STEM Education

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

STEM Education, 6(5): 974–997
DOI:10.3934/steme.2026039
Received: 10 March 2026
Revised: 30 May 2026
Accepted: 01 July 2026
Published: 08 September 2026

Perspective

Structured AI-guided education (SAGE): A six-step pedagogy for AI orchestration and assessment assurance in higher education

Mahmoud Elkhodr*and Ergun Gide

School of Engineering and Technology, Central Queensland University, Sydney, Australia

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

Academic Editor: Stephen Xu

Abstract: Generative artificial intelligence (GenAI) has altered the conditions under which higher education assessment is expected to evidence student capability. Detection-based responses are unreliable, while prohibition alone does not prepare graduates for AI-integrated workplaces. This perspective presents Structured AI-Guided Education (SAGE), an empirically developed and iteratively evaluated six-step pedagogy for AI orchestration and assessment assurance. SAGE requires students to begin with a human baseline, evaluate AI output against authoritative disciplinary anchors, document accept/modify/reject decisions, refine the work, use AI as a bounded critic, reflect on the transferability of their judgment, and defend the reasoning behind the final artifact. Across the authors' reported program, SAGE has been directly implemented with more than 1500 students in undergraduate, postgraduate, and transnational contexts. The framework distinguishes AI permission from four observable task-level orchestration capabilities: Passive Acceptor, Selective Adapter, Balanced Integrator, and Critical Synthesiser. These profiles describe the quality of human judgment demonstrated rather than fixed learner identities. Defend is positioned as a proportionate, task-matched confirmation point within a distributed assurance architecture rather than as a single terminal examination or oral-only requirement. The paper provides the canonical account of the six-step cycle, its terminology, staged progression, required learner artifacts, comparative positioning, empirical evidence base, cross-disciplinary Defend formats, and evaluation blueprint. SAGE thereby connects institutional permission settings with disciplined AI use, visible student judgment, and proportionate assurance of individual attainment.

Keywords: generative AI, AI orchestration, assessment assurance, assessment design, academic integrity, AI literacy, higher education, SAGE

1. Introduction

Generative artificial intelligence (GenAI) tools can produce coherent text, code, and analytical outputs at low cost and high speed, creating new opportunities for learning support alongside new threats to assessment validity [9, 10]. When GenAI is permitted without an explicit learning workflow, assessment artifacts may cease to reliably represent a student's reasoning, judgment, and problem-solving process. The difficulty observed in the teaching contexts from which SAGE emerged was more specific than the now familiar language of over-reliance and reduced critical thinking. Students were still learning the university habit of questioning authority, qualifying claims, and reasoning against evidence, yet AI produced polished and apparently authoritative responses within seconds. This was visible in first-year transition, where students move from school into university, and in some linguistically and culturally diverse cohorts where previous educational experience had emphasized receiving authoritative answers more than challenging them. SAGE was therefore designed to teach students how to question an apparently authoritative answer, not simply to warn them that AI may be wrong. This diagnosis, that the difficulty is dispositional and disciplinary rather than a matter of tool access alone, is the insight from which the framework is designed.

Educational responses to GenAI have typically been positioned as a binary choice between restriction and permissiveness [1, 2, 9]. A third direction has become increasingly necessary: structured orchestration, in which students are required to interact with GenAI in ways that make their reasoning explicit, testable, and teachable [3, 8]. The Structured AI-Guided Education (SAGE) framework was developed for this purpose. Its design begins from a simple pedagogical rule: AI should receive the student's thinking, not replace the starting point of that thinking. Students are therefore asked to bring an initial draft, outline, bullet points, sketch, interpretation of the task, or statement of uncertainty before AI is used. This initial human baseline turns AI from a substitute author into a comparator, coach, and critic. SAGE then teaches students to evaluate AI output against disciplinary anchors and to justify every major accept, modify, or reject decision. It provides a practical, adoptable pedagogy that can be implemented within existing courses without requiring redesign of the entire curriculum.

Australia's national regulator, the Tertiary Education Quality and Standards Agency (TEQSA), has urged universities to rethink assessment rather than rely on detection [9, 10]. In its 2025 guidance, TEQSA called for a balance between "open" tasks, where students use AI to build capability, and "secure" tasks, where they demonstrate individual attainment under supervised conditions. The University of Sydney similarly introduced a two-lane assessment model separating secure and open tasks [12]. However, policy frameworks alone do not show educators how to teach with AI in a way that develops genuine disciplinary judgment. SAGE addresses this pedagogical gap, but it is not itself an assessment-security classification or an AI-permission scale. These are distinct but connected design decisions. Assessment security determines how strongly individual attainment must be independently assured; AI permission determines what forms of GenAI use are allowed within a task; a SAGE orchestration level identifies the quality of human judgment that the learning design is intended to develop and the rubric may assess; and the assessment design specifies the process artifacts, controlled conditions, and direct-assurance activities through which that capability is demonstrated. Permission scales such as the AI Assessment Scale specify how much AI use is permitted, but do not prescribe how students should exercise and evidence judgment during that use [23]. SAGE provides this pedagogical and evidentiary layer by structuring how AI is used and, through Defend, confirming ownership of the

relevant reasoning where an assured claim about individual attainment is made [18, 19].

1.1. Positioning SAGE among adjacent approaches

SAGE occupies a complementary position among several established responses to GenAI in higher education. Permission frameworks govern what use of AI is allowed; program-level assurance approaches govern where capability-building and secure evidence of attainment should occur; AI-literacy models describe the knowledge and dispositions students require; and AI-resistant or secure assessment designs seek to constrain inappropriate substitution. SAGE addresses a different but connected question: how students should exercise, document, and demonstrate disciplinary judgment when AI participates in learning and assessment. Table 1 summarizes this positioning.

Table 1. Comparative positioning of SAGE among adjacent approaches to AI-integrated assessment.

Adjacent approach	Primary question governed	Complementary contribution of SAGE
AI permission frameworks, including the AI Assessment Scale [23]	What forms or degree of AI use are permitted within an assessment?	Specifies how students exercise, document, and defend disciplinary judgment within permitted AI use through a repeatable learning process and required evidence artifacts.
Open–secure, two-lane, and program-level assurance approaches [9, 10, 12]	Where and under what conditions should AI-integrated capability development and evidence of individual attainment occur?	Provides a hybrid, learning-outcome-mapped pedagogy rather than a separate two-lane structure. AI-integrated capability development and task-matched direct assurance may coexist within or across assessments, with intended learning outcomes determining what must be defended.
AI-literacy and critical AI-literacy models [24, 25]	What knowledge, competencies, and critical dispositions should learners develop when engaging with AI?	Operationalizes evaluative and metacognitive capability through six ordered steps, required learner artifacts, accept/modify/reject decisions, and observable orchestration profiles.
AI-resistant or secure assessment design [9, 10]	How can unauthorized substitution, inappropriate assistance, or unverified learning claims be constrained?	Provides a capability-building design for contexts in which AI use is permitted, professionally relevant, and pedagogically valuable, while retaining direct assurance of individual reasoning.

SAGE does not replace permission scales, critical AI-literacy models, or program-level assurance architectures. It supplies the pedagogical and evidentiary process connecting those policy and curriculum decisions to observable student judgment.

This paper presents the SAGE framework as an adoptable pedagogy for educators seeking to integrate GenAI into teaching and assessment in ways that preserve student judgment rather than merely regulate tool access. The paper specifies the six-step cycle, the two-stage progression, required learner artifacts, discipline-specific implementation patterns, and an assurance-by-design model that distributes evidence of learning across the process [19]. A summary of the empirical evidence base is included to support adoption decisions. The presentation remains implementation-centered and is intended for direct use by unit coordinators, course designers, academic integrity leads, and educators designing assessment in AI-integrated learning environments.

SAGE is best understood as a program rather than a single technique. As shown in Figure 1, it comprises a six-step pedagogy, a permission-to-pedagogy alignment layer, and a research-integrity extension, supported by a suite of student-facing and educator-facing resources. The pedagogy, which

is the subject of this paper, structures how students generate, evaluate, refine, critique, reflect upon, and defend AI-assisted work.

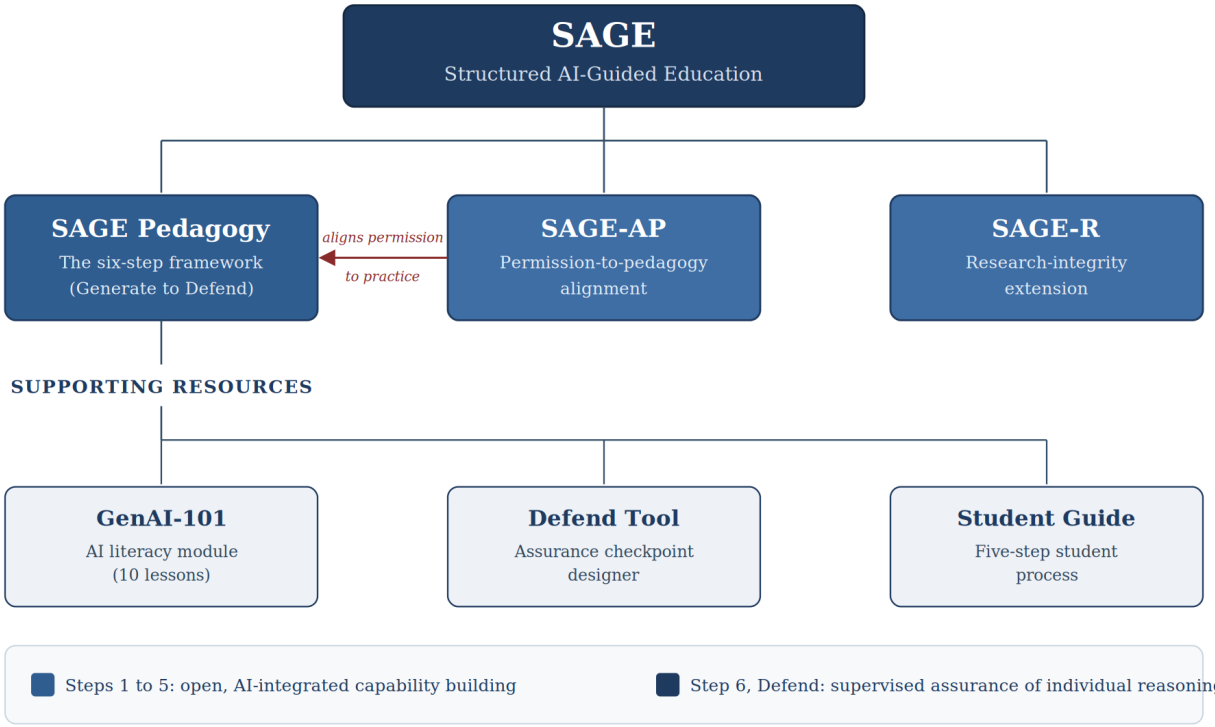

Figure 1. The SAGE program. The six-step pedagogy, the SAGE-AP permission-to-pedagogy alignment layer, and the SAGE-R research-integrity extension sit within a single program, supported by the GenAI-101 module, the Defend Tool, and the Student Guide.

The Structured AI-Guided Education Assessment Policy (SAGE-AP) translates institutional permission settings into explicit conditions for AI-supported assessment, thereby connecting policy to practice [18]. The SAGE Research extension (SAGE-R) applies the same principles to the supervision of research candidature [22]. The present paper provides the canonical account of the pedagogy and its assurance model, while the associated policy and assurance treatments are developed in the companion publications [18, 19].

SAGE is theoretically grounded in several established traditions rather than in Bloom's revised taxonomy alone. The cycle operationalizes self-regulated learning by requiring students to set evaluation criteria, monitor AI output against those criteria, and adjust their work through documented revision, mirroring the forethought, performance, and self-reflection phases described by Zimmerman [15]. The Evaluate, AI Critic, and Reflect steps draw on metacognition and reflective practice, requiring students to externalize and interrogate their reasoning rather than merely describe a workflow [16, 17]. Constructive alignment and programmatic assessment further inform the assurance model: the evidence students produce during the open learning process must align with the intended learning outcomes, and the strongest assurance should occur at points where the relevant reasoning can be directly observed. Bloom's taxonomy is retained only as a coarse sequencing heuristic for cognitive

demand across the steps. Against this background, the contribution of SAGE is integrative rather than the invention of a single technique. Iterative refinement, structured reflection, adversarial critique, and defense each exist in the prior literature; what is distinctive is their integration into a repeatable assessment sequence that couples open, AI-integrated capability-building with distributed, task-matched assurance of student judgment.

2. The SAGE framework

SAGE is an empirically developed and iteratively evaluated pedagogy for developing students' capacity to orchestrate generative AI outputs rather than passively accept them. AI orchestration is defined here as the deliberate human direction, evaluation, correction, critique, synthesis, and defence of AI-generated or AI-assisted outputs against domain evidence, such that the human retains decision authority over what is accepted, modified, or rejected and can account for each decision. The framework is built around a discipline-learning claim: novices cannot usually see AI errors in the way an expert can. An expert may recognize a faulty HTML pattern, a missing NIST control, a weak risk statement, or an inappropriate policy recommendation almost immediately; students must be taught the standards and reference points that make such recognition possible. SAGE therefore uses anchors, including rubrics, professional standards, laws, policies, source texts, datasets, code tests, case constraints, and disciplinary principles, as the mechanism through which students build the knowledge required to judge AI output. SAGE treats GenAI as an assistive tool whose outputs must be accepted, modified, or rejected with explicit, evidence-based justification [8]. The framework originated in classroom studies that began in 2022 and were first reported in 2023 [1], and was subsequently developed through classroom-based studies in ICT, cybersecurity, data analytics, systems analysis and design, and professional communications [2–7]. It has since expanded into a broader evidence base involving more than 1,500 students, first-year and postgraduate cohorts, external Australian implementation, and transnational implementation [6, 18–21]. SAGE therefore functions as a transferable pedagogy for building visible, accountable AI judgment across contexts, rather than as a single assessment technique.

2.1. Canonical terminology

Table 2 consolidates the principal terms used throughout the SAGE program. The definitions state how each term is used within SAGE. Where a concept is developed more fully in a companion paper, that source remains the citation of record for the underlying argument.

Table 2. Canonical terminology used within the SAGE program.

Term	Canonical definition within SAGE
AI orchestration	The deliberate human direction, evaluation, correction, critique, synthesis, and defense of AI-generated or AI-assisted output against disciplinary evidence, with the human retaining decision authority over what is accepted, modified, or rejected.
Human baseline	A representation of the student's own thinking produced before substantive AI assistance, such as notes, an outline, interpretation, sketch, preliminary analysis, partial solution, or statement of uncertainty. Its purpose is to prevent AI from becoming the first thinker in the task and to make the student's starting position visible.
Disciplinary anchor	An authoritative external reference point used to evaluate an AI output or decision, including a professional standard, law, regulation, peer-reviewed source, clinical guideline, dataset, code test, source text, rubric, policy, or case constraint.
Corroborative evidence	Process evidence that supports an interpretation of student engagement but does not, by itself, establish individual attainment. Examples include prompts, AI transcripts, evaluation logs, change records, accept/modify/reject decisions, and reflections [19].
Controlled assurance condition	A supervised, time-bounded, restricted, or otherwise controlled environment that reduces opportunities for substitution or external assistance, although it may not directly reveal the student's underlying reasoning [19].
Direct assurance	Task-matched evidence in which the student personally demonstrates, explains, adapts, reproduces, or applies the relevant reasoning under conditions that permit an inference of individual attainment [19].
Distributed assurance	An assessment architecture in which corroborative process evidence, controlled assurance conditions, and direct assurance are distributed across the learning sequence rather than concentrated in one terminal event [19].
Assurance debt	The accumulation of claims about student learning or attainment for which sufficiently direct evidence has not yet been obtained, increasing the risk that assurance is postponed to an inadequate or excessively high-stakes endpoint [19].

As shown in Figure 2, the core of SAGE is a six-step learning process that moves students from passive AI consumption to critical AI orchestration. The sequence is deliberately ordered, following six steps: Generate, Evaluate, Refine, AI Critic, Reflect, and Defend. Generate protects the student's starting point by requiring human contribution before or within the prompt; Evaluate anchors the AI output against disciplinary authority; Refine converts that comparison into documented decisions; AI Critic teaches students to seek challenge without surrendering judgment; Reflect consolidates the anchoring habit as a transferable professional skill; and Defend confirms that the reasoning can be explained, adapted, and justified by the student. The process does not simply culminate in Defend as a final examination. Instead, evidence of judgment is generated across the preceding steps: students prompt, compare, correct, critique, document, and reflect before they are asked to confirm ownership of the reasoning. Defend is therefore a proportionate, task-matched confirmation point within a distributed assurance layer, not the sole source of assurance.

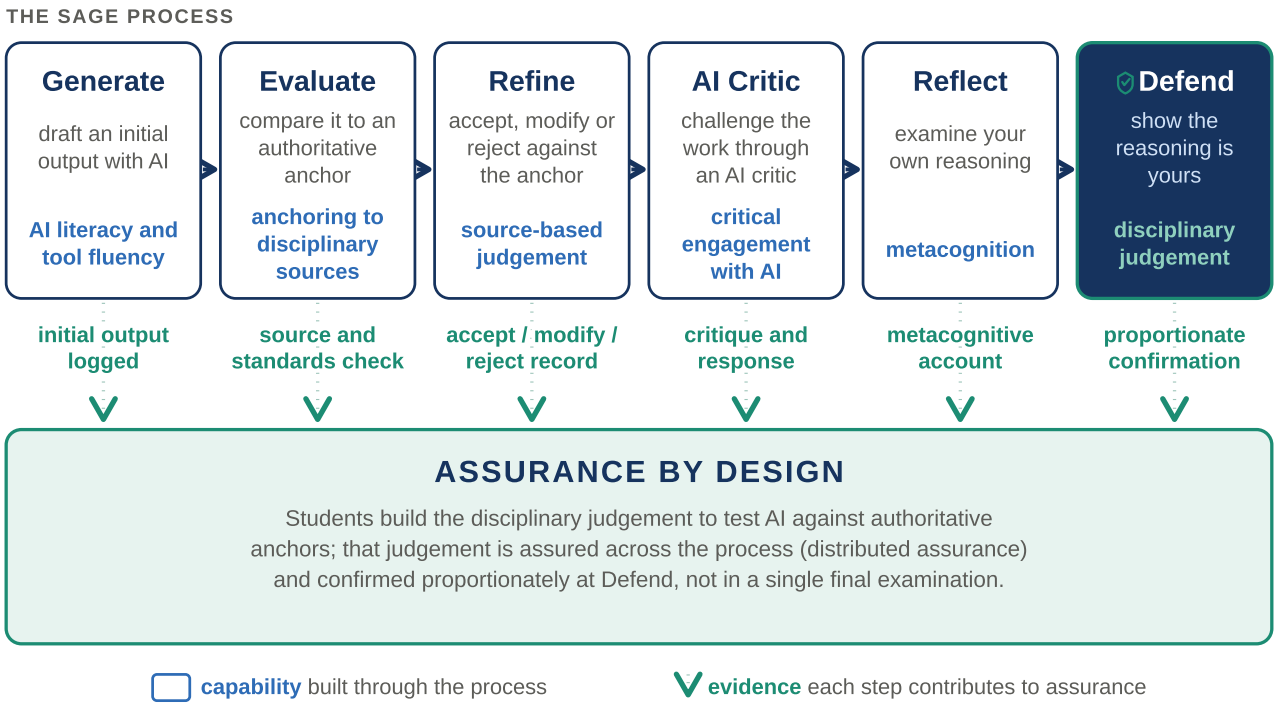

Figure 2. The SAGE six-step process within a distributed assurance architecture. The horizontal representation depicts a hybrid, learning-outcome-mapped design: evidence of judgment is generated throughout the AI-integrated sequence, while Defend provides task-matched direct confirmation for the intended learning outcomes that require assurance rather than functioning as a separate lane or a single final examination.

2.2. The six-step cycle

Step 1: Generate. The first act in SAGE is not delegation to AI but the preservation of the student's own starting point. In instructor-scaffolded tasks, this may occur through a carefully designed base prompt that asks AI to produce a deliberately useful but imperfect draft for critique. In more independent tasks, students must produce a human baseline, such as an outline, notes, bullet points, sketch, partial solution, problem interpretation, or statement of uncertainty, before using the AI tool. The baseline need not be polished; its function is to make the student's prior understanding visible and to prevent the tool from becoming the first thinker in the task. In the early stages, standardized prompts are provided by the instructor to ensure comparable outputs across the cohort. These prompts are derived from the task's intended learning outcomes and from the domain criteria against which the output will later be evaluated, so that each prompt elicits an output rich enough to support a meaningful evaluation in Step 2. A small prompt bank developed by the teaching team is reused across the cohort to control variability. In later stages, students design their own prompts, but their prompt must incorporate the human baseline and any standards, readings, or task constraints supplied by the educator. This baseline demonstrates independent engagement with the problem and prevents complete delegation.

Step 2: Evaluate. Students compare the AI output against authoritative anchors relevant to their discipline. An anchor is an external reference point that gives the student a basis for judgment: an

industry standard, regulatory framework, professional code, peer-reviewed paper, clinical guideline, rubric, textbook, dataset, code test, law, policy, or case-study constraint. This step is the core disciplinary skill-building mechanism in SAGE. Experts can often identify an error because they already know the standard; students need a structured way to learn what the standard requires and then to hold AI output accountable against it. In the formative stage, anchors are supplied directly by the instructor as a curated resource set, accompanied by an evaluation checklist and worked exemplars that show what an adequate comparison looks like; in the summative stage, students locate and justify their own anchors. They identify what aligns, what is missing, what is unsupported, what is oversimplified, and what is incorrect. This step is where critical thinking becomes practical: students do not merely ask whether an AI answer is good; they ask what it is good against.

Step 3: Refine. Students modify the AI output on the basis of their evaluation. Each modification is documented as an accept, modify, or reject decision with a brief evidence-based justification citing specific anchors, standards, sources, or contextual constraints. Accept means that the AI output aligns with the anchor and can be explained; modify means that the output is partly useful but requires correction or contextualization; reject means that the output is wrong, misleading, unsupported, or inappropriate for the task. If the AI overlooked jurisdictional requirements, ignored cultural safety considerations (as defined by the relevant disciplinary or professional standard, for example, the applicable code of practice or regulatory guideline, rather than by the student or instructor alone), or fabricated references, the student must identify and correct each issue. The refinement log is therefore not a supplementary compliance document; it is the primary evidence that the student has transformed AI output through disciplinary judgment. The refinement produces a professional-standard artifact that the student can defend as their own intellectual product.

Step 4: AI Critic. Students assign the AI tool a domain-specific persona (for example, "You are a senior cybersecurity auditor reviewing this policy for regulatory compliance") and submit their refined output and decision log for critical evaluation. They then analyze the AI's feedback, determining which critiques are valid, which reflect the AI's own limitations, and which require human judgment to resolve. The AI Critic step is intentionally bounded: the AI is used as a critical coach, not as a second source of truth. It asks difficult questions and surfaces possible weaknesses, but the student must still adjudicate the critique against the anchor. This step therefore builds evaluative agency and intellectual confidence. Students learn not only to receive critique, but also to reject AI suggestions when they are generic, misaligned, or unsupported. Conceptually, the AI Critic step combines adversarial prompting with metacognitive scaffolding rather than simple feedback generation. The persona instruction is an adversarial device that elicits challenge to the student's work; the subsequent appraisal of which challenges to accept is the metacognitive component, in which the AI critique functions as an external prompt for the student to reconsider and defend their own reasoning. The pedagogical object of interest is the student's discrimination among critiques, not the critique itself. This step inverts the AI relationship: the student must evaluate the evaluator, operating at Bloom's Evaluate level [11]. It develops metacognitive sophistication and strengthens the student's authority over the AI.

Step 5: Reflect. Students produce a metacognitive analysis documenting what the AI contributed, where it failed, the corrective reasoning applied, the ethical and disciplinary considerations that informed their decisions, and the broader implications for AI reliability in their discipline. Functional reflections, namely those that critique domain content and consequential decisions directly, are

distinguished from procedural reflections that merely describe workflow benefits. The reflection has two purposes. First, it records the student's intellectual contribution by linking the final artifact to the accept/modify/reject and AI Critic decisions that shaped it, including any relevant ethical, professional, or contextual judgment. Second, it transfers the anchoring habit beyond the immediate task: the anchor used in one assessment may later become a professional code, organizational policy, dataset, design standard, legal requirement, or ethical obligation in the workplace. Over time, students begin to anticipate AI weaknesses, risks, and contextual limitations before they occur, which constitutes an essential professional skill in AI-integrated workplaces.

Step 6: Defend. Students demonstrate, under supervised or otherwise controlled conditions, that they can explain, justify, critique, reproduce, adapt, extend, or apply the reasoning documented in their AI-assisted work. The intended learning outcomes determine what must be directly confirmed, while the epistemic form of the assessed competency determines the appropriate form of Defend. The defining criterion is ownership of judgment: students must be able to identify risks, explain trade-offs, justify decisions, critique limitations in their own final work, and respond to task-matched challenges or altered conditions. Full SAGE assessment adoption includes Defend, but this does not require every formative exercise or every component of an assessment to conclude with a separate supervised event. Within the hybrid architecture, Defend may be embedded within the same assessment as the AI-integrated work, linked to a related assessment, or distributed proportionately across the task, unit, or course. Its form may be oral, written, practical, technical, demonstrative, or artifact-based. Defend is therefore neither a generic viva nor a separate secure lane appended to an otherwise open assessment. It provides direct confirmation of the reasoning associated with those learning outcomes for which an assured claim about individual attainment is made, while Generate, Evaluate, Refine, AI Critic, and Reflect develop and make that judgment visible. This task-matched approach avoids treating a return to high-stakes final examinations as the only alternative to unsupervised take-home assessment. The purpose of Defend is to confirm individual attainment and accountability for AI-assisted reasoning, while critical-thinking, ethical judgment, and metacognitive capability are developed across the full cycle.

2.3. Design principle

The design principle underlying SAGE is assurance by design, as developed in the companion paper [19]. SAGE does not divide assessment into two separate lanes in which capability development occurs in one and assurance occurs only in the other. It uses a hybrid, learning-outcome-mapped architecture in which open, AI-integrated capability development and direct assurance may coexist within the same task, be sequenced across related tasks, or be distributed across a unit or course. The intended learning outcomes determine which elements of knowledge, reasoning, judgment, or performance require direct confirmation; the epistemic form of those outcomes determines the appropriate Defend format; and the stakes and delegation risk determine the intensity and location of assurance [19]. Steps 1–5 develop capability through open, AI-integrated work in which students begin from their own thinking, test AI output against authoritative anchors, make accept/modify/reject decisions, use AI as a bounded critic, and reflect on the reliability of AI in the discipline. Step 6 provides task-matched direct confirmation of the reasoning associated with the learning outcomes for which an assured claim about individual attainment is made. Assurance is distributed because corroborative evidence is generated throughout the process and direct assurance is positioned where

it provides the strongest valid evidence, rather than being deferred automatically to a single high-stakes endpoint. Process documentation supports this inference but is not treated as a substitute for direct assurance where individual attainment must be certified. The model is epistemically matched: a coding outcome may be defended through a code walkthrough or supervised modification, a design outcome through a design justification, a cybersecurity outcome through a risk-and-control explanation, and a written-analysis outcome through reasoning about sources and claims. Bloom's taxonomy remains useful for describing increasing cognitive demand, but the underlying mechanism is better described through self-regulated learning, reflective practice, constructive alignment, and programmatic assessment [15–17].

2.4. SAGE student orchestration levels

The six-step process is intended to move student behavior from passive acceptance toward increasingly deliberate orchestration of AI output. Table 3 defines four observable task-level capability profiles. These profiles describe the quality of judgment demonstrated in a particular task rather than permanent learner identities. A student may therefore demonstrate different profiles across tasks, disciplines, stages of study, and levels of scaffolding.

Table 3. SAGE student orchestration levels and observable interpretations.

Orchestration level	Definition	Observable interpretation
Passive Acceptor *Baseline, not a target*	Accepts AI output with minimal critical evaluation or justification.	The student reproduces or endorses AI output without testing substantive claims against an authoritative disciplinary anchor. Decisions are absent or explained by restating the AI response.
Selective Adapter	Filters AI output using domain knowledge, accepting, modifying, or rejecting selected elements.	The student identifies substantive strengths and weaknesses and justifies selected accept/modify/reject decisions against sources, standards, tests, data, rubrics, or contextual constraints.
Balanced Integrator	Systematically combines human and AI contributions through explicit trade-off reasoning, contextual judgment, and relevant disciplinary and ethical checks.	The student systematically evaluates and refines AI contributions, performs relevant disciplinary and ethical checks, documents significant human decisions and edits, explains points of agreement and divergence, and justifies how the final position or artifact integrates AI output with personal and domain knowledge while retaining human decision authority.
Critical Synthesiser	Identifies gaps and biases and strategically customizes prompts, tools, or workflows to develop a traceable synthesis, reframing, solution, or human-centered contextual adaptation extending beyond both starting inputs.	The student proactively identifies gaps, biases, and contextual limitations; strategically adapts prompts, tools, or workflows and, where relevant, orchestrates multiple AI tools; and produces a substantive development present in neither the initial human contribution nor the AI output. The process includes a traceable explanation of how and why the development was produced and its ethical, creative, and disciplinary implications.

The AI guidance or permission level states what GenAI use is permitted. The SAGE orchestration level states the capability that the learning design is intended to develop and the rubric may assess. Once the relevant institutional security and permission settings have been established, the target orchestration profile identifies the quality of judgment to be developed, and the SAGE implementation protocol identifies the evidence through which that judgment must become observable. Permission,

capability, and evidence should therefore be constructively aligned: more extensive or strategically customized AI use requires correspondingly stronger evidence of evaluation, human decision authority, ethical judgment, and contextual adaptation. This alignment does not create two fixed assessment lanes. Within the hybrid SAGE architecture, corroborative evidence and direct confirmation are mapped to the intended learning outcomes and may be combined within one assessment, sequenced across assessments, or distributed across the unit or course. Earlier SAGE materials used the term *Corrective Editor* to describe transitional behavior in which a student detects and corrects errors in AI output. In the consolidated model presented here, corrective editing is treated as an observable behavior within the broader Selective Adapter profile rather than as a separate target capability.

3. Two-stage progression

SAGE operates through two stages that move students from guided practice to independent application with proportionate assurance. The stages describe changes in scaffolding and expected independence rather than an automatic classification of every student at a particular orchestration level. The progression is informed by Bloom's revised taxonomy [11] and aligns with TEQSA's distinction between open tasks that develop capability and secure tasks that verify individual attainment [9]. Within SAGE, open capability-building and direct assurance are connected through one assessment architecture rather than treated as unrelated activities.

Stage 1 (Formative): Tutorial-based scaffolded learning. Students practice the SAGE process under instructor guidance using structured prompts, worked examples, instructor-supplied disciplinary anchors, and immediate feedback from the instructor and peers. Students still produce a brief human baseline before interacting with AI, although the instructor may provide a standardized prompt to control variability across the cohort. The instructor supplies evaluation checklists and standards, facilitates peer review during Refine, guides the AI Critic interaction, and sets short metacognitive reflection tasks of approximately 150–250 words. The principal capability aim is to move students beyond Passive Acceptor behavior and establish Selective Adapter capability through evidence-based identification, correction, and rejection of AI output.

Stage 2 (Summative): Assessment-driven independent application. Students apply the complete SAGE cycle with reduced scaffolding. In Generate, students produce a human baseline and design prompts that incorporate the baseline, task constraints, and relevant disciplinary context. In Evaluate, they independently identify, justify, and apply authoritative anchors. Refine requires documented accept/modify/reject decisions with evidence-based justification. AI Critic is conducted using a domain-specific persona, followed by student adjudication of the critique rather than automatic acceptance. Reflect produces an extended evidence-based account of approximately 400–500 words. Stage 2 is intended to support Balanced Integrator capability by requiring students to coordinate human and AI contributions through explicit trade-off reasoning and contextual judgment. Critical Synthesiser capability is demonstrated only where the student produces a traceable synthesis, reframing, solution, or contextual adaptation extending beyond both the initial human contribution and the AI output; completion of Stage 2 does not automatically establish this highest profile.

Defend confirms ownership of the judgment associated with the learning outcomes for which the assessment makes an assured claim about individual attainment. The architecture is hybrid rather than divided into separate open and secure lanes: open AI-integrated work and direct assurance may

coexist within the same task, be sequenced across connected tasks, or be distributed across the unit or course. The intended learning outcomes determine what reasoning, judgment, or performance must be directly confirmed; the epistemic form of the competency determines whether confirmation should be oral, written, practical, technical, demonstrative, or artifact-based; and the stakes and delegation risk determine the intensity and location of the checkpoint [19]. Process evidence remains embedded in the AI-integrated work for learning and corroboration, while Defend or an equivalent direct-assurance checkpoint provides stronger evidence of individual attainment for the outcomes that require it. This distributed approach avoids assurance debt: the accumulation of unverified learning claims when verification is postponed to a single final event [19].

4. Step-by-step implementation protocol

Whereas Section 2.1 establishes the rationale and design intent of each step, Table 4 provides the operational counterpart: a procedure that educators can adapt directly into an assessment brief without reconstructing the design rationale. For each step, the table specifies the student task, the required artifact that renders the work traceable, and an assessment interpretation stating both what should receive credit and what should be avoided. These operational cues support marking and troubleshooting and are therefore additional to, rather than a restatement of, the conceptual account in Section 2.1. The step intent and sequencing are drawn from the SAGE implementation guide [8] and refined through empirical studies [1–7].

Selected SAGE steps may be used as formative scaffolds within tutorials and developmental activities. A minimum viable capability-building implementation requires Generate–Evaluate–Refine and a short Reflect statement. Full SAGE assessment adoption uses all six steps and formalizes both AI Critic and Defend. Where an assessment makes an assured claim about individual attainment, Defend or an equivalent direct-assurance checkpoint must be included proportionately at the task, unit, or course level. Educators are encouraged to adopt incrementally, beginning with Stage 1, scaffolded exercises before progressing to complete summative implementation.

Table 4. SAGE implementation protocol: student tasks, required artifacts, and assessment interpretation.

Step	Student task	Required artifact	Assessment interpretation
1. Generate	Produce and retain an initial human baseline, such as notes, an outline, preliminary interpretation, sketch, partial solution, or statement of uncertainty. Use that baseline and the task constraints to prompt GenAI for a comparator, draft, or alternative response.	Human baseline plus a prompt–output record sufficient to demonstrate how the student's initial thinking informed the AI interaction.	**Credit:** Evidence of prior student engagement and appropriate alignment among the human baseline, prompt, and task. AI is used as a comparator, coach, or contributor rather than as the first thinker. **Avoid:** Omission of the human baseline or submission only of AI output or final text.
2. Evaluate	Evaluate the GenAI output against explicit disciplinary criteria, including accuracy, relevance, completeness, alignment, evidence quality, and disciplinary correctness.	Evaluation log identifying substantive issues, their significance, and the authoritative anchor used to assess each issue.	**Credit:** Identification of substantive, domain-specific strengths, omissions, errors, and unsupported claims, together with an explanation of why they matter. **Avoid:** Superficial or generic evaluation without reference to disciplinary anchors.
3. Refine	Revise the work on the basis of the evaluation and document each major accept, modify, or reject decision against the relevant anchor or contextual constraint.	Change log recording the major modifications and the evidence-based rationale for each decision.	**Credit:** Student-directed refinement demonstrating disciplinary understanding and traceable judgment. **Avoid:** Re-prompting for a replacement draft without demonstrating purposeful revision or ownership of the changes.
4. AI Critic	Use GenAI as a bounded critical reviewer through a domain-specific persona, then adjudicate its critique against the relevant disciplinary anchor.	AI Critic prompt and output, together with the student's record of which critiques were accepted, modified, or rejected and why.	**Credit:** Discrimination among critiques and evidence that the student retains decision authority over the AI critic. **Avoid:** Uncritical acceptance of all AI feedback or unsupported dismissal of the critique.
5. Reflect	Produce a structured reflection on the AI contribution, domain errors identified, corrective reasoning applied, ethical and disciplinary considerations encountered, and implications for AI reliability in the discipline.	Structured reflection aligned with the course outcomes: approximately 150–250 words in Stage 1 and 400–500 words in Stage 2.	**Credit:** Specific linkage among the reflection, disciplinary anchors, accept/modify/reject decisions, AI Critic appraisal, and relevant ethical, professional, or contextual judgment. **Avoid:** Generic statements about convenience, speed, or workflow that are not connected to concrete decisions.
6. Defend	Demonstrate ownership of the reasoning associated with the learning outcomes requiring direct confirmation. Explain, justify, critique, reproduce, adapt, extend, or apply the submission logic in a task-matched format under supervised or otherwise controlled conditions.	Defend record, such as a rubric outcome, transcript note, supervised written artifact, technical demonstration, practical performance record, or documented response to an altered condition.	**Credit:** Ownership of reasoning, justification and critique of key decisions and trade-offs, and transfer or extension of the reasoning to a question, variation, or altered condition. **Avoid:** Describing the submitted output without being able to justify, critique, adapt, or extend its major decisions, corrections, or trade-offs.

5. The Defend step: Rationale and discipline implementation

The inclusion of a Defend step is not a theoretical preference; it is an empirically grounded response to a documented structural limitation of unsupervised assessment under GenAI-rich conditions. In SAGE, Defend is one component of distributed assurance rather than a stand-alone oral examination [19]. It provides direct assurance of individual reasoning at a selected point, while earlier

steps provide corroborative process evidence and, where appropriate, controlled assurance conditions. This matters for educators because take-home artifacts alone may not sufficiently evidence individual attainment; it also matters for students because a wholesale retreat to final examinations can narrow the ways in which genuine capability is demonstrated. Defend is therefore positioned as a proportionate assurance and equity mechanism: it checks ownership while preserving the learning value of open, AI-integrated assessment. This claim rests on the authors' own audit study [7] and is consistent with wider sector guidance that treats assessment redesign, rather than detection, as the sustainable response to GenAI [9, 10, 20].

TEQSA's 2025 resource identifies structural assessment redesign, not detection, as the sustainable response [9]. The University of Sydney's two-lane model provides an adjacent program-level response by distinguishing open tasks for learning from secure tasks for assurance [12]. SAGE does not reproduce that architecture as two separate lanes. Its assurance-by-design architecture is hybrid and learning-outcome-mapped: AI-integrated capability development and direct assurance may be combined within one assessment, sequenced across connected assessments, or distributed across a unit or course. The intended learning outcomes determine what must be assured; the epistemic form of the required competency determines how it should be defended; and the stakes and delegation risk determine where and with what intensity assurance should occur [19]. Within this architecture, SAGE distinguishes three evidence functions: corroborative evidence, such as AI logs, process trails, accept/modify/reject records, and reflections; controlled assurance conditions, such as supervised task windows or restricted environments; and direct assurance tasks, such as viva voce, live artifact walkthroughs, code explanations, practical demonstrations, or written reasoning checks. These forms may be combined in a hybrid design, but corroborative process evidence is not treated as a substitute for direct assurance where an intended learning outcome requires confirmation of individual attainment.

The Defend step is held to evidence ownership of reasoning on the following basis. Authorship of text can be delegated to a tool, but the capacity to reconstruct, justify, adapt, and defend the reasoning behind that text under task-matched challenge cannot be delegated in real time. Requiring the student to do so under appropriate conditions, therefore, tests the individual cognition that the open task alone cannot assure. To make this assurance reliable rather than impressionistic, three controls are recommended. First, Defend should be marked against a small, shared analytic rubric tied to the same criteria assessed in the open task, such as evaluation quality, refinement justification, critique discrimination, and transfer to a new condition. Second, examiner standardization should be supported through a shared prompt bank, common rubric descriptors, calibration on sample responses, and, where stakes are high, double or sampled second marking. Third, the risks of oral formats should be acknowledged and mitigated through non-oral Defend variants of equivalent demand, prompt structure in advance where appropriate, reasonable adjustments, and marking against criteria rather than fluency or confidence. Defend is thus a proportionate assurance checkpoint within a distributed design, not a high-stakes oral examination.

The Defend step is format-agnostic. The principle remains constant: the student demonstrates ownership of the reasoning, product, or decision pathway under conditions where the process cannot simply be simulated. Table 5 illustrates how Defend may be implemented across representative disciplines. The table should be read as a set of task-matched examples rather than as a prescription for a particular assessment format.

Table 5. Discipline-specific Defend tasks within a distributed-assurance design.

Discipline	What must be assured	Illustrative Defend formats	Corroborating process evidence
Cybersecurity	Threat prioritization and control-selection reasoning.	**Live/oral:** Timed risk-scoring of an unseen scenario, with examiner challenge concerning control selection and regulatory alignment. **Non-oral:** Supervised written reasoning check in which control choices are justified against the applicable standard.	Evaluation log and accept/modify/reject decisions recorded against the NIST or ISO anchor used in the open task.
Programming	Command of design logic and the capacity to diagnose and correct code.	**Live/oral:** Supervised code walkthrough or live debugging of a seeded error, with explanation of design decisions and trade-offs. **Non-oral:** Timed in-class modification of unseen code, submitted with a written justification of each change.	Change log summarizing refinements and AI Critic outputs showing which critiques were accepted, modified, or rejected.
Health Sciences	Clinical decision-making and recognition of where AI recommendations require human override.	**Live/oral:** Structured clinical reasoning viva applied to a new case. **Non-oral:** Supervised written case interpretation identifying where AI-generated recommendations must be overridden and explaining why.	Evaluation log against the relevant clinical-guideline anchor and reflection linking corrections to that guideline.
Business	Evidence-based strategic reasoning under stakeholder constraints.	**Live/oral:** Oral defense of a strategic recommendation with examiner challenge. **Non-oral:** Timed written response to an altered stakeholder constraint in which the revised position is defended with evidence.	Refinement log and reflection documenting how the recommendation was tested against the source evidence.
Engineering	Justification of design decisions and trade-offs, including safety and regulatory considerations.	**Live/oral:** Design defense with interrogation of the relevant constraints. **Non-oral:** Supervised written trade-off analysis responding to a modified constraint with reference to the governing standard.	Accept/modify/reject decisions recorded against the applicable design code, together with the AI Critic appraisal.
Law	Independent legal reasoning distinct from an AI-generated summary.	**Live/oral:** Moot argument or case-analysis viva citing relevant authorities. **Non-oral:** Annotated case memo produced in a controlled window, distinguishing an independently reasoned argument from an AI-suggested summary.	Refinement log recording where AI summaries were corrected against cited authorities.
Education	Pedagogical judgment and the rationale for accepting or rejecting AI-suggested approaches.	**Live/oral:** Teaching demonstration with reflective justification. **Non-oral:** Supervised written lesson rationale explaining which AI-suggested approaches were modified or rejected and why.	Reflection linking pedagogical choices to the Evaluate, Refine, and AI Critic decisions made during the task.

The Defend step does not require a full examination. A 5–10 minute micro-viva, a timed in-class application task, a live artifact walkthrough, a supervised debugging activity, a design explanation, or a short written reasoning check may be sufficient, provided the task requires the student to apply or justify the reasoning behind the AI-assisted work. Three defensible variants are recommended: a micro-viva targeting evaluation choices, corrections, and trade-offs; an in-class application task applying the same frameworks to a new prompt; or an annotated walkthrough with structured explanation of critical sections and spot checks. The critical point is proportionality: assurance should be strong enough to confirm ownership, but not so large that it recreates the inequities, anxiety, and workload of a high-stakes final examination.

6. Evidence base

The SAGE evidence base is a cumulative program of classroom design and evaluation rather than a single controlled trial. Across direct teaching and study implementations from 2022 to 2026, more than 1500 students completed SAGE-structured learning activities designed or delivered by the authors across undergraduate, postgraduate, first-year, cybersecurity, data-analytics, systems-analysis, professional-communication, and transnational contexts. Table 6 maps the principal empirical studies to their context, design, evidence, and contribution to the canonical framework. The table should not be summed to derive the program-level total because some publications examine overlapping cohorts, while some directly taught implementation cohorts are documented at program level rather than as separate study samples. The SAGE program comprises 12 core empirical, conceptual, and implementation studies within a broader portfolio of more than 40 publications and related scholarly outputs [8, 22].

Table 6. Core empirical evidence informing the development of the SAGE framework.

Study and context	Design and scale	Principal evidence	Contribution to the canonical framework
Origin study [1]	Two ICT units in Term 3 2022; experimental reflective-laboratory analysis incorporating three classroom case studies.	The reported case studies produced score gains of 85.6–92.9%, while incomplete or confusing designs fell to zero. Student responses also exposed over-reliance on fluent AI output and the need for explicit evaluative guidance.	Diagnosed the pedagogical problem from which SAGE developed and established Generate–Evaluate–Refine scaffolding, instructor guidance, and structured reflection as the initial core.
Cross-institutional Australian study [2]	Data-analytics cohort at the Australian Institute of Business Intelligence; cohort $n = 74$, including 67 complete responses; case-study and survey design.	Students rated the learning benefit at 4.14/5. The study identified variability in AI literacy and patterns of over-reliance outside the originating university context.	Provided early cross-institutional evidence and confirmed that students required a structured method rather than permission or access alone.
Cybersecurity implementation [3]	Undergraduate and postgraduate cybersecurity units, COIT12212 and COIT20263; $n = 105$; curriculum-embedded classroom implementation.	The study reported improvements of up to 87% in assessed performance and a 78% critical-thinking result under the study's operationalization; assignment completion exceeded 90% across the two cohorts.	Formalized the SAGE framework and demonstrated source anchoring, accept/modify/reject decisions, reflective evaluation, and AI Critic within disciplinary assessment.
Systems-analysis implementation [4]	Eighteen student groups across four campuses within a systems-analysis unit; full unit cohort $n = 281$; cross-campus embedded study.	Eighty-four per cent demonstrated balanced or selective, justified AI use, and every analyzed group detected systematic AI failure within the task sequence.	Established observable orchestration behaviors, the three-zone collaboration model, and the competency ceiling that motivated further development of AI Critic and Critical Synthesiser capability.
AI as Critic study [5]	Longitudinal follow-up involving $n = 280$ across five campuses, with qualitative analysis of six representative groups; preprint and manuscript under review.	Adversarial AI critique elicited contextual authority grounded in scope, constraints, risk, and regulation. Students produced theory-informed synthesis, while accessibility awareness remained dependent on recurrent prompting.	Examined AI Critic as a reproducible mechanism, developed the Apply–Defend–Synthesise pattern, supported the Critical Synthesiser profile, and established the need for recurrent scaffolding.

Continued on the next page.

Table 6 continued.

Study and context	Design and scale	Principal evidence	Contribution to the canonical framework
First-year verification study [6]	First-year ICT cohort, $n = 167$; supervised assessment in which AI use was permitted, followed by a 12-item task-proximal reflection instrument.	Seventy-three per cent reported systematic verification against the source article, 81% revised deeply, 58.1% worked iteratively with repeated source checks, and 77.8% identified verification accuracy as the most demanding aspect of the task.	Documented the verification difficulty and policy–practice divide, strengthened explicit source-anchoring guidance, and reinforced the need to distinguish the reported process from assured individual attainment.
Assessment-process audit [7]	Twenty-five group submissions representing 100 undergraduate and postgraduate cybersecurity students; retrospective five-check audit of required AI-process artifacts.	Only 3 of the 25 submissions were fully auditable. Logs, decision tables, and reflections could be incomplete, internally inconsistent, or insufficient to establish ownership when produced under unsupervised conditions.	Provided the direct empirical rationale for Defend, distributed assurance, and the distinction among corroborative evidence, controlled assurance conditions, and direct assurance.
UK–China transnational implementation [21]	Level 4 computing cohort in the Oxford Brookes College and Chengdu University of Technology Sino-British Collaborative Education Programme; $n = 131$; SAGE-guided graph-theory intervention.	Mean confidence increased from 2.84 to 3.63 on a five-point scale, with a statistically significant change and large effect size. Students also reported more positive perceptions of AI critique and greater caution toward unverified output.	Demonstrated transfer of the SAGE learning process beyond the originating institution and national context and supported its use for developing critical AI literacy in technical education.

The findings in Table 6 are study-specific and should not be interpreted as pooled effects. The studies differ in discipline, unit of analysis, instruments, cohort structure, and outcome measures. Their collective value lies in the cumulative design sequence: early studies diagnosed over-reliance and established structured evaluation; cybersecurity and systems-analysis implementations formalized source anchoring, orchestration, and AI Critic; the first-year study exposed verification difficulty; the submission audit established the limits of unsupervised process evidence; and the transnational implementation provided early evidence of transferability. The broader active-learning literature provides a compatible foundation for this evaluation-centered approach [13].

The empirical studies are complemented by design and policy extensions. SAGE-AP translates institutional permission settings into explicit pedagogical conditions for AI-supported assessment [18]. The assurance-by-design paper formalizes Defend, distributed assurance, assurance debt, and task-matched confirmation of student reasoning [19]. The implementation guide and SAGE resource suite provide the open implementation layer through which the combined program is made available to educators [8, 22].

6.1. Recognition, institutional integration, and external implementation

Three SAGE resources, covering curriculum design, student guidance, and the empirical evidence base, are listed in the Tertiary Education Quality and Standards Agency Gen AI Knowledge Hub, a curated national collection of higher-education resources for generative AI practice [20].

CQUniversity has adapted SAGE within its Guided-Assured Assessment Model and associated staff-development resources, incorporating the orchestration profiles and the six-step cycle [26].

Collaborative Australian implementation has included the Australian Institute of Business Intelligence [2]. International implementation has included a Level 4 computing cohort in the

Oxford Brookes College and Chengdu University of Technology Sino-British Collaborative Education Programme [21].

7. Assessment alignment and evaluation blueprint

Educators adopting SAGE should first map each intended learning outcome to the form and strength of evidence required. Outcomes developed through AI-integrated activity may be supported by prompts, evaluation logs, refinement records, AI Critic appraisals, and reflections; outcomes for which individual attainment must be certified require a task-matched Defend activity or equivalent direct-assurance checkpoint. This mapping determines what must be defended rather than requiring every component of the submitted artifact to receive identical verification. The educator should then define three components for the assessment package. First, a student submission checklist specifying the required artifacts: human baseline, prompt/output evidence, evaluation log, refinement change log, AI Critic log, structured reflection, and, where required by the outcome mapping, a Defend completion record. Second, a marking alignment or rubric skeleton that allocates marks to evaluation quality, refinement quality, critique discrimination, reflection specificity, ethical and disciplinary judgment, and Defend performance. The rubric should reward the quality of human judgment demonstrated, not the polish of GenAI outputs. Third, an outcome-mapped Defend prompt bank consisting of a small set of standardized questions or equivalent task variants, such as "Identify the most serious error in the initial draft and explain the correction", "Explain why this AI critique was rejected", or "Apply the reasoning to this altered condition and identify what must change."

A worked example illustrates how these components combine. In a cybersecurity policy task, the brief requires students to generate a draft access-control policy using GenAI (Generate), compare it against ISO/IEC 27001 controls and the relevant jurisdictional privacy legislation (Evaluate), revise it with a documented accept/modify/reject log (Refine), submit it to an AI persona acting as a compliance auditor and appraise which findings are valid (AI Critic), and reflect on which regulatory gaps the AI failed to detect (Reflect). The Defend checkpoint is a ten-minute micro-viva in which the student is given an unseen variant of the scenario and must justify two control selections and explain one error they corrected. The rubric awards marks for the soundness of the regulatory justification and the discrimination shown over the AI critique, not for the polish of the generated policy.

SAGE can be evaluated at the course level using indicators that do not require a full research study. Learning-process indicators include the proportion of submissions showing substantive evaluation findings, the frequency and quality of meaningful refinements, evidence of critique discrimination, specificity of reflection, and the quality of accept/modify/reject decisions. Assurance indicators include alignment between open artifact quality and Defend performance, the rate of Defend discrepancies where reasoning ownership is weak, and the distribution of corroborative, controlled, and direct assurance evidence across the unit. Feasibility indicators include marker workload per submission, Defend time per student, scalability constraints, and the availability of equivalent non-oral Defend formats. If the intention is later publication of an empirical evaluation, these indicators provide the basis for a replicable study design incorporating rubric coding, inter-rater reliability, cohort comparisons, and analysis of student progression across tasks.

8. Practical considerations

SAGE is designed to be adopted in resource-constrained environments, but educators should plan for several practical considerations. Workload management can be addressed through sampling or targeted Defend for large cohorts, standardized Defend prompts (which assess the quality of a student's reasoning about a position rather than requiring a single correct answer, and so accommodate differing opinions and perceptions provided the reasoning is defensible against the criteria), and structured logs that expedite marking. Equity and access considerations require that adoption does not mandate premium tools; educators should permit low-bandwidth alternatives where feasible. Policy coherence requires transparent tool permissions and disclosure rules aligned to institutional integrity requirements. Adaptation should preserve four invariants regardless of discipline: students must evaluate and justify accept/modify/reject decisions; students must produce traceable artifacts; direct assurance must be mapped to the intended learning outcomes rather than applied uniformly to every component; and, where an assured claim about individual attainment is made, students must demonstrate ownership of the relevant reasoning through a proportionate, epistemically matched Defend activity or equivalent direct-assurance checkpoint.

The SAGE framework, including worked discipline examples, assessment rubrics, marking guides, tutorial templates, implementation walkthroughs, a GenAI foundations module, student guidance, SAGE-R guidance for research students and supervisors, and a Defend/assurance planner, is freely available under Creative Commons licensing through the SAGE resource suite [8, 22]. A complementary teaching resource, SAGE 101: Generative AI Foundations Module, provides a ready-to-deploy introduction to responsible GenAI use for students, and related student-facing guides support implementation in coursework and research contexts [22].

8.1. Limitations and constraints

Several limitations and constraints should be weighed before adoption. On scalability, the Defend step adds supervised or task-controlled assurance time that does not reduce linearly with cohort size; in very large cohorts, full individual Defend may be infeasible, and educators may need sampling, group-then-individual probes, short standardized checkpoints, or targeted Defend for high-risk components, each of which trades assurance coverage against workload. On reliability, supervised judgments are only as consistent as the rubric and marker calibration behind them; without shared descriptors, a prompt bank, and calibration on sample responses, Defend outcomes may vary between markers. On equity and access, students differ in access to AI tools, prior prompting skill, English-language confidence, and comfort with oral formats; mitigations include permitting low-bandwidth tools, providing non-oral Defend variants of equivalent demand, and offering reasonable adjustments. On cognitive load, the full six-step cycle imposes substantial demand and may overwhelm novice learners if introduced at once; staged implementation and minimum viable adoption remain important. On transferability, the evidence base is strongest in computing, cybersecurity, systems analysis, and related information-systems contexts, although external Australian and international implementations now provide early evidence of broader applicability [2, 21]. Finally, as students increasingly combine multiple AI tools, additional guidance will be needed on cross-checking outputs across tools to detect hallucinations and reconcile conflicting results.

8.2. Boundary conditions

SAGE is not intended to replace secure assessment where a learning outcome explicitly requires unaided individual performance, nor should it be used where legal, ethical, clinical, professional, or institutional conditions prohibit AI use. It is also unnecessary where the assessed capability is predominantly psychomotor, interpersonal, or performative and an AI-mediated artifact is not relevant to the intended learning outcome. Contested disciplinary knowledge is not, by itself, a boundary to SAGE: comparing competing authorities may constitute the evaluative work the task is intended to develop. SAGE is less suitable where neither the educator nor the student can establish a legitimate basis for adjudicating among competing sources, standards, or interpretations. In all contexts, the framework should be adopted only where the use of AI, the selected disciplinary anchors, the required evidence, and the assurance method remain constructively aligned with the learning outcome.

A further opportunity, rather than a limitation, is that SAGE generates rich process data: prompts, evaluation logs, accept/modify/reject decisions, AI Critic appraisals, reflections, and Defend outcomes. These artifacts are amenable to learning-analytics treatment, for example, coding the depth of evaluation, tracing the alignment between documented reasoning and Defend performance, or modeling how evaluative discrimination develops across successive tasks. Such analysis could both inform formative feedback at scale and provide the basis for more rigorous empirical evaluation of the framework in future work.

9. Conclusions

This Perspective has presented SAGE as an adoptable GenAI pedagogy defined by a six-step orchestration process, a staged progression from scaffolded practice to autonomous application, four observable student orchestration levels, canonical terminology, and an assessment-assurance architecture that distributes evidence of student judgment across the learning process. Its central contribution is not that it permits students to use AI, nor that it asks them to verify AI output in general terms. Its contribution is that it teaches students how to verify by beginning with their own thinking, holding AI output against authoritative disciplinary anchors, documenting accept/modify/reject decisions, using AI as a bounded critic, and defending the reasoning behind the final work. The orchestration levels make the intended capability visible, while the distinction among assessment security, AI permission, orchestration capability, and evidence of assurance prevents these separate design decisions from being collapsed into a single scale. The approach specifies student actions, required artifacts, assessment interpretations, task-matched Defend formats, and a mechanism for confirming ownership of AI-assisted reasoning without reducing assessment reform to detection, prohibition, or a wholesale return to high-stakes final examinations. The expanded evidence base, direct implementation with more than 1,500 students, 12 core empirical, conceptual, and implementation studies within a broader portfolio of more than 40 publications and related scholarly outputs, TEQSA recognition, institutional integration, open resources, and collaborative external implementations support adoption while providing a foundation for further empirical evaluation.

A graduate who can type a prompt into a GenAI tool has no special advantage. A graduate who can start with their own thinking, identify the right anchor, test AI output against professional standards, recognize risks and false information, revise the work responsibly, reject flawed AI suggestions, and explain the reasoning behind the final decision is far more valuable. SAGE is designed to produce the

latter by making human judgment visible, teachable, transferable, and defensible when AI participates in learning and assessment.

Author contributions

Mahmoud Elkhodr: Conceptualization, methodology, investigation, resources, writing–original draft, writing–review and editing, visualization, and project administration. Ergun Gide: Conceptualization, methodology, investigation, validation, writing–review and editing, and supervision. Both authors have read and approved the final version of the manuscript.

Use of Generative-AI tools declaration

Generative AI tools (Anthropic Claude and OpenAI Codex) were used during the preparation and final proofreading of this Perspective to assist with structural formatting, identification of unresolved template text, and editorial refinement of author-drafted text. All intellectual contributions, including the research design, methodology, analysis, interpretation of findings, and pedagogical framework, are solely the work of the authors. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of this publication.

Acknowledgments

The authors gratefully acknowledge the contributions of all students who participated in SAGE-integrated units, workshops, studies, and associated learning activities from 2022 to 2026. Their engagement with AI-assisted learning, disciplinary verification, critique, reflection, and defense contributed directly to the iterative development and evaluation of the framework described in this Perspective.

Conflict of interest

The authors declare no conflicts of interest in this paper.

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

Ethics declaration

This Perspective does not report new research involving human participants or animals. It synthesizes previously published studies and implementation evidence; therefore, no new ethics approval or consent to participate was required.

References

1. Elkhodr, M., Gide, E., Wu, R. and Darwish, O., ICT students' perceptions towards ChatGPT: An experimental reflective lab analysis. *STEM Education*, 2023, 3(2): 70–88. https://doi.org/10.3934/steme.2023006

2. Sandu, R., Gide, E. and Elkhodr, M., The role and impact of ChatGPT in educational practices: insights from an Australian higher education case study. *Discover Education*, 2024, 3: 71. https://doi.org/10.1007/s44217-024-00126-6

3. Elkhodr, M. and Gide, E., The SAGE framework for developing critical thinking and responsible generative AI use in cybersecurity education. *Discover Education*, 2025, 4: 517. https://doi.org/10.1007/s44217-025-00935-3

4. Elkhodr, M. and Gide, E., AI leads, humans lead, or collaborate? Empirical findings and the SAGE roadmap for embedding GenAI in systems analysis and design education. *STEM Education*, 2026, 6(2): 194–229. https://doi.org/10.3934/steme.2026009

5. Elkhodr, M. and Gide, E., AI as critic: validating SAGE pedagogy for human authority and responsible GenAI use in systems analysis and design education. *EdArXiv Preprints*, 2025. Available from: https://osf.io/preprints/edarxiv/8j3xf

6. Elkhodr, M., Azra, A. and Gide, E., How first-year students actually use ChatGPT in permitted assessments: empirical typologies, verification gaps, and the policy-practice divide. *Research Square Preprint*, 2026. https://doi.org/10.21203/rs.3.rs-8628653/v1

7. Elkhodr, M. and Gide, E., Embedding assurance within learning: empirical evidence from the SAGE framework for repositioning take-home assessment in AI-integrated higher education. *STEM Education*, 2026, 6(4): 584–605. https://doi.org/10.3934/steme.2026024

8. Elkhodr, M. and Gide, E., Embedding generative AI in curriculum: the SAGE framework and evidence-based implementation guide, Version 2. Zenodo, 2026. https://doi.org/10.5281/zenodo.18383951

9. Tertiary Education Quality and Standards Agency, Enacting assessment reform in a time of artificial intelligence. Australian Government, 2025. Available from: https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/enacting-assessment-reform-time-artificial-intelligence

10. Lodge, J.M., Howard, S., Bearman, M. and Dawson, P., Assessment reform for the age of artificial intelligence. Tertiary Education Quality and Standards Agency, Australian Government, 2023. Available from: https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/assessment-reform-age-artificial-intelligence

11. Anderson, L.W. and Krathwohl, D.R., Eds., *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, New York, NY, USA: Longman, 2001.

12. Bridgeman, A., Moorhouse, A. and Davison, C., Program level assessment design and the two-lane approach. *Teaching@Sydney*, 2024. Available from: https://educational-innovation.sydney.edu.au/teaching@sydney/program-level-assessment-design-and-the-two-lane-approach/

13. Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., et al., Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, 2014, 111(23): 8410–8415. https://doi.org/10.1073/pnas.1319030111

14. Elkhodr, M. and Gide, E., Eds., *Generative Artificial Intelligence Empowered Learning: A New Frontier in Educational Technology*, 1st ed., New York, NY, USA: Chapman and Hall/CRC, 2025. https://doi.org/10.1201/9781003422433

15. Zimmerman, B.J., Becoming a self-regulated learner: an overview. *Theory Into Practice*, 2002, 41(2): 64–70. https://doi.org/10.1207/s15430421tip4102_2

16. Flavell, J.H., Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. *American Psychologist*, 1979, 34(10): 906–911. https://doi.org/10.1037/0003-066X.34.10.906

17. Schön, D.A., *The Reflective Practitioner: How Professionals Think in Action*, New York, NY, USA: Basic Books, 1983.

18. Elkhodr, M. and Gide, E., From permission to pedagogy: the Structured AI-Guided Education Assessment Policy (SAGE-AP) for generative AI in higher education. *Education Sciences*, 2026, 16(6): 986. https://doi.org/10.3390/educsci16060986

19. Elkhodr, M. and Gide, E., Assurance by design: embedding the SAGE Defend step in AI-integrated higher education assessment. *Frontiers in Education*, 2026, 11: 1872630. https://doi.org/10.3389/feduc.2026.1872630

20. Tertiary Education Quality and Standards Agency, Gen AI Knowledge Hub. Australian Government, 2026. Available from: https://www.teqsa.gov.au/guides-resources/higher-education-good-practice-hub/gen-ai-knowledge-hub

21. Fakirah, M., Elkhodr, M. and Hussain, S., Fostering critical AI literacy in UK–China transnational computing education: an implementation of the SAGE framework. *ICAIES 2026*, in press.

22. Elkhodr, M. and Gide, E., SAGE resource suite: GenAI-101 module, student guide, SAGE-R guidance, Defend/assurance planner, publications evidence page, and implementation resources. 2026. Available from: https://sage-framework.com

23. Perkins, M., Furze, L., Roe, J. and MacVaugh, J., The Artificial Intelligence Assessment Scale (AIAS): a framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, 2024, 21(6): 49–66. https://doi.org/10.53761/q3azde36

24. Long, D. and Magerko, B., What is AI literacy? Competencies and design considerations. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, 1–16. https://doi.org/10.1145/3313831.3376727

25. Ng, D.T.K., Leung, J.K.L., Chu, S.K.W. and Qiao, M.S., Conceptualizing AI literacy: an exploratory review. *Computers and Education: Artificial Intelligence*, 2021, 2: 100041. https://doi.org/10.1016/j.caeai.2021.100041

26. CQUniversity, *CQUniversity Guided–Assured Assessment Model Educator Guide*. Available from: https://www.cqu.edu.au/about-us/excellence-education/the-cquniversity-learning-and-teaching-strategy-futurenow/cqu-guided-assured-assessment-model-educator-guide

Authors' biographies

Dr Mahmoud Elkhodr is a Senior Lecturer in the School of Engineering and Technology at Central Queensland University, Sydney campus. His research spans cybersecurity, the Internet of Things, artificial intelligence in education, and smart cities. He is the developer of the SAGE framework and lead author of the SAGE evidence-based implementation guide. He serves in leadership roles within the IEEE NSW Section and is co-editor of *Generative Artificial Intelligence Empowered Learning* (Chapman and Hall/CRC, 2025) [14].

Professor Ergun Gide is a Professor of Information Systems in the School of Engineering and Technology at Central Queensland University, Sydney campus. His research interests include information-systems management, e-business, and technology-enhanced learning. He is co-editor of *Generative Artificial Intelligence Empowered Learning* (Chapman and Hall/CRC, 2025) and a co-investigator across the SAGE empirical research program.

AIMS Press