further research and concluding remarks are provided in Section 8.

## 2. Related work

Contemporary higher education research has moved beyond the prohibition of GenAI tools toward their structured or regulated integrations. For example, the work in [8] proposed a comprehensive policy framework coordinating pedagogical, governance, and operational dimensions, explicitly linking assessment redesign to institutional capability-building. Other work, such as [9], introduced the Artificial Intelligence Assessment Scale. It translates permissible AI use into discrete levels intended to reduce ambiguity and normalise ethical disclosure. While works such as [10] formalised assessment adaptation as an iterative institutional process, foregrounding proactive task design and review cycles rather than ad hoc enforcement.

These frameworks share a common epistemological assumption that the students will engage cooperatively with the transparency requirements imposed on them. In [18], the authors directly challenged this assumption, arguing that purely discursive controls are structurally fragile because they rely on behaviours that students remain free to ignore. Their call for structural assessment changes rather than policy refinement provides an important precursor to the present study.

Recent integrity scholarship treats GenAI as a qualitative escalation of contract cheating. Leaton Gray et al. [19] argued that AI lowers cost, increases plausible deniability, and scales the production of plausible academic prose. Research in [20] reviewed assessment guidelines from top-ranking universities and found that most recommend stress-testing tasks against what AI can produce, with some advocating the integration of AI into assessment processes to reduce incentives for covert outsourcing. The implication is that a fluent product can no longer be safely equated with authentic cognition. This ultimately strengthens the case for process-anchored evidence.

This literature signals a pivot toward process evidence, with metacognitive commentary increasingly treated as a primary marker of genuine engagement. A cognitive mirror framework in which AI serves as a reflective partner that rewards explanation quality over answer production was

proposed in [21]. However, measurement remains the key issue. Reflective assessments are typically graded through narrative self-reporting, and, from what we have observed in this study, GenAI can plausibly produce procedurally convincing reflections. Similarly, the authors in [22] showed that log-based process data can reveal response patterns and latent traits. However, they also highlighted the need to validate process indicators in specific assessment contexts. Therefore, the literature gap is not the absence of frameworks that mandate process documentation. Rather, it is the absence of empirical evidence that such documentation reliably distinguishes genuine engagement from simulated compliance when submitted under unsupervised conditions.

To address this gap, the present study audited 25 group submissions from two cybersecurity management cohorts in which assessments were designed using the SAGE framework [12], a structured AI integration pedagogy validated across prior empirical studies [12–15]. The assessment architecture mandated AI interaction logs, accept/modify/reject decision tables, and reflective commentary—producing the exact category of process artefacts that the literature positions as the solution to AI-era integrity challenges. A five-check audit (The SAGE Audit Protocol) was applied to test whether these artefacts, when produced under unsupervised take-home conditions, constituted reliable evidence of the process they claimed to document.

## 3. Research design and methods

SAGE Stage 1 is a tutorial-based and scaffolded pedagogy. Students practise evaluating GenAI outputs against provided domain standards, research evidence, or discussions with project stakeholders, refining them with evidence, and reflecting on the process under instructor guidance. Stage 2 is assessment-based and progressively independent. Students apply the same cycle autonomously in summative tasks with reduced scaffolding. The core cycle operationalises five steps (Generate, Evaluate, Refine, AI Critic, Reflect), with explicit emphasis on verification against authoritative sources and evidence-based accept, modify, or reject decisions documented in structured templates [12]. The framework is documented in the SAGE Implementation Guide [23], with current resources, sample templates, and the policy page maintained on the framework website [24].

The assessment under audit was a Stage 2 task. Students were required to develop a risk register for a fictional healthcare organisation (Koala Health), propose risk treatment strategies aligned to the Australian Cyber Security Centre (ACSC) Essential Eight [25], and respond to a third-party data breach incident scenario. The task was designed with industry validation which confirmed that the security challenges reflected contemporary risks across the technology sector. At each stage, students were required to generate AI-assisted outputs, critically evaluate them against the Kim et al. telehealth threat taxonomy [26] and the Essential Eight maturity model, and document their accept, modify, or reject decisions with professional justification in structured reporting tables. Complete AI interaction logs, including prompts and unedited AI responses, were mandated as appendices.

The unit of analysis was each submitted report together with its appendices. The appendices were expected to contain raw AI interaction artefacts (prompts and complete AI outputs), which served as primary evidence for the audit. Group identifiers and student names were removed from the analysis narrative. All findings are reported at the group submission level.

The study sample comprised 25 group submissions drawn from two cybersecurity units offered at an Australian regional university during the third teaching term of 2025: COIT12212 Cybersecurity

Management (undergraduate) and COIT20263 Information Security Management (postgraduate). Groups were formed by the instructor and comprised four students each, yielding a total participant pool of 100 students across both cohorts. Ten submissions were drawn from the undergraduate cohort and fifteen from the postgraduate cohort. The assessment task examined was identical in design across both cohorts, with postgraduate students expected to demonstrate a higher level of critical engagement with the domain literature in their evaluation and reflection components. The marker responsible for the audit was also the designer of the assessment and the instructor of record for both units, a dual role that is acknowledged as a source of interpretation bias and is discussed further in the limitations. This study was conducted under the standard ethical provisions applicable to analysis of existing assessment submissions in Australian higher education and did not require separate ethics committee approval.

### 3.1. The SAGE Audit Protocol and analytical approach

The SAGE Audit Protocol, developed for this study, comprises five checks designed to assess not merely whether AI had been used, but whether the documented process evidence was internally consistent, traceable, and substantive.

1. **Check 1: Primary evidence presence.** This check assessed whether AI interaction evidence was provided as raw artefacts (complete prompts and unedited AI outputs) rather than paraphrased summaries or narrative reconstructions. Raw artefacts are essential because they provide verifiable primary evidence.
2. **Check 2: Traceability.** This check assessed whether each AI critique claim appearing in the main-body tables (the accept, modify, or reject columns) could be traced to a specific artefact in the appendices. A traceable submission allows the marker to locate the exact AI output that informed each documented decision.
3. **Check 3: Internal data consistency (logical checksum).** This check tested whether the assets, threats, risk ratings, and evaluative claims presented in the human-authored tables were consistent with the content of the supplied AI artefacts. Specifically, it examined whether a table referenced elements that never appeared in the recorded AI outputs, and whether a submission attributed a position to the AI tool that was not supported by any recorded output. This check functions as a logical checksum: if the human evaluation column states that the AI "underestimated" a particular risk, the AI output in the appendix should contain a rating for that risk that is demonstrably lower than the human rating.
4. **Check 4: Modification provenance.** This check assessed whether there was evidence of what was changed between the AI output and the final submission, and why. Adequate provenance required that rationales be tied to specific standards (e.g., ACSC Essential Eight controls), case-study constraints (e.g., Koala Health's legacy authentication infrastructure), or recognised control frameworks (e.g., the Kim et al. threat taxonomy) rather than generic assertions.
5. **Check 5: Reflection specificity.** This check distinguished between functional reflections (those containing domain-specific judgement, identification of specific AI errors, and documented corrective reasoning) and procedural reflections (those describing workflow benefits such as time savings or avoidance of a blank page, without identifying domain errors or corrective steps).

Although the checks are presented conceptually as a sequence, they were applied analytically to all submissions for marking and cross-submission comparison. Failure on an earlier check did not

preclude examination of subsequent checks, as each submission still required evaluation across the full audit matrix.

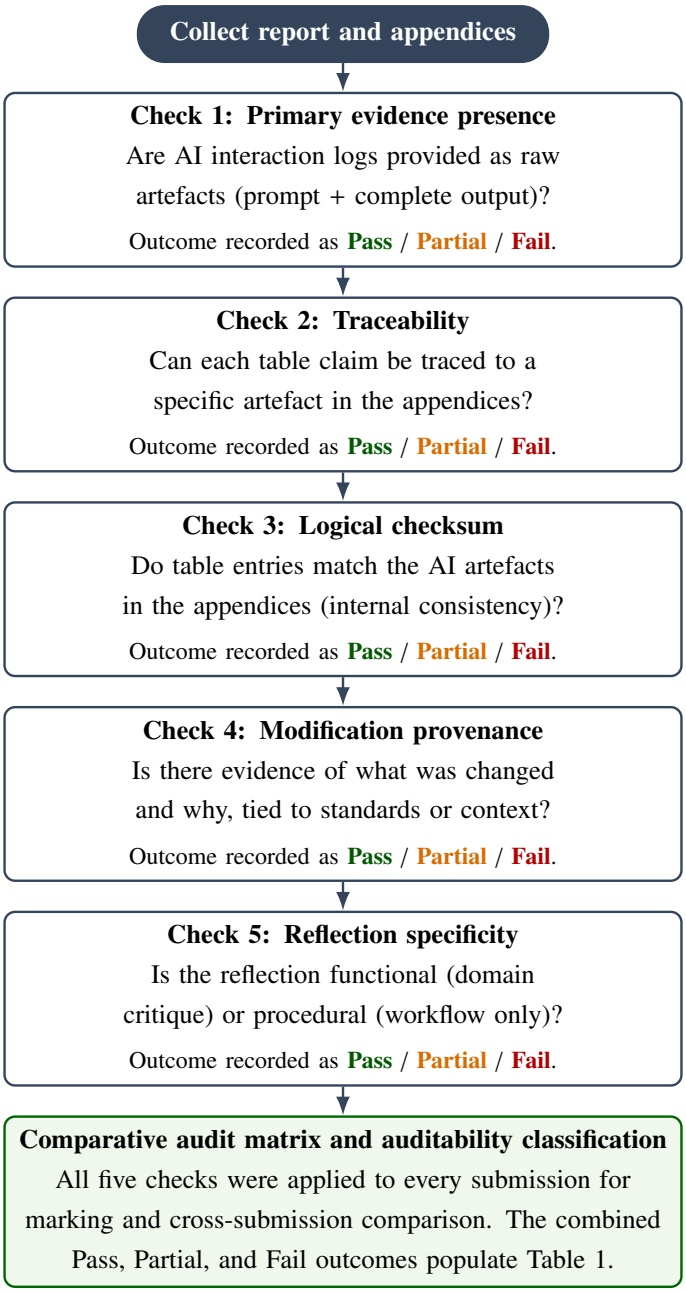


**Figure 1.** The SAGE five-check audit protocol. Although presented sequentially, all five checks were applied to every submission for marking and comparative audit purposes.

The analysis proceeded iteratively. Each submission was first marked against the assessment rubric, then re-examined using the five-check protocol. Patterns were coded inductively and consolidated into themes after the seventh submission. The remaining submissions were then examined deductively against the established themes. The marker was also the designer of the assessment, which introduces a potential interpretation bias acknowledged in the limitations. However, the audit checks are designed to be replicable. They test observable properties of submitted documents rather than subjective quality

judgements.

## 4. Findings

Five themes emerged from the audit. They are presented below with representative examples drawn from the de-identified dataset, which are also summarised in Table 1. Across 125 individual checks (25 submissions $\times$ 5 checks), 17 checks (14%) were passed, 44 (35%) were partial, and 64 (51%) were failed. Only 3 of 25 submissions (12%) produced evidence chains that were substantially auditable. Full Traceability (C2) compliance was the weakest dimension, with 0 passes across the full dataset.

### 4.1. Only 12% of submissions produced substantially auditable evidence chains

Of the 25 submissions analysed, only 3 provided a near-complete evidence map connecting AI artefacts to human modifications across the submitted assessment components. These submissions generally included raw AI prompts and outputs in the appendices and provided stronger evidence of modification provenance and reflection specificity than the rest of the dataset. However, they did not achieve full traceability across every table entry. Some links between risk-register or treatment-matrix claims and appendix artefacts remained incomplete, non-verifiable, or dependent on marker interpretation. For this reason, these submissions are classified as substantially auditable rather than fully traceable.

The remaining submissions exhibited varying degrees of incompleteness. Several provided partial artefacts, such as truncated AI outputs or selected excerpts rather than complete interaction logs. Others replaced primary evidence with narrative descriptions of the AI interaction, stating in prose what the AI had recommended without providing the raw output. In these cases, the marker was required to accept the student's account of the AI output on trust, as no verifiable primary evidence was available. This created a substantial verification burden. On average, marking each submission under the SAGE audit protocol required approximately twice the time needed for conventional marking, owing to the need to cross-reference tables against appendices and to identify gaps in the evidence chain.

**Table 1.** Audit Results Matrix.

| Group | C1 Evidence | C2 Traceability | C3 Checksum | C4 Provenance | C5 Reflection | Note |
|---|---|---|---|---|---|---|
| ***Cohort A: Undergraduate (n = 10 groups, 4 students per group)*** | | | | | | |
| **G1** | Pass | Fail | Pass | Fail | Fail | |
| **G2** | Partial | Fail | Partial | Fail | Pass | |
| **G3** | Pass | Fail | Pass | Pass | Pass | a |
| **G4** | Partial | Fail | Fail | Partial | Fail | |
| **G5** | Fail | Fail | Fail | Fail | Fail | |
| **G6** | Partial | Fail | Fail | Partial | Fail | |
| **G7** | Pass | Partial | Pass | Pass | Pass | b |
| **G8** | Partial | Fail | Partial | Fail | Fail | |
| **G9** | Fail | Fail | Fail | Partial | Fail | |
| **G10** | Partial | Fail | Fail | Fail | Partial | |
| ***Cohort B: Postgraduate (n = 15 groups, 4 students per group)*** | | | | | | |
| **G11** | Partial | Fail | Fail | Partial | Partial | |
| **G12** | Partial | Fail | Fail | Partial | Fail | |
| **G13** | Pass | Partial | Partial | Partial | Partial | |
| **G14** | Fail | Fail | Fail | Partial | Partial | |
| **G15** | Partial | Partial | Partial | Partial | Pass | |
| **G16** | Pass | Partial | Pass | Pass | Pass | a |
| **G17** | Partial | Fail | Fail | Fail | Fail | |
| **G18** | Fail | Fail | Fail | Partial | Fail | |
| **G19** | Partial | Fail | Partial | Fail | Partial | |
| **G20** | Partial | Fail | Fail | Partial | Fail | |
| **G21** | Fail | Fail | Fail | Fail | Fail | |
| **G22** | Partial | Partial | Fail | Partial | Partial | |
| **G23** | Partial | Fail | Fail | Fail | Fail | |
| **G24** | Fail | Fail | Fail | Partial | Partial | |
| **G25** | Partial | Fail | Partial | Fail | Fail | |
| **Summary** | **Pass: 5<br>Partial: 14<br>Fail: 6** | **Pass: 0<br>Partial: 5<br>Fail: 20** | **Pass: 4<br>Partial: 6<br>Fail: 15** | **Pass: 3<br>Partial: 12<br>Fail: 10** | **Pass: 5<br>Partial: 7<br>Fail: 13** | |

*Note.* Five-check protocol applied to 25 group submissions. Pass = satisfied, Partial = incomplete, and Fail = absent or non-verifiable. C1 = primary evidence presence; C2 = traceability; C3 = logical checksum; C4 = modification provenance; C5 = reflection specificity.
Superscript a indicates submissions that were substantially auditable across most checks but not fully traceable.
Superscript b indicates an incomplete submission due to group member withdrawal; submitted components passed all assessable checks.

### 4.2. Logical checksum failures revealed internal inconsistencies

A recurring empirical pattern was internal inconsistency between the human-authored tables and the AI artefacts provided in the appendices. The assessment required students to generate an AI-assisted risk assessment for Koala Health, a fictional Australian healthcare network. Students were then required to critically evaluate that AI output against the Kim et al. telehealth threat taxonomy and their textbook's risk management methodology, producing a risk register of 12 to 15 critical assets. For each asset, students documented the AI's original risk rating alongside their own adjusted rating, with a justification column explaining why they accepted, modified, or rejected the AI's assessment. This design created a natural checksum: the AI position recorded in the appendix should correspond to the AI position as reported in the table.

In several submissions, the risk register introduced assets, threat categories, or organisational constraints that were not present in the recorded AI output. For example, in one case, a table entry referenced a device-use asset category (such as Bring Your Own Device policies) and attributed a specific risk position to the AI tool. However, when the appendix was examined, the recorded AI output did not contain that asset category. In another case, the AI comparison column stated that the AI had "underestimated" a particular risk, but the appendix contained no rating for that asset from which an underestimation could be inferred.

These are not simple documentation omissions. They are verifiability failures. The submission cannot support its own process claim. It is important to note that a definitive reason for these inconsistencies cannot be established from document evidence alone. The log may have been incomplete because an interaction was conducted but not recorded. The asset may have been introduced through a subsequent AI interaction that was not appended. Alternatively, the table may have been constructed or embellished without a corresponding AI interaction. What can be stated is that the submitted evidence does not verify the submitted claim. The distinction matters.

### 4.3. Human evaluation cells frequently functioned as compliance text

Across the dataset, table cells labelled as human judgement (the "AI Comparison and Justification" and "Response (A/M/R)" columns) were frequently populated with polished, generic justificatory language. The text in these cells often did not cite a concrete AI error, a specific correction, or a context-specific trade-off tied to Koala Health's documented circumstances. Instead, a recurring pattern involved formulaic statements such as "the AI underestimated the risk due to the complexity of the healthcare environment" or "AI did not fully account for regulatory requirements" without specifying which regulatory requirement, which AI output was deficient, or what the corrective action entailed.

This pattern was consistent with treating the evaluation column as a formatting requirement, a cell to be filled, rather than as a decision record that documents a genuine evaluative act. The pattern was also consistent with the use of an AI tool to generate or polish the human evaluation text itself, although this cannot be established definitively from the submitted documents. It is equally possible that some students genuinely attempted the evaluation but lacked the domain depth to produce specific justifications, or that time pressure led to superficial completion. Regardless of cause, the effect is the same: the evaluation cells in these submissions did not provide evidence of the critical thinking and domain-specific reasoning that the assessment was designed to elicit.

### 4.4. Reflection quality separated into procedural versus functional modes

Two distinct reflection modes were observed across the 25 submissions.

Procedural reflections described workflow-level benefits of AI use. Characteristic statements included references to avoiding a blank page, accelerating initial drafting, and generating comprehensive starting points. These reflections rarely identified a specific domain error in the AI output and rarely documented a corrective step grounded in cybersecurity standards or case-study constraints. At least nine of the 25 submissions exhibited predominantly procedural reflection.

Functional reflections, by contrast, critiqued domain content directly. They evaluated the adequacy of the AI's threat coverage against the Kim et al. taxonomy, questioned assumptions embedded in the AI's risk ratings, and assessed the suitability of recommended controls for Koala Health's specific operational context. Only three submissions combined functional reflection with broader auditability across the remaining checks. It is acknowledged that these submissions may reflect higher pre-existing competency, stronger prompt engineering skills, or greater familiarity with the SAGE process rather than a generalisable pedagogical outcome. The sample does not permit causal attribution.

A representative functional reflection in the auditable subset identified a misclassification of a threat category under the Kim et al. taxonomy and documented both the corrective reasoning and the authoritative source used to justify the revised position. The distinction was therefore not one of length or sophistication of expression, but of whether domain judgement was visible and traceable in the reflection itself.

### 4.5. Submissions showed indicators consistent with audit-trail simulation risk

Several submissions did not provide raw AI interaction logs. Instead, they presented third-person summaries of what the AI tool had produced, written in narrative form within the main body of the report. Statements such as "the AI recommended implementing multi-factor authentication across all clinical systems" appeared without a corresponding verbatim AI output in the appendices.

A related pattern was observed in submissions where the appendix did contain AI interaction content, but presented it as a single undifferentiated block rather than as demarcated exchanges. When a specific table claim, for example that the AI had recommended a particular control for a named asset, was traced back to the appendix, no corresponding interaction could be located. The claim existed in the table; the evidentiary link to it did not.

The reasons for this pattern cannot be determined from document evidence. Students may have conducted the AI interactions but failed to save or append the logs. They may have been uncertain about whether raw logs would attract penalisation. They may have been under time pressure and prioritised the main report over appendix completeness. They may also have constructed the narrative without a genuine underlying interaction. Intent cannot be established from this analysis alone.

The structural consequence is that a third-person narrative of an AI interaction can be produced without maintaining a genuine process record. This structure reduces auditability and increases the plausibility of reconstructed compliance. The term "audit-trail simulation" is used here not to assert that simulation occurred, but to describe a structural risk. The artefact format permits simulation to be indistinguishable from genuine but poorly documented process adherence. This is a design vulnerability in any assessment that relies on self-reported process evidence submitted under unsupervised conditions.

## 5. Discussion

The findings of this study are best understood not as a critique of AI-integrated pedagogy, but as a critique of the current assurance role assigned to conventional unsupervised take-home assessment. Within the SAGE framework, Stage 1 and Stage 2 were designed to make AI-supported reasoning more visible through structured generation, evaluation, refinement, critique, and reflection. What the present audit shows is that, when this pedagogical logic is expressed through unsupervised take-home artefacts, the resulting evidence is difficult to verify at scale and does not reliably function as sufficient standalone assurance of individual attainment. The issue, therefore, is not the educational value of the learning design itself, but the assumption that an open assessment, however carefully scaffolded, can serve simultaneously as a vehicle for learning and as conclusive evidence of authentic student learning.

### 5.1. Sector implications

The empirical evidence from the 25 submissions supports a broader concern that is increasingly visible in the higher education sector: submitted artefacts produced under open conditions are becoming more difficult to rely upon as credible assurance evidence in the GenAI era. TEQSA's 2023 discussion paper on assessment reform observed that non-invigilated tasks are difficult to design in ways that preclude substantial GenAI use [27]. More recent sector responses have similarly recognised the limits of relying on submitted products alone as proof of individual attainment. The present study contributes concrete empirical evidence to this discussion by identifying how those limitations manifest in practice through traceability failures, logical checksum mismatches between tables and appendices, compliance-pattern text in evaluation cells, procedural rather than functional reflection, and the structural feasibility of audit-trail simulation.

These mechanisms are not speculative. They were observed repeatedly across the dataset and documented through the SAGE five-check audit. In this respect, the study extends current policy discussion by showing not only that open assessment is increasingly difficult to assure, but also why. Importantly, the findings do not support a blanket return to low-tech examinations as the default institutional response. Rather, they point to the need for assessment designs in which assurance is more deliberately embedded within the learning process itself, through opportunities for students to make their reasoning observable, explainable, and defensible.

### 5.2. Compliance gradient: Honest effort is penalised

The dataset revealed a perverse incentive structure that warrants explicit discussion. Process documentation created a substantial workload for both students and markers. It also created unequal outcomes that disadvantaged conscientious students. Students who attempted to follow the full SAGE process in good faith produced lengthy appendices containing raw AI interaction logs, multiple prompt iterations, and detailed cross-referencing between their tables and the AI outputs. This effort was time-consuming, cognitively demanding, and produced long documents that were themselves difficult to navigate and consequently difficult to grade.

Students who did not follow the process with equivalent rigour could still submit polished tables and reflections with reduced or absent evidence. The quality of the final product, namely the tables and narrative, could be comparable or even superior because AI tools can generate or polish the ostensibly

human-authored components. This produces what we termed as a compliance gradient: honesty is costly in terms of effort and document length, while the simulation of compliance remains feasible and carries a lower production burden. This gradient is not specific to SAGE. It is a structural property of assessment designs that rely on self-reported process evidence under unsupervised conditions in the GenAI era.

### 5.3. The marker's dilemma

The audit revealed a structural difficulty for the marker that warrants explicit treatment because it has not been adequately documented in the assurance literature. In the present setting, the marker faced an uncomfortable and arguably unsustainable choice. Either the submitted product is accepted without credible assurance of the process that produced it, or the process is audited at scale, which imposes a marking burden that is not viable for large cohorts.

The empirical character of this dilemma can be specified. The audit reported here covered 25 group submissions. Even at this modest scale, the cross-referencing required between main-body tables and appendices approximately doubled the time per submission compared with conventional marking. The additional time was not consumed by content evaluation, which is the legitimate intellectual work of assessment. It was consumed by verification activity: locating which prompt produced which output, determining whether a table claim could be traced to a specific artefact, identifying whether an evaluation column reflected a genuine human judgement or compliance-pattern text, and recording the resulting gap for the audit. Across the 125 individual checks in this study, 51% returned Fail and 35% returned Partial. The marker was therefore engaged in gap documentation, rather than in the assessment of learning, for the majority of the time spent per submission.

Scaling this approach to cohorts of 50, 100, or 200 students is unlikely to be practical without dedicated auditing resources that are unavailable in most institutional contexts. The implication is not that the audit is wrong, but that the audit role cannot be sustainably performed by the unit coordinator within the existing marking workflow. The activities involved are structurally a different category of work. They are not the evaluation of student understanding against criteria, which is what marking has historically been. They are the interpretation of process evidence, the verification of evidence chains, the adjudication of edge cases, and the management of uncertainty about intent. These are the activities of an auditor or compliance officer, and they sit poorly on top of the existing pedagogical responsibilities of the unit coordinator.

A consequence worth naming explicitly is that the burden of maintaining assessment integrity under AI-integrated conditions has, in the current default arrangement, been transferred largely to individual educators rather than being supported through systemic redesign. Unit coordinators are expected to absorb the new interpretive load on top of their existing workload, often without commensurate workload recognition, additional resources, or institutional process. The audit reported here demonstrates that this transfer is not only unequal but quantitatively unsustainable at typical cohort sizes. The arrangement implicitly assumes that the educator can perform the audit role, the marking role, and the teaching role simultaneously. The empirical evidence indicates that the audit role alone is sufficient to consume the available marking time even at small scale.

There is a further qualitative dimension. When the marker's role shifts toward investigation, the tone of the teacher-student relationship shifts with it. Interpretive marking of the kind documented here requires the educator to make repeated judgements about whether a student's reflection is

genuine or formulaic, whether an evaluation cell reflects critical thinking or compliance, and whether documentation gaps reflect oversight, time pressure, or audit-trail simulation. These judgements are cognitively demanding, frequently inconclusive, and not pedagogically rewarding. Sustained over an entire cohort, they shift the dominant register of marking from formative evaluation toward suspicion management. This is not a sustainable register for higher education marking, nor is it consistent with the trust-based relationship that effective teaching requires.

It is also important to note that the SAGE Audit Protocol, while effective at identifying inconsistencies, does not produce definitive findings about student intent or behaviour. It identifies gaps and mismatches in the evidence chain. Determining why those gaps exist would require additional evidence, such as interviews, screen recordings, or supervised replication, that is not available in a standard marking workflow for take-home assessment. The method is therefore useful for diagnostic purposes but insufficient as a standalone assurance mechanism. This limitation does not weaken the empirical claim that the audit burden is structurally unsustainable. It reinforces it. The interpretive workload documented here is incurred even when the method cannot conclusively establish what occurred. Educators are asked to perform high-cost interpretation under conditions of irreducible uncertainty, which is the worst possible combination of demand and reward in any audit role.

The implication of the marker's dilemma for assessment design is taken up in Section 6. The short form of the argument is that the current default of expecting individual educators to perform process audit at scale, in addition to marking and teaching, is not a sustainable response to AI-integrated assessment. A different distribution of assurance work is required.

### 5.4. The contribution: An empirical warning and a design implication

The contribution of this paper is an evidence-based warning about the current role of take-home assessment in AI-integrated higher education. The study shows that even highly scaffolded take-home tasks, supported by structured templates, AI logs, evaluation tables, and reflective prompts, do not reliably yield sufficient standalone evidence of authentic student judgement under unsupervised conditions. This does not diminish the pedagogical value of open AI-integrated learning. Rather, it clarifies that pedagogy and assurance cannot be assumed to coincide within the same submitted assessment.

At the same time, the audit method used in this study did make inconsistencies visible without recourse to probabilistic AI detection tools. This is an important practical distinction. The SAGE audit checks operate on internal document consistency rather than statistical inference about authorship, which means they are transparent, replicable, and explainable to students. In that respect, the study contributes both an empirical warning and a more credible basis for rethinking how assurance should be designed within future AI-integrated assessment.

The broader implication is that take-home assessment should not be discarded, but repositioned. Open tasks remain valuable because they allow students to explore, iterate, draft, verify, and refine work in ways that reflect contemporary professional practice. What the present findings indicate is that such tasks can no longer bear the full burden of assurance when treated as terminal proof of individual attainment. More credible assurance will require assessment designs in which the outputs of open learning feed into subsequent opportunities for students to explain, justify, and defend key decisions in ways that make reasoning and ownership more directly observable. At a broader level, this aligns with program-level assessment design and distributed assurance approaches in which different tasks serve

different pedagogical and assurance purposes across the curriculum [28].

## 6. Recommendations

The findings of this study suggest that the problem exposed by GenAI is not simply that students can use new tools in unsupervised assessment. Rather, it is that many conventional take-home tasks can no longer be relied upon to verify authentic engagement or individual attainment when submitted as standalone artefacts under open conditions. At the same time, the findings do not support abandoning AI-integrated learning, nor do they justify a blanket return to examination-dominated assessment. Such a retreat would risk allowing assurance concerns to override educational design, narrowing the range of capabilities that higher education is able to foster and recognise.

What is needed instead is a more deliberate repositioning of take-home assessment within a learning architecture in which assurance is embedded, designed, and demonstrated through fit-for-purpose forms of defensible judgement. On this basis, four recommendations are offered.

### 6.1. Recommendation 1: Reposition take-home assessment rather than treating it as standalone proof of learning

The present findings indicate that take-home assessment in its conventional form should no longer be treated as sufficient evidence of individual attainment in AI-integrated higher education. Even where structured templates, AI logs, evaluation tables, and reflective components are mandated, the audit showed that traceability remained weak, internal consistency frequently broke down, and submitted artefacts often failed to verify the process they purported to document. These limitations do not mean that take-home assessment has lost its pedagogical value. They do mean, however, that its assurance role must be reconsidered.

Accordingly, take-home tasks should be repositioned as important developmental components within the broader learning arc rather than as self-sufficient proof of learning. Their value lies in enabling exploration, drafting, iterative refinement, evidence use, and authentic engagement with the kinds of AI-supported workflows students are likely to encounter beyond university. Used in this way, take-home assessment remains central to contemporary pedagogy. What must change is not its existence, but the expectation that an unsupervised final artefact can, by itself, certify that the underlying thinking is genuinely the student's own.

### 6.2. Recommendation 2: Treat structured AI pedagogy as the foundation, not the endpoint, of assurance

The findings of the present study should not be interpreted as evidence against frameworks such as SAGE. On the contrary, SAGE remains pedagogically valuable because it teaches students to work with AI in a critical and evidence-based manner through generation, evaluation, refinement, critique, and reflection. The difficulty identified here is not that this pedagogical structure lacks value, but that pedagogical artefacts generated under unsupervised conditions do not automatically become credible assurance evidence merely because they are documented.

This distinction is important. Policy can define whether AI use is permitted, but policy alone does not teach students how to interrogate AI output, nor does it show educators how to make student reasoning visible. Structured pedagogies such as SAGE address that gap by building evaluative

judgement into the learning process. However, the present evidence suggests that pedagogy must be carried one step further. The development of AI-supported work should remain visible and scaffolded, but assurance should not be assumed from logged process alone. Instead, pedagogy should provide the foundation on which more credible forms of demonstrated ownership can later be built.

### 6.3. Recommendation 3: Embed assurance through defensible judgement rather than defaulting to low-tech exams

The results of this study do not support a blanket return to traditional examinations as the primary response to GenAI. Such a move would risk reviving exactly the forms of assessment reform has long sought to move beyond, privileging recall under pressure over explanation, judgement, revision, and authentic application. Exam reversion also reintroduces known equity costs. Closed-book examination formats systematically disadvantage students with anxiety, language barriers, and diverse learning needs, and a generation of assessment reform has been directed at reducing rather than restoring those barriers [29]. More fundamentally, the current institutional move back toward invigilated formats is being driven by anxiety about AI rather than by an updated theory of how students learn or what graduates are now expected to demonstrate. The challenge is not how to recreate a pre-AI classroom, but how to verify whether students can exercise responsibility, ownership, and disciplinary judgement in an AI-rich environment. Graduates will enter professional contexts in which AI use is expected, interrogated, and accountable. The competency at stake in those contexts is not the production of polished text, which AI can supply, but evaluative judgement: the capacity to identify where AI is wrong, to test its claims against authoritative sources, to revise its output, and to defend those revisions under appropriate scrutiny [30]. This is what assurance should now be designed to verify.

For this reason, assurance should be embedded through moments of defensible judgement rather than being outsourced either to the final submitted artefact or to a blanket revival of closed-book exams. In direct response to the limitations documented in this study, a sixth step has been added to the SAGE framework: Defend, a short supervised checkpoint at which students demonstrate ownership of the reasoning embedded in their AI-supported work under conditions that cannot be simulated or delegated. Defend is not a full examination, nor is it a parallel secure task running alongside the open work. It is integrated into the assessment sequence itself, immediately following the open developmental cycle (Generate, Evaluate, Refine, AI Critic, Reflect). Its format varies by discipline and context, and may take the form of a brief oral defence, a live walkthrough, a timed scenario-based exercise, a code explanation, a practical demonstration, or other format that makes reasoning and ownership observable. The principle is not that one format is universally superior. The principle is that assurance should arise through a fit-for-purpose moment of demonstrated understanding that is embedded within the AI-integrated work rather than appended to it as a separate stream. Defend also addresses the marker's dilemma documented in Section 5.3. By embedding a supervised assurance moment within the assessment sequence, the verification work that would otherwise fall to the individual marker as post hoc audit is shifted into a structured, time-bounded interaction with the student. The educator is no longer required to interpret whether a reflection is genuine or whether an evaluation cell reflects compliance. Those questions are answered directly through the student's demonstrated reasoning at the Defend checkpoint. This redistributes the assurance work from interpretive marking, which is cognitively demanding and structurally inconclusive, to direct evaluation, which is the activity educators are trained for and structurally suited to perform.

### 6.4. Recommendation 4: Design integrated learning-and-assurance sequences across units and programs

The broader implication of the present study is that assurance should not be improvised as a late-stage control added after learning has taken place. It should be designed as part of an integrated sequence in which students first engage openly with AI in developmental tasks and are then required, at structured points, to demonstrate ownership of the reasoning embedded in those outputs. Under such a model, learning and assurance are connected through a single progression in which open work feeds forward directly into later acts of explanation, justification, and defence, rather than being relegated to separate streams or compressed into a single high-stakes terminal artefact.

This conception is related to, but conceptually distinct from, the program-level two-lane approach in which open and secure assessment tasks are allocated to parallel streams across the curriculum [28]. Both positions share the principle that not every task should carry the same assurance burden, and both reject the assumption that a single open artefact can serve as conclusive proof of learning. The difference lies in the locus of assurance. Under the two-lane approach, assurance is allocated to a separate, parallel task. Under the model supported by the present findings, assurance is embedded as a step within the AI-integrated work itself, so that open developmental engagement and supervised demonstration of ownership belong to the same assessment sequence rather than to different ones. Embedding assurance within learning, rather than running it alongside, preserves the pedagogical continuity that structured AI integration is intended to provide while closing the verification gap documented in the present audit.

The SAGE framework, with the addition of the Defend step, operationalises this principle at the level of assessment design rather than at the level of policy aspiration. The open developmental phase, in which students engage with AI under structured scaffolding, accept-modify-reject decision logs, and a critique cycle, is followed by a supervised checkpoint at which ownership is demonstrated under conditions the educator can directly evaluate. Assurance is therefore neither an interpretive task superimposed on marking nor a parallel stream of secure work running alongside the open task. It is a structural feature of the assessment itself, performed by the design rather than by the individual marker. The marker's role at the Defend checkpoint is the evaluation of demonstrated reasoning at a defined and time-bounded point in the sequence, which is the activity educators are trained and structurally suited to perform.

At the program level, this principle scales without requiring uniformity. Not every unit needs to carry the same assurance burden or use the same Defend format. More credible assurance emerges when programs are designed so that students encounter multiple, context-appropriate opportunities to demonstrate evaluative judgement under supervised conditions, with the Defend format calibrated to the cognitive demand of the preceding task. Under this arrangement, assurance becomes a deliberate, programmatic property of the curriculum rather than an ad hoc response improvised by individual unit coordinators in reaction to a perceived integrity risk. The submitted artefact ceases to bear the burden of standalone proof, and defended understanding becomes the credible marker of attainment in AI-integrated higher education.

## 7. Limitations

Several limitations must be acknowledged. First, this study is based on two small cohorts at a single institution with a dataset of 25 group submissions (approximately 100 students). The audit findings reported here are cohort-specific and should be generalised with caution.

Second, the analysis relies entirely on document artefacts and internal consistency checks. It cannot observe drafting behaviour. It cannot determine the sequence in which document components were produced, and cannot establish intent. The audit identifies evidence gaps and inconsistencies, but cannot determine why those gaps exist. Supplementary methods such as student interviews, think-aloud protocols, or screen recordings would be required to understand the behavioural mechanisms underlying the observed patterns.

Third, group work complicates attribution. Each submission was produced by a group of four students, and it is not possible to attribute specific process fidelity decisions to individual group members from the submitted documents. A submission that appears inconsistent may reflect the work of one group member who deviated from the group's process rather than a collective decision.

Fourth, the marker was also the designer of the assessment and the lead developer of the SAGE framework. This dual role introduces a potential interpretation bias: the marker may have applied more rigorous scrutiny to process evidence than an independent marker would, or conversely, may have been inclined to interpret ambiguities favourably. Future studies should include independent coding to mitigate this concern.

Fifth, the study cannot determine whether documented process fidelity predicts individual competence. The three submissions that demonstrated a complete audit chain may or may not reflect higher individual capability among their members. A secure performance measure, administered independently, would be required to triangulate whether students who document their process thoroughly also perform better under supervised conditions.

Despite these limitations, the failure modes identified are structural rather than incidental. They arise from the interaction between unsupervised conditions, capable AI tools, and self-reported process evidence. Researchers seeking to further validate these findings are encouraged to replicate the SAGE audit protocol with independent coding, include student interviews on workload and decision-making, and incorporate a secure performance measure to examine the relationship between process documentation quality and demonstrated individual competence.

## 8. Conclusions

The evidence presented in this study does not constitute an argument against take-home assessment as a pedagogical mode. The five-check SAGE audit protocol, applied across 25 group submissions from undergraduate and postgraduate cybersecurity cohorts, demonstrates something more precise: that unsupervised take-home artefacts, in their conventional form, can no longer function as sufficient standalone instruments of learning assurance under GenAI conditions. The compliance gradient identified across the cohort reveals not a failure of student capability, but a structural misalignment between what take-home assessment was designed to produce and what assurance now requires. Within a structured pedagogical architecture such as SAGE, take-home assessment retains substantial learning value. The iterative, AI-engaged process it affords — when students are guided through disciplined

cycles of application, reflection, and critique — continues to develop the higher-order competencies that tertiary education is obligated to cultivate. What it can no longer do, without further design intervention, is serve as the terminal evidence of that development. The submission artefact and the assurance of learning have been decoupled by the conditions this study documents. The design challenge that follows from this decoupling is not a retreat to supervised examination as a default, but the deliberate embedding of defended, observable judgement within AI-integrated assessment sequences. Students must be positioned to account for their reasoning in contexts that cannot be simulated or delegated. The Defend step now embedded as a sixth step within the SAGE framework is the direct response to the assurance gap this study documents. The present paper establishes the empirical foundation that motivated that response: the demonstration that, under unsupervised conditions, process documentation alone cannot bear the assurance burden assigned to it.

## Author contributions

Mahmoud Elkhodr: Conceptualisation, methodology, assessment design, data collection, SAGE audit protocol development, formal analysis, writing – original draft, writing – review and editing, visualisation. Ergun Gide: Supervision, writing – review and editing, validation. Both authors have read and approved the final version of the manuscript and take full responsibility for its content.

## Use of generative AI tools declaration

The authors declare that GenAI tools (ChatGPT by OpenAI and Prism) were used during the preparation of this manuscript for the following purposes: improving the readability and language of selected passages, formatting references, LaTeX formatting, and assisting with the structural organisation of draft sections. All AI-generated or AI-assisted content was critically reviewed, and verified by the authors. The authors retain full responsibility for the accuracy, integrity, and originality of the work. No AI tool was used in the data collection, analysis, or interpretation of findings.

## Acknowledgments

The authors wish to thank the students who participated in the cybersecurity management units at Central Queensland University during the third teaching term of 2025, whose submitted work formed the dataset for this study.

## Conflict of interest

The authors declare no financial conflict of interest. The lead author is the developer of the SAGE framework evaluated in this study; this dual role is disclosed here and discussed as a potential interpretation bias in Section 7.

Professor Ergun Gide is an editorial board member of STEM Education and was not involved in the editorial review or the decision to publish this article.

## Ethics declaration

This study involved a retrospective audit of de-identified group assessment submissions collected as part of standard coursework requirements. It was conducted in accordance with CQUniversity's Human Research Ethics Procedure, Reference 3127, which recognises exemption from ethical review for lower-risk research satisfying the relevant National Statement conditions. No participants were interviewed, surveyed, or observed, and no individually identifiable information is reported. The relevant procedure is available through CQUniversity's public policy register under Human Research Ethics Procedure, Reference 3127.

## Data availability

The audit dataset consists of de-identified student assessment submissions and cannot be made publicly available due to student privacy obligations under Australian higher education regulations. The five-check audit protocol is described in full in Section 3 and is designed to be independently replicable. Requests for aggregated, de-identified data may be directed to the corresponding author.

## References

1. M. Elkhodr and E. Gide, AI leads, humans lead, or collaborate? empirical findings and the SAGE roadmap for embedding GenAI in systems analysis and design education, *STEM Education*, **6** (2026), 194–229.
2. S. Rafiq, Qurat-ul-Ain and A. Afzal, The role of AI detection tools in upholding academic integrity: An evaluation of their effectiveness, *Contemporary Journal of Social Science Review*, **3** (2025), 901–915. https://contemporaryjournal.com/index.php/14/article/view/379.
3. J. Fleckenstein, J. Meyer, T. Jansen, O. Köller, S. D. Keller and J. Möller, Do teachers spot AI? evaluating the detectability of AI-generated texts among student essays, *Computers and Education: Artificial Intelligence*, **6** (2024), 100209.
4. J. Luo, A critical review of GenAI policies in higher education assessment: a call to reconsider the "originality" of students' work, *Assessment & Evaluation in Higher Education*, **49** (2024), 651–664.
5. Y. An, J. H. Yu and S. James, Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration, *International Journal of Educational Technology in Higher Education*, **22** (2025), 10.
6. Y. Jin, L. Yan, V. Echeverria, D. Gašević and R. Martinez-Maldonado, Generative AI in higher education: A global perspective of institutional adoption policies and guidelines, *Computers and Education: Artificial Intelligence*, **8** (2025), 100348.
7. Y. Dai, S. Lai, C. P. Lim and A. Liu, University policies on generative AI in Asia: Promising practices, gaps, and future directions, *Journal of Asian Public Policy*, **18** (2025), 260–281.
8. C. K. Y. Chan, A comprehensive AI policy education framework for university teaching and learning, *International Journal of Educational Technology in Higher Education*, **20** (2023), 38.

9. M. Perkins, L. Furze, J. Roe and J. MacVaugh, The artificial intelligence assessment scale (AIAS): A framework for ethical integration of generative AI in educational assessment, *Journal of University Teaching and Learning Practice*, **21** (2024), q3azde36.
10. Z. Quince, J. Munn and R. Greenaway, *Adapting assessment in the age of generative AI: The AAM-GenAI framework (practice report)*, Scholarship of Learning and Teaching Paper 28, Southern Cross University, 2025.
11. M. Elkhodr, E. Gide, R. Wu and O. Darwish, ICT students' perceptions towards ChatGPT: An experimental reflective lab analysis, *STEM Education*, **3** (2023), 70–88.
12. M. Elkhodr and E. Gide, The SAGE framework for developing critical thinking and responsible generative AI use in cybersecurity education, *Discover Education*, **4** (2025), 517.
13. M. Elkhodr and E. Gide, Embedding generative AI into systems analysis and design curriculum: Framework, case study, and cross-campus empirical evidence, arXiv preprint arXiv:2511.17515, 2025. https://arxiv.org/abs/2511.17515.
14. M. Elkhodr and E. Gide, AI as critic: Validating SAGE pedagogy for human authority and responsible GenAI use in systems analysis and design education, EdArXiv Preprints, 2025. https://osf.io/preprints/edarxiv/8j3xf.
15. M. Elkhodr, A. Azra and E. Gide, How first-year students actually use ChatGPT in permitted assessments: Empirical typologies, verification gaps, and the policy-practice divide, Research Square Preprints, 2026.
16. H. Ranasinghe, E. Gide and M. Elkhodr, The significance of GenAI empowered ERP systems course teaching in quality education, in *2024 21st International Conference on Information Technology Based Higher Education and Training (ITHET)*, 2024, 1–7.
17. R. Sandu, E. Gide and M. Elkhodr, The role and impact of ChatGPT in educational practices: insights from an Australian higher education case study, *Discover Education*, **3** (2024), 71.
18. T. Corbin, P. Dawson and D. Liu, Talk is cheap: why structural assessment changes are needed for a time of GenAI, *Assessment & Evaluation in Higher Education*, **50** (2025), 1087–1097.
19. S. Leaton Gray, D. Edsall and D. Parapadakis, AI-based digital cheating at university, and the case for new ethical pedagogies, *Journal of Academic Ethics*, **23** (2025), 2069–2086.
20. B. L. Moorhouse, M. A. Yeo and Y. Wan, Generative AI tools and assessment: Guidelines of the world's top-ranking universities, *Computers and Education Open*, **5** (2023), 100151.
21. H. Tomisu, J. Ueda and T. Yamanaka, The cognitive mirror: A framework for AI-powered metacognition and self-regulated learning, *Frontiers in Education*, **10** (2025), 1697554.
22. S. He and Y. Cui, A systematic review of the use of log-based process data in computer-based assessments, *Computers & Education*, **228** (2025), 105245.
23. M. Elkhodr and E. Gide, Embedding generative AI in curriculum: the SAGE framework and evidence-based implementation guide, 2026. https://doi.org/10.5281/zenodo.18383951.
24. M. Elkhodr and E. Gide, SAGE framework: structured AI-guided education, https://sage-framework.com, 2026, Accessed: 19 May 2026.

25. Australian Cyber Security Centre, *Essential Eight Maturity Model*, Australian Cyber Security Centre, 2023. https://www.cyber.gov.au/sites/default/files/2023-11/PROTECT%20-%20Essential%20Eight%20Maturity%20Model%20%28November%202023%29.pdf.

26. D.-W. Kim, J.-Y. Choi and K.-H. Han, Risk management-based security evaluation model for telemedicine systems, *BMC Medical Informatics and Decision Making*, **20** (2020), 106.

27. J. M. Lodge, S. Howard, M. Bearman, P. Dawson and Associates, *Assessment reform for the age of artificial intelligence*, Technical report, Tertiary Education Quality and Standards Agency, 2023. https://www.teqsa.gov.au/sites/default/files/2023-09/assessment-reform-age-artificial-intelligence-discussion-paper.pdf.

28. A. Bridgeman, D. Liu and R. Weeks, Program level assessment design and the two-lane approach, Teaching@Sydney, The University of Sydney, 2024. https://educational-innovation.sydney.edu.au/teaching%40sydney/program-level-assessment-two-lane/.

29. M. Elkhodr, AI era must not become excuse to default to low-tech exams, Future Campus, 2026. https://futurecampus.com.au/2026/04/11/ai-era-must-not-become-excuse-to-default-to-low-tech-exams/.

30. M. Elkhodr and E. Gide, Students are asking for AI guidance, not just policy, Times Higher Education Campus, 2026. https://www.timeshighereducation.com/campus/students-are-asking-ai-guidance-not-just-policy.

**Author's biography**

Dr. Mahmoud Elkhodr is a Senior Lecturer in the School of Engineering and Technology at Central Queensland University, Australia. His research focuses on cybersecurity, the Internet of Things, artificial intelligence, digital health, smart cities, and technology-enhanced education.

Dr. Ergun Gide is a Professor in the School of Engineering and Technology at Central Queensland University, Australia. His research and teaching interests include AI, project management, digital transformation, higher education, and technology-enhanced learning.