



The AI Authorship Trap: A Zero-Trust Competence Attestation Protocol for Generative-AI Education

Kwan Hong TAN

Singapore University of Social Sciences, Singapore

DOI: <https://doi.org/10.68050/JAMS.2026.870>

Received: 14 October 2026; Accepted: 20 October 2026; Published: 26 October 2026

ABSTRACT

Generative artificial intelligence (GenAI) weakens the inference from a polished academic submission to independent learner competence. The AI authorship trap arises when an institution treats the origin of an artifact as a substitute for evidence of learning. This design-science article proposes the Competence Attestation Protocol (CAP), a zero-trust-inspired, standards-compatible workflow linking declared AI-assisted work to a fresh independent challenge, delayed transfer task, accountable assessment, and scoped, revocable credential. The contribution is an evidence-binding and issuance policy, not a new cryptographic primitive or a claim that digital signatures prove knowledge. The architecture specifies attempt-bound records, nonce consumption, authorization guards, calibrated challenge banks, accessible assessment paths, and a governed dispute process. A reproducible Python abstract controller with 11 Boolean flags and 13 event types reached 24 states and exhaustively checked 312 outgoing state-event transitions. No violations of its four specified safety conditions were detected; four deliberately inconsistent states were also correctly flagged by the checking routine. Because the checker assumes truthful assessment events and trusted guards, these findings establish limited model consistency, not verified human identity, assessment validity, production security, or educational effectiveness. The paper distinguishes what standards already provide from CAP's proposed orchestration, analyses collusion, compromised identity and assessor services, task leakage and revocation abuse, and sets out a falsifiable multi-course evaluation programme. It thereby shifts assurance from asking whether AI wrote an answer to evaluating what a learner can independently demonstrate under an explicit and contestable policy.

Keywords: Generative artificial intelligence; competency assessment; zero-trust architecture; verifiable credentials; academic integrity; learning analytics

1. INTRODUCTION

Generative AI can produce essays, programs, explanations, design artifacts, and feedback at a quality that blurs the conventional relation between completed work and the knowledge of its apparent author. The educational problem is not that every assisted submission is illegitimate. Rather, an institution may award credit for a polished product while lacking defensible evidence that the learner can explain, adapt, or independently reproduce the associated competence. The distinction between performance during learning and durable learning is well established (Soderstrom & Bjork, 2015). More recently, Bastani et al. (2025) found in a school mathematics setting that access to an unguarded AI tutor could improve assisted practice performance yet worsen subsequent unaided performance relative to relevant comparators. This is context-specific evidence, not proof that AI assistance is generally harmful.

The AI authorship trap is defined here as a category error: substituting a claim about how an artifact was produced for a claim about what a person has learned. An AI detector that correctly classifies authorship still cannot establish competence; a student who legitimately used AI may nevertheless demonstrate it. Conversely, a demonstrably human-written answer can be memorized, purchased, copied from another person, or based on fragile understanding. Empirical studies also identify detection unreliability and disparate error risks, including concerns for non-native English writers (Liang et al., 2023; Weber-Wulff et al., 2023). These observations

motivate a change in the unit of assurance from text provenance to learner performance under independently specified conditions.

Existing work offers complementary responses. The Learning Assurance Frontier describes capacity constraints on independent verification and transfer (Tan, 2026c), the Cognitive Sovereignty Threshold examines human reasoning and contestability (Tan, 2026b), and the Adaptive Opportunity Floor addresses distributional risks in AI tutoring (Tan, 2026a). These accounts motivate criteria that CAP proposes to implement within an institutional workflow. However, digital badges and comprehensive learner records can already contain assessment criteria, evidence and competencies, while secure online assessment can already authenticate sessions and score performance (1EdTech Consortium, 2024). CAP's proposed contribution is therefore narrower and more testable: an explicit cross-stage issuance guard linking a versioned competence claim to fresh independent demonstration, delayed transfer, status control and accountable exception handling. Its combined workflow is a design proposal, not evidence that no existing institution has implemented similar combinations.

The research question is: How can an institution issue a portable, independently checkable claim of learner competence in a GenAI-rich environment without relying on opaque AI-authorship classification? Three subsidiary questions concern (RQ1) the minimally sufficient evidence chain, (RQ2) enforceable safeguards against replay and unauthorized issuance, and (RQ3) limits of trust, privacy, accessibility, and pedagogical feasibility.

The contribution is fivefold. First, the article separates three propositions that are often conflated: an artifact has an attributable history; an assessment event occurred under a stated policy; and a person possesses a defined competence. Second, it specifies CAP, an original state-gated protocol that connects learning events to a competence credential. Third, it derives four conditional protocol invariants and checks them in a finite abstract reference model. Fourth, it maps adversarial failure modes to evidence and revocation controls without claiming to solve the problem of dishonest or mistaken human assessors. Fifth, it proposes a falsifiable programme for testing validity, cost, fairness, and learner outcomes before institutional deployment.

2. BACKGROUND AND RESEARCH GAP

2.1 Product provenance is not a learning outcome

The traditional take-home assignment bundles several forms of evidence: the quality of a finished artifact, the history of its production, the learner's grasp of its contents, and the portability of that grasp to an altered task. GenAI decouples these components. For this paper, artifact provenance means a trace of declared contributors and operations; independent competence means demonstrated performance on an appropriately selected task without the prohibited assistance; and transfer means use of the relevant understanding on a materially changed task, including after a delay. The distinction accords with the learning-versus-performance literature and evidence for retrieval practice (Karpicke & Roediger, 2008; Soderstrom & Bjork, 2015).

Kasneci et al. (2023) describe both educational opportunities and challenges created by large language models. Effective governance therefore cannot consist solely of excluding AI. In CAP, assistance may be permitted and documented for the production phase, while a separate event, not the artifact itself, must justify the competence claim. This distinction avoids penalizing legitimate assistance and gives instructors a route to value process, reflection, and revision without treating any one of those signals as a complete proxy for learning.

2.2 Why AI authorship detection cannot carry the burden

Weber-Wulff et al. (2023) evaluated detection tools under varied writing conditions and found important weaknesses, particularly when text had been transformed or paraphrased. Liang et al. (2023) reported a bias risk for non-native English writing samples. Detection probabilities depend on models, samples, languages, editing, and prevalence; they do not directly quantify a student's mastery. In addition, a declaration of AI use

is not independently authenticated by classifier output. CAP therefore treats classifier scores, if collected at all, as investigatory leads requiring human review and never as issuance or punitive decision predicates.

2.3 Borrowing carefully from computer security and standards

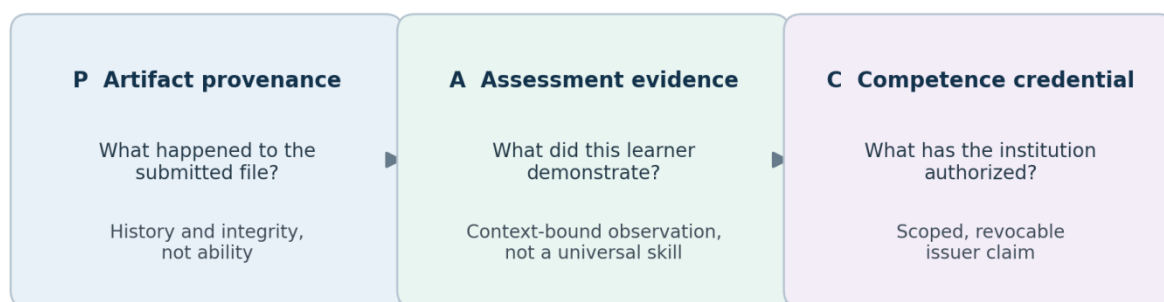
Zero-trust architecture replaces implicit trust based on network location with explicit, contextual authorization (Rose et al., 2020). CAP adapts that principle to educational claims: a signed submitted file, an authenticated session, or a prior high grade is not enough to authorize a new competence credential. Each critical transition requires fresh, scoped evidence. This is an analogy and design adaptation; NIST SP 800-207 does not define education assessment rules.

C2PA Content Credentials support assertions about content history and provenance, not truth, identity of the thinker, or mastery (Coalition for Content Provenance and Authenticity [C2PA], 2025). W3C PROV-O supplies a conceptual vocabulary for entities, agents, activities, and derivations (W3C, 2013). W3C Verifiable Credentials Data Model v2.0 specifies a machine-readable model for issuer assertions and verifiability (W3C, 2025); 1EdTech Open Badges 3.0 supports portable achievement claims (1EdTech Consortium, 2024). Learning Tools Interoperability (LTI) 1.3 supplies an education-tool launch and integration pathway (1EdTech Consortium, n.d.). CAP composes such components while keeping the pedagogical inference and the computer-security assertions explicitly separate.

2.4 Specific research gap

Authentic assessment asks students to do consequential tasks, digital badges encode recognition, content provenance describes artifacts, and AI detection estimates a source category. None, by itself, creates an evidence chain that demands a fresh unaided challenge, a delayed transfer observation, an explicit assessor decision, and revocable issuance attached to a specific competence definition. The proposed novelty is the conjunction of these conditions as one transparent protocol, not the invention of hashing, credential signing, challenge-response security, or retrieval-based assessment. Figure 1 previews the resulting trust distinction.

Three distinct questions. No automatic entailment.



CAP binds the three evidence types through explicit assessment and issuance gates.

Figure 1. Provenance, assessment evidence and competence credentials support different assertions.

2.5 What is new, and what is inherited

Open Badges 3.0 and the W3C Verifiable Credentials model already support portable, machine-verifiable achievement statements, and Open Badges permits criteria and evidence references. PROV-O and C2PA support different provenance use cases; learning platforms and testing services supply assessment workflows. CAP should not claim novelty in any of these component capabilities. Its proposed innovation is an assess-before-issue composition: a verified statement of independent challenge completion and subsequent transfer must exist under a stable rubric, approved assistance policy and bounded learner-to-attempt linkage before an issuer can sign the specific claim. The explicit distinction between an authenticated session and the person who

actually performs the work is another essential boundary. This composition could be implemented using existing standards, which is a strength for interoperability rather than a deficiency in technical originality.

The difference is operational rather than merely terminological. A badge may truthfully encode that an institution awarded a skill, yet may not require any delayed demonstration as a condition for issuance. An authenticated examination can observe a performance, yet may not bind that result to a portable credential and its later revocation status. CAP requires all three decision layers to remain separately inspectable: evidence generation, interpretation against a rubric, and credential authorization. The novelty claim is thus a proposed normative policy plus traceable reference implementation. Table 6 compares the different obligations; the comparison is qualitative and does not show superior learning or security outcomes (Kane, 2013; W3C, 2025).

3. METHODOLOGY AND DESIGN BOUNDARIES

3.1 Design-science procedure

The study follows design-science research, which treats a rigorously specified artifact and its evaluation as a research contribution (Hevner et al., 2004; Peffers et al., 2007). The process comprises: (1) identifying the authorship-to-competence inference problem; (2) deriving education, security, privacy, and accessibility requirements; (3) constructing the CAP objects, transitions, and issuer policy; (4) checking structural properties in a small executable state model; and (5) specifying empirical and security tests required for external validation. Evidence was synthesized from peer-reviewed AI education and assessment literature, security standards, and credential specifications. This is neither a human-subjects experiment nor a performance benchmark of deployed software.

The unit of analysis is one *attempt* to verify one explicitly described competence for one learner under one institutional assessment policy. Distinct assessments, retakes, accommodations, and revised syllabuses require separate versioned attempts. The validity of a claim is conditional on its rubric, challenge sampling, authentication, evaluator quality, and explicit assistance policy, not merely on the success of cryptographic checks.

3.2 Adversary model and trust assumptions

The protocol considers a learner who may submit GenAI-generated work, reuse an old assessment response, alter a record, collude with another person, or seek a credential without completing the independent task; a compromised client or LMS integration; and an attacker who replays or forges an API request. It assumes institution-managed signing keys remain uncompromised; an approved identity service correctly links the learner to the relevant session at the chosen assurance level; assessment staff and challenge platforms apply the declared assistance restrictions with known, imperfect reliability; and status services are sufficiently available to check revocation. The reference model abstracts these assumptions as trustworthy transition predicates.

The protocol does *not* guarantee that a learner has an internal mental state called ‘understanding’. A sophisticated impostor, inattentive assessor, coercive proctor, leaked item bank, inaccessible task, or corrupted issuer can defeat the semantic truth of the competence claim. CAP can make an evidentiary chain inspectable and guard the issuance workflow; it cannot turn assessment into a cryptographic proof of cognition. This boundary is essential to interpreting the results.

3.2.1 Authentication is not proof of the acting learner

Digital identity proofing, authentication and federated session assertions are related but distinct assurance activities (National Institute of Standards and Technology [NIST], 2025). A successfully authenticated institutional account establishes only that approved credentials or authenticators were presented at a specific assurance level. It does not exclude credential sharing, substitute test-takers, coerced participation, session hand-off, remote control or unauthorized off-camera assistance. Accordingly, a production CAP would record the institution’s selected identity proofing level, session authentication mode, assessment-condition attestation, risk flags and the accountable decision maker. Higher-stakes claims require stronger, proportionate controls

and an accessible alternative when the default control is unsuitable. No identity assurance level in the abstract model should be interpreted as empirical confirmation of the natural person's continuous presence.

3.2.2 Trust boundaries and failure domains

CAP distinguishes five trust domains: (1) the learner device and AI-assisted artifact, which are untrusted input; (2) the identity provider and session-binding service, which can be compromised or mistaken; (3) the challenge generator and assessment environment, which can leak tasks or fail to enforce permitted aids; (4) assessor, moderator and authorizing officer, whose judgments can be wrong or collusive; and (5) the evidence vault, issuer key service and credential status endpoint, which can be altered or unavailable. In the executable model, domains 2 to 5 are compressed into truthful event predicates; no network, human actor, keys, access control, random generator, race condition or multi-tenant data store is simulated. A deployable threat model must separately evaluate all five domains, including malicious insiders and colluding institutional staff.

3.3 Design requirements

Table 1 translates the problem into functional and non-functional requirements. A requirement is satisfied at protocol-design level only if an explicit object, guard, or governance rule implements it; pedagogical validity and operational performance remain open empirical questions.

Table 1. Traceable CAP design requirements and corresponding controls

ID	Requirement	Design control	Residual concern
R1	Separate evidence of artifact production from competence	Distinct provenance and competence records	Assessors may overvalue the artifact
R2	Bind each attempt to a learner and a policy	Context-bound authenticated session and policy version	Impersonation and authentication error
R3	Prevent reuse of a prior challenge	One-time nonce, expiry, signed outcome	Leaked challenge bank or collusion
R4	Require independent performance and transfer	Unaided challenge and delayed altered task	Task sampling, accommodation validity
R5	Prevent stale, incomplete or tampered issuance	Typed guard conditions, invalidate/hold states	Compromised trusted services
R6	Issue portable, inspectable claims	Signed credential with bounded scope and status	Verifiers may overinterpret the claim
R7	Minimize exposure of learner evidence	Pseudonymous IDs, commitments, segregated vault	Metadata leakage and re-identification
R8	Preserve accessible alternatives and appeal	Equivalent task modes, human review and dispute flags	Unequal burden and assessor bias
R9	Permit interoperable implementation	LTI launch, PROV-compatible mapping, VC/Open Badge export	Local integration and versioning costs

3.4 Evaluation strategy and status

Three analyses were carried out: requirements-to-controls traceability (Table 1), deterministic adversarial walkthroughs (Table 4), and an executable finite state-space exploration of the abstract issuance controller. These analyses can reveal missing workflow guards, but they cannot estimate learning gains, detection sensitivity, runtime performance, false accusation rates, or the effectiveness of identity verification. Section 7 proposes those prospective tests; the article does not present them as completed.

4. COMPETENCE ATTESTATION PROTOCOL (CAP)

4.1 Three independent categories of assertion

CAP differentiates (P) a *provenance assertion* that records how an artifact or submission was handled; (A) an

assessment-event assertion that documents an identified learner's performance under a stated constraint; and (C) a *competence-credential assertion* that an authorized institution evaluated sufficient evidence against a versioned rubric and awarded a bounded claim. The first does not entail the second, and the second does not automatically entail the third. For example, a perfectly signed record of a failed oral defence provides authentic evidence of failure, not grounds for awarding the credential. A credential's authentic signature verifies attribution to its issuer; it does not establish that the issuer assessed well.

Figure 1 illustrates the distinction; Figure 2 depicts CAP's proposed architecture. The assessment policy specifies whether AI may be used for preparation, where it must not be used, what assistance constitutes an accommodation, what counts as independent demonstration, and the degree of uncertainty that triggers human review rather than automatic issuance.

Competence Attestation Protocol (CAP): proposed evidence architecture

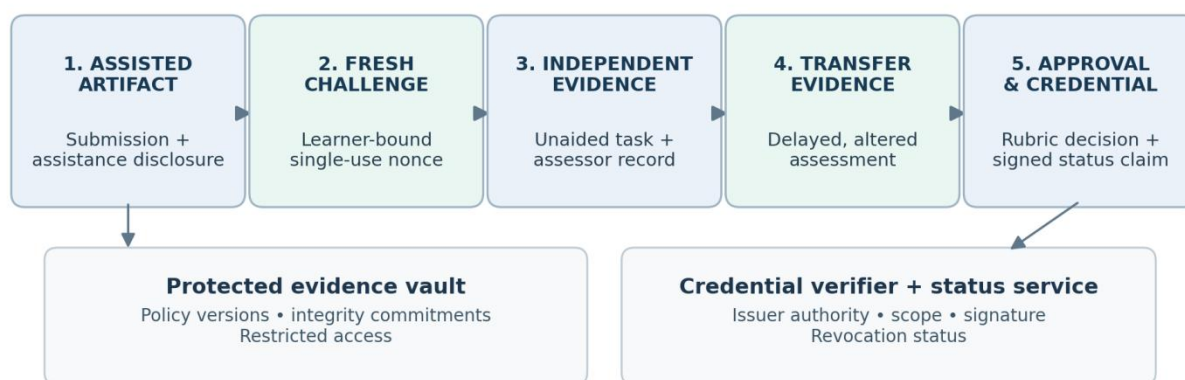


Figure 2. CAP binds independent learning evidence to a scoped credential through gated stages.

4.2 Data contracts and evidence minimization

A CAP *attempt descriptor* contains a randomly generated attempt identifier, pseudonymous learner handle, competence URI and version, rubric version, assessment-policy hash, accessibility mode, evidence-class requirements, and status. Its *artifact record* contains content digest, declared assistance category, derivation or source pointers when voluntarily provided, timestamp, and submitter identity binding. A *challenge record* contains a short-lived nonce, task-set version, policy constraints, assessment mode, and evaluator authority. An *outcome record* contains rubric-level judgments, assessor signature or institutional authorization proof, time, challenge reference, and exceptional conditions. Finally, the *attestation record* holds the resulting claim scope, issuer, issue time, validity conditions, evidence references, and revocation/status location.

Table 2. Minimal typed record schema for one CAP attempt

Record	Required fields	Verification rule
Attempt	attempt_id; learner_ref; competence_id; rubric_v; policy_v; status	Context and versions must agree across all records
Artifact	attempt_id; digest; declared_assistance; source_ref	Digest and provenance references match stored manifest
Challenge	challenge_id; attempt_id; nonce; task_v; expires_at; mode	Fresh, unspent, time-bounded, learner-bound
Outcome	challenge_id; rubric_v; assessor_ref; result; proof_ref; timestamp	Trusted assessor, correct task and policy version



Transfer	attempt_id; altered_task_v; result; assessor_ref; timestamp	Independent reassessment after a predeclared delay
Credential	competence_id; issuer; learner_ref; scope; evidence_root; status_uri	Issue only if all guards hold; verify issuer and status

The evidence_root is a cryptographic commitment to a canonical manifest of evidence references and policy versions, not a public repository of essays, video, biometric recordings, or detailed grade histories. Implementations should use domain-separated, keyed or randomly salted commitments where low-entropy values could otherwise be guessed, preserve the key/salt securely, and define deterministic canonicalization before signing. A bare hash is integrity evidence, not encryption, confidentiality, or proof of originality. The sensitive assessment objects remain in an access-controlled evidence vault with retention limits, authorized disclosure, correction procedures, and auditable access logs. W3C credential verification and provenance records are complementary layers, not one substitute for the other (W3C, 2013, 2025).

4.3 Evidence sequence and authorization gates

CAP's lifecycle has eight normal-stage events: registration; artifact submission; assistance disclosure; issuance of a fresh independent challenge; assessment of the unaided response; delayed transfer assessment; accountable assessor approval; and credential issuance. An assessor may reject, hold, request a repeat, accommodate, or refer for appeal at designated gates. At each step, a server verifies attempt ID, principal authorization, assessment-policy version, signed context, and validity interval. Critical updates must be transactionally bound to their unique attempt, so two parallel clients cannot both consume the same nonce or create contradictory active states.

Crucially, CAP does not define 'independent' as physical isolation by default. In one programming course it may mean a live alteration of a codebase without an LLM, under an approved environment; in another it may mean a structured oral explanation, a simulation, or an equivalent accessible format. The academic authority must predeclare the allowed aids and accommodations, consistent with human-centred deployment principles (Miao & Holmes, 2023). The protocol can record that these conditions were attested by an authorized assessor; it cannot sense every covert channel.

4.3.1 Challenge generation, confidentiality and comparability

Every target competence should have a versioned blueprint identifying construct, expected reasoning operations, permissible aids, task complexity, time limit, marking rubric and decision threshold. Authorized experts author multiple equivalent task families and unfamiliar surface variations. At issuance time, a controlled service draws a task family and instance using a cryptographically secure random source, binds the instance to a new unpredictable nonce and attempt identifier, sets a short expiry and records a signed task-set commitment. Learners receive only the task variant required for their session; repository access, authoring privileges and item exposure are logged and independently reviewed. Previous assessment artifacts, exact answer keys and rubric secrets must not be transmitted through a public credential. In a leaked-bank incident, administrators suspend affected task versions and provide a new, properly moderated assessment rather than silently accepting compromised results.

Difficulty must not be equated with randomness. Before high-stakes use, task families should undergo expert content review, pilot studies, inter-rater moderation, differential item functioning analysis where sample sizes permit, and monitoring of item exposure and unexpected pass-rate drift (American Educational Research Association et al., 2014). Comparability is defined at the level of an interpretation-use argument for the specified competence (Kane, 2013), not merely identical number of questions. Programmatic tasks may sample debugging or test-writing, while humanities tasks may require synthesis and oral explanation, and professional simulations may use observed decisions. Difficulty and scoring evidence should be reported by task family, modality and authorized accommodation, with performance standards established before operational use. A candidate who receives an accessibility-equivalent modality should be judged against the same underlying construct, with validated mode-specific procedures rather than automatic penalty.



4.3.2 Scoring, human accountability and escalation

A rubric should separate analytic criteria and identify critical errors, acceptable reasoning variation, partial-credit rules and evidence requirements. Independently trained assessors should apply the rubric, with double marking or adjudication for borderline, contested or high-impact claims. The assessment service records the assessor's authority, rubric version, evidence digest, reason codes, timing and any conflict-of-interest declaration. Automated scoring may support triage but cannot substitute for declared human responsibility without separately demonstrated validity. To limit assessor collusion, deployments should separate item-bank administration, marking, appeal handling and credential signing where feasible; randomly sample issued attestations for blind re-marking; and create protected whistleblowing and incident-reporting routes. Compromised authorizations trigger an issuance freeze, scope-specific review and a reasoned notice and appeal before permanent status change wherever circumstances permit.

An instance is not considered 'fresh' simply because it has a new timestamp. Its novelty must be evaluated against known exposures and replayable solutions. A nonce alone prevents an identical token from being consumed twice in a conforming transactional controller, but cannot prevent a learner from memorizing a leaked answer to a newly presented version. Policy authors must define minimum semantic variation, monitor task duplication and occasionally sample unseen tasks from a quarantined bank. These are governance and measurement controls outside the supplied Boolean model.

4.4 State transitions, invalidation, and verifier semantics

A single attempt progresses through enrolled, submitted, disclosed, challenge-open, challenge-passed, transfer-passed, approved, and active. Invalidated attempts are not silently repaired, and a new attempt must be opened for a retake. A reported compromise, record mismatch, expiry of an unfinished challenge, consent withdrawal where applicable, or assessment irregularity places the attempt into hold or invalidated status. Revocation changes the external credential's currently trusted status even when a previously exported credential retains a valid historical signature.

An issuance service must recheck current policy and evidence versions immediately before release. Once the service has confirmed an approved assessment, it may construct a W3C Verifiable Credential or an Open Badge conforming to the chosen profile (W3C, 2025; 1EdTech Consortium, 2024). Independent verifiers must check issuer authorization for the claimed competence, proof integrity, subject binding, credential status, expiration, and the precise scope of what was assessed. A valid signature from an arbitrary issuer cannot establish recognized competence, and offline checks of a stale status snapshot cannot guarantee that a previously issued credential remains in good standing.

Algorithm 1. CAP issuance gate, expressed as protocol-level pseudocode

```
function issue_competence(attempt_id, requested_claim):  
    a = load_attempt_for_update(attempt_id)  
    require authenticated_issuer() and issuer_authorised(requested_claim)  
    require a.status == APPROVED and not a.revoked and not a.invalidated  
    require exact_version_match(a.competence, a.rubric, a.policy)  
    require verify_bound_artifact_manifest(a) and disclosure_complete(a)  
    require challenge_is_fresh_and_consumed_once(a)  
    require verify_independent_outcome(a) and verify_delayed_transfer(a)  
    require assessor_signature_valid(a) and no_unresolved_dispute(a)  
    require current_evidence_root(a) == a.approved_evidence_root  
    credential = sign_scoped_claim(a, requested_claim, status_endpoint(a))  
    atomically_record_issuance_once(a.id, credential.id)  
    return credential
```

The conceptual algorithm relies on a database transaction or equivalent atomic compare-and-set mechanism. Merely adding a nonce field or a hash without enforcing uniqueness, freshness, and concurrency control would



not satisfy R3 or R5. The event log records verified transitions and issuer decisions, not a ground truth of knowledge. To avoid using an educational credential as a surveillance log, source responses and personal metadata are not included in publicly presented proofs by default.

4.5 Institution-scale integration

A deployment can place CAP behind an LTI 1.3 tool launched from a learning management system, with competency rubrics owned by the academic department and outcomes written to a protected institutional service. An internal PROV-compatible representation links artifact entities to assessment activities and responsible roles. Credential issuers translate approved outcomes to a VC/Open Badge profile. The record schema should be versioned, and consumers should reject unknown mandatory fields or unsupported credential schemas instead of interpreting them as successful evidence. The architecture does not require a blockchain, perpetual storage of learner work, or universal biometric monitoring.

5. ANALYSIS AND RESULTS

5.1 Conditional structural invariants

The controller enforces four narrowly defined safety requirements. I1 (issuance completeness): an active attestation requires all enrollment, submission, disclosure, passed-challenge, nonce-spent, delayed-transfer and approval flags, with neither invalidated nor revoked set. I2 (abstract challenge non-reuse): a passed challenge has a spent nonce flag and a replay-labelled event cannot advance the abstract controller. I3 (invalidation): a tamper event on a submitted attempt clears approval and active status; an expire event on a challenge that is open but not passed invalidates the attempt. I4 (revocation precedence): a revoke event on a submitted attempt clears approval and active status, and a revoked state never remains active. These are predicates over Boolean records and selected transitions, not proofs that a real nonce is unpredictable, a signature is authentic, or a student demonstrably learned.

The informal preservation argument is straightforward but conditional. The only transition that creates active checks all I1 prerequisites. The only pass-challenge transition marks the nonce spent and requires that it was previously unused. The tamper, unfinished-challenge expiry, and revoke transitions explicitly clear the relevant state, and no later transition can bypass invalidation or revocation. Other transitions either preserve the relevant flags or fail without changing the state. This argument holds only for an implementation whose transition predicates, atomic writes, external credential status semantics, and trusted actor checks actually satisfy the described design.

5.2 Bounded executable state-space exploration

To identify missing internal transition guards, the revised standard-library Python controller represents one attempt with 11 Boolean fields and 13 discrete events. Breadth-first search starts at all-false state, applies every event to every newly discovered reachable state and halts at fixed-point closure. It finds 24 distinct states; independently iterating all 13 event labels over every reached state yields 312 distinct state-event evaluations. Across those 312 evaluations, the encoded safety predicates and event-specific checks reported no violations. An eight-event legitimate issuance path reaches active. The abstract replay and forged-approval event labels are defined as no-ops; their non-effect is part of the specification rather than evidence of measured resistance to real forged requests. Four deliberately inconsistent negative-control states each trigger at least one safety failure. These checks establish that the checker detects simple inconsistent state assignments, not that it detects all implementation vulnerabilities.

This is exhaustive exploration of the *reachable abstract state graph*, not a proof over production systems or the 2,048 mathematically possible assignments to 11 Boolean flags. The model has no learner, assessor, identity-provider, time, network, cryptographic key or challenge-content representation. The breadth-first search, event definitions, guards, state-bit order, safety predicates, witness path and negative controls are supplied in the revised executable supplement. The 312-line transition trace can be regenerated as a CSV,



including no-op events; closure is checked by asserting that each target is among the 24 reachable states. The reference script was executed using Python 3.13.5 (standard library only); it is designed for Python 3.10 or newer. Appendix A provides replication steps and interprets every event guard without treating successful assertions as an empirical security benchmark.

Table 3. Structural verification outcomes in the abstract reference model

Check	Observed result	Interpretive boundary
Reachability closure	24 reachable states; 312 transitions checked	Model state space only
Issuance completeness (I1)	No violation in any reachable transition	Assumes trusted event predicates
One-time challenge (I2)	Replay event never advanced state	Does not model stolen live sessions
Tamper/pending-expiry invalidation (I3)	Tamper removed active/approval; pending expiry barred issuance	External caches must consult live status
Revocation precedence (I4)	Revoked attempt never remained active	Relies on verifier checking status
Positive issuing path	Valid eight-event path reached active	Does not imply genuine learning

5.3 Threat walkthroughs and residual failure

Table 4 executes conceptual attack traces against CAP's defined controls, distinguishing events the protocol structurally blocks from attacks that are outside the abstract trust boundary. 'Blocked' refers to rejection within a conforming implementation under the listed assumptions, not to experimentally demonstrated resistance to live attackers.

Table 4. Adversarial walkthroughs and trust limits

Scenario	CAP decision or evidence	Status of claim
Learner uploads fluent AI-written code and seeks a credential without a live skill check	Artifact submission is allowed; no fresh challenge outcome, so issuance guard fails	Structurally blocked
Former student replays a previously successful challenge token	Used nonce is rejected; challenge/attempt binding must match	Structurally blocked
Attacker alters an archived rubric result after assessment	Manifest mismatch forces hold or invalidation	Blocked if integrity/checking services are trusted
Client submits a forged assessor approval	Signer authority and signature validation required	Blocked under key-security assumption
AI assistance is concealed during a supposedly unaided session	Identity and approved-session controls may detect irregularity, but covert assistance can pass	Not solved by cryptography
An assessor mistakenly approves an inadequate explanation	A perfectly authentic credential can encode an erroneous judgment	Requires moderation, sampling and appeal
Learner uses an approved accessibility aid	Aid is recorded in assessment policy, without automatically voiding independence	Valid only if equivalence is established
Employer verifies an exported badge after revocation	Current credential status must be fetched	Not guaranteed under offline or stale verification



The important negative result is conceptual: CAP can prevent a credential from being issued *without required recorded evidence*, but it cannot ensure that recorded evidence proves true competence. This limitation does not defeat the protocol's purpose. It directs institutional investment to assessment design, staff calibration, and challenge security rather than to claims that content signatures or classifiers are epistemically decisive.

5.3.1 Additional institutional and human threats

Assessor collusion permits a genuine authorized account to approve fabricated performance; two-person review, role separation, conflict declarations and randomized post-issuance audits are deterrents, not formal guarantees. Compromise of an identity provider can transform a signed session assertion into a false identity binding; risk-tiered identity proofing and independent incident processes limit impact, but the state model assumes the session linkage is sound. Coercion and substitution can similarly produce 'valid' assessment records about the wrong individual. Challenge-bank leakage allows authentic but rehearsed answers to pass fresh-token checks; item rotation, exposure analytics and secure authoring reduce but do not eliminate this risk.

Accessibility-related false failures introduce a different threat: an otherwise competent learner may fail because the challenge format, proctoring requirement, timing or permitted aids obstruct the intended skill. Institutions should preregister equivalent modes, monitor differential error and burden, and distinguish technical session failure from non-mastery. Revocation abuse is also possible, for example when an administrator uses credential status to punish dissent or conceal an institutional error. Status decisions require least-privilege authority, dual review for disputed actions, durable explanations, appeal rights and audit sampling. A verifier who relies on cached status, an unrecognized issuer or an overly broad credential interpretation remains exposed even when the issuer's internal state is correct. Appendix B presents the corresponding controls, residual risks and recommended tests.

5.4 Worked learning scenario

Consider a second-year programming course where learners may use GenAI to draft a Python API integration. The submitted module passes unit tests and is accompanied by a declaration that a language model suggested the initial retry logic. The artifact record therefore truthfully captures permitted assistance, but the grade alone does not answer whether the student understands idempotency, exception handling, or timeout recovery.

CAP launches a fresh challenge asking the learner to diagnose a previously unseen duplicate-request bug, explain what can and cannot be safely retried, and modify the code under the course's declared no-LLM conditions. The instructor applies a versioned rubric covering diagnosis, reasoning, and implementation. A later assessment introduces a changed service contract or another language-independent task. Only after evidence for both demonstrations has been reviewed, with accommodations or appeals resolved, may CAP issue a scoped statement such as 'demonstrated foundational reasoning about safe API retries under the stated assessment policy.' The credential does not claim general software-engineering expertise, nor does it retrospectively label the assisted original as fraudulent.

If the student passes the first challenge but not delayed transfer, the attempt remains uncredentialed for this claim. If the institution decides to credential narrower achievement, it must define and assess a separate rubric rather than silently weaken the existing claim. This makes a competence assertion more interpretable to educators, learners, and downstream verifiers.

5.5 Comparison with alternative assurance patterns

Table 5. Comparative scope of competing assessment architectures

Approach	What it can substantiate	Central unresolved issue
AI-authorship classifier	Statistical source classification under tested conditions	Not equivalent to competence; domain shift and fairness



Traditional submitted artifact	Quality of deliverable under submission rules	Assistance and durable transfer unknown
Human oral defence alone	Contemporaneous explanation on sampled questions	Recording, replay, consistency, delay and portability
Signed digital badge alone	Issuer authenticity and declared claim	Assessment sufficiency remains hidden
CAP (proposed)	Auditable linkage between policy, fresh tasks, delayed transfer, approval and scoped credential	Authentic evidence may still reflect poor or biased assessment

The comparison is architectural, not a measured ranking. CAP deliberately inherits the costs of independent verification. The claimed contribution is not automatic superiority on every metric; it is sharper separation of what different evidence types can legitimately support.

6. DISCUSSION

6.1 From authorship policing to competence evidence

The practical shift is to stop asking a classifier to determine whether a finished answer deserves trust. Institutions should instead make the assessed claim explicit: ‘Can this student solve, explain, and adapt this type of problem under the specified conditions?’ This reframing permits productive GenAI use during learning while protecting an institution’s obligation to certify achievement. It also makes educational failure observable in a form that can guide remediation rather than escalating every AI-detection alert into an allegation of misconduct.

CAP operationalizes the independent-transfer emphasis in Tan’s Learning Assurance Frontier (Tan, 2026c). It also creates a possible computational input for the Cognitive Sovereignty Threshold’s emphasis on self-explanation and human oversight (Tan, 2026b). It does not merely rename either framework: its research object is the authorization and evidence-binding protocol at credential issuance, including nonce freshness, version matching, integrity commitments, revocation, and third-party verification.

The originality test should be strict: CAP is not a new content-authentication standard, a general zero-trust security theorem, an automatic competence detector or an improved psychometric instrument. It defines a versioned decision workflow and its evidence contracts, while inheriting existing trust anchors. Table 6 separates inherited features from proposed CAP-specific obligations. A prior solution implementing every CAP obligation would narrow the claim of novelty, and a systematic comparative implementation study is a priority for future research.

Table 6. Explicit boundary between existing infrastructure and CAP obligations

Component	What exists independently	Distinct additional CAP requirement
Content provenance and hashing	Artifact handling and integrity assertions	Never accept provenance as sufficient evidence of competence
Secure exam or oral defence	Authenticated assessment event and human score	Require fresh attempt-bound evidence plus delayed transfer and claim scope
Open Badges / verifiable credentials	Portable signed achievement with optional evidence and criteria	Enforce a prior multi-event issuance policy and live status checking
Learning assurance theories	Normative independent-transfer and human-agency criteria	Instantiate a testable gate and stateful authorisation controller
Zero-trust access control	Explicit contextual service authorization	Apply per-claim authorization to academic evidence with appeal safeguards



6.2 Trade-off between assurance and educational friction

Independent challenges consume instructor attention and can feel high-stakes if poorly designed. Bastani et al. (2025) show why the separation between assisted practice and independent performance can matter, but they do not prescribe CAP's design. Bućinca et al. (2021) found that cognitive forcing strategies can reduce overreliance on AI in decision contexts, while also creating user-experience trade-offs. CAP should therefore target high-value competence claims, sample tasks rather than audit every keystroke, reuse standards, and allow cohort-level scheduling. Where the marginal assurance benefit is smaller than student and instructor burden, a lighter protocol may be warranted. No cost-effectiveness threshold is asserted without institution-specific data.

6.3 Equity, accessibility and proportionality

Uniform surveillance is neither a necessary nor a sufficient assurance mechanism. Remote proctoring can disadvantage learners with disabilities, unstable connectivity, shared housing, or language differences. CAP's policy layer must permit equivalent tasks, additional time, assistive technology, alternate communication modes, and review by an accessible human channel. 'Unaided' means without unauthorized substantive solution assistance, not without every accommodation. An accessibility accommodation should be represented as a policy-compliant modality, not as an automatic credibility penalty. This extends the distributional caution in Tan (2026a). Institutions must validate that alternate challenges assess the same constructs and test for subgroup differences in completion, credentialing, appeals, and burden.

6.4 Privacy and governance of educational evidence

A hash of a student's answer, a device fingerprint, and a face recording have different privacy properties. Credentials should normally expose the minimum necessary claim, issuer, validity, and status; supporting responses remain protected for authorized audit. Learners should know what is collected, why it is necessary, how long it is retained, and how inaccurate records may be contested. The issuer must segregate credential-verification metadata from detailed assessment data and implement role-based or attribute-based access, retention, key rotation, and incident response. The NIST AI Risk Management Framework offers a broader lens for governing AI-related risk, not a certification of CAP itself (Tabassi, 2023).

A selective-disclosure credential can reduce unnecessary data sharing when supported by the chosen technical profile, but it cannot eliminate inference from the disclosed claim or patterns of verification. A public append-only ledger is not required: tamper-evident institutional audit logs and revocation endpoints can provide needed accountability without publishing educational trajectories.

6.5 Where the protocol must not be used

CAP should not be used to infer neurological capacity, assign universal intelligence scores, eliminate human due process, or force standardized assessment where a profession requires plural forms of expertise. It is unsuitable for mechanical credential automation when rubrics are unreliable or assessors cannot identify task-specific competence. Systems must not label students dishonest merely because they cannot produce an unaided response at one moment. A failed or disputed attempt is an assessment outcome that merits appropriate feedback and review, not a moral diagnosis.

7. EVALUATION PROGRAMME AND RESEARCH IMPLICATIONS

The next phase should combine technical testing with an educational effectiveness trial. First, independent engineers should implement and audit CAP's authorization, concurrency, nonce consumption, signature handling, credential status, schema downgrade resistance, revocation, and evidence vault. Security outcomes would include unauthorized issuance rate under seeded faults, replay acceptance rate, time to invalidate a compromised credential, cross-tenant data exposure, and verification latency. These outcomes must be measured rather than inferred from the abstract state exploration.



Second, educators should preregister a multi-course field study comparing (A) conventional artifact assessment with transparent AI-use rules, (B) artifact plus independent live challenge, and (C) CAP with live challenge, delayed transfer, and linked credential. The primary educational endpoint should be blinded performance on a novel delayed task assessed by trained independent raters, not apparent assignment fluency or detector score. Secondary outcomes should include retained skill, learner agency, appeal outcomes, educator workload, student perceived fairness, assessment time, and subgroup-specific exclusion or completion rates. Allocation should be randomized where feasible at an appropriate clustering level; analysis should account for course, assessor, prior achievement, accommodation, and missing outcomes.

The design yields testable hypotheses rather than unearned conclusions. H1: adding independently sampled challenges will improve the correspondence between awarded credentials and delayed transfer performance compared with artifact-only grading. H2: integrity-linked credential issuance will reduce unauthorized issuance in a fault-injected implementation compared with an otherwise identical implementation lacking freshness, version, and status guards. H3: unless accessibility and human review are built in, stronger assessment controls may disproportionately increase burden or rejection in particular learner subgroups. H4: requiring both independent and delayed evidence may improve claim specificity at the cost of longer credential completion time. Negative, null, and heterogeneous results should all be reported.

Statistical plans should predefine competence rubrics, minimum important differences, error rates, sample sizes, attrition procedures, assessor agreement statistics, and measures of false reassurance. A qualitative component should interview learners and educators about due process and whether the protocol improves confidence or merely shifts administrative burden. Publication of anonymized aggregate outcomes, the protocol version, instrument descriptions, and the reference model would permit meaningful replication without disclosing sensitive student records.

7.1 Empirical validation and preregistration thresholds

A credible validation trial must specify the exact competence claims, candidate pool, learning support allowed in each experimental arm and transfer assessment horizon before outcome inspection. The recommended primary measure is independently and blindly scored delayed transfer on a concealed task family. Secondary measures are decision agreement between awarded competence and independently observed performance, calibration curves, false assurance and false exclusion rates, completion time, staff hours, appeals and subgroup disparities. Statistical uncertainty should be reported with confidence intervals and a prespecified missingness strategy. For human scoring, publish adjudication rules, inter-rater reliability and results stratified by task family and approved modality. Predefine success and non-inferiority thresholds with subject experts; this manuscript does not invent an effect size, sample size or institutional cost estimate.

7.2 Independent security assessment and counterexamples

Separately, a controlled implementation trial should fault-inject compromised assessor roles, malicious signed identity assertions, challenge leakage, parallel nonce consumption, lost status connectivity, corrupted evidence objects and abusive revocation. An evaluator should report which attack reached which trust boundary and whether the intended safeguard failed open or failed closed. The present reference model cannot adjudicate these attacks because its event predicates assume truthful external checks. The explicit research programme thus distinguishes (i) computational consistency of a specified controller, (ii) implementation robustness under adversarial faults and (iii) psychometric validity of the substantive credential, three evidentiary layers that must never be conflated.

8. CONCLUSION

GenAI makes authorship a tempting but misleading shortcut for judging student learning. The proposed Competence Attestation Protocol connects AI-assisted artifacts to bounded, fresh independent challenges, delayed transfer, accountable assessment and revocable credentials, using established standards wherever possible. Its distinct contribution is the explicit authorization policy linking these events, not the discovery of



new cryptographic assurances. A complete exploration of 24 reachable states and 312 event transitions found no violation of the four encoded safeguards, and deliberately inconsistent negative controls were detected. Because external identity, task validity, assessor honesty and production system behaviour are assumptions rather than modeled facts, stronger confidence requires independent implementation tests and educational trials. CAP is therefore an auditable design hypothesis for ‘What can this learner demonstrate under a specified policy?’, not proof of competence by signature alone.

DECLARATIONS

Funding: No external funding was received for this conceptual and reference-model study.

Competing interests: The author declares no competing interests.

Ethics and participant data: No human participants were recruited and no learner-level data were collected for this study. The worked scenario is illustrative.

Data and code availability: No human-participant dataset was collected. The revised Python 3.10+ standard-library source code accompanies the submission as `CAP_Reference_Model.py`; running the documented export command generates the complete 312-row transition edge list (`CAP_312_Transitions.csv`). All reported state counts and negative controls are independently reproducible from this model. Model results are not student-outcome observations, cryptographic verification or a deployed security evaluation.

Appendix A. Reproducible abstract specification

State definition: $S = (\text{enrolled}, \text{submitted}, \text{disclosed}, \text{challenge_open}, \text{challenge_passed}, \text{nonce_used}, \text{delayed_passed}, \text{approved}, \text{active}, \text{invalidated}, \text{revoked})$. Each component is Boolean. The initial state is $(0,0,0,0,0,0,0,0,0,0,0,0)$. A single simulated state represents one assessment attempt; a repeated attempt is outside the current model and would receive a new identifier and state. An event that fails its stated guard is defined to leave the state unchanged. No external identity assurance, cryptographic verification, learner-performance measurement or challenge-generation randomness is executed by these transitions.

The abstract event alphabet is $E = \{\text{enrol}, \text{submit}, \text{disclose}, \text{open_challenge}, \text{pass_challenge}, \text{pass_delayed_transfer}, \text{approve}, \text{issue}, \text{replay_nonce}, \text{forge_approval}, \text{tamper}, \text{expire}, \text{revoke}\}$. The first eight events support a positive issuance trace. Tamper, expire and revoke are conditional terminalizing transitions. The two named adversarial events `replay_nonce` and `forge_approval` are encoded as no-ops for all states, explicitly expressing assumed rejection rather than testing any real cryptographic defence.

Replication protocol: save the supplied Python source and run ‘python `CAP_Reference_Model.py` --export-transitions `CAP_312_Transitions.csv`’. Confirm 11 components, 13 events, 24 reachable states, 312 outgoing transitions, four negative controls, one positive issuance path and closed reachability. CSV records each source state as an 11-bit string in the ordered variable list above, the event name, target state and a no-op flag. The script uses breadth-first enumeration to a fixed point and checks every outgoing transition for both state predicates and relevant event-specific postconditions. Results can vary only if the model source or the interpreter’s behaviour changes; real assessment quality is not measured.

Invariant interpretation: I1 is a state predicate requiring prerequisites whenever active is true. I2 combines a state predicate (`challenge_passed` implies `nonce_used`) with checks that `replay_nonce` cannot change state and a spent nonce cannot be used to set `challenge_passed`. I3 requires invalidated states not to remain active and checks that tamper of a submitted attempt, or expiry of an open unpassed challenge, terminalizes the attempt. I4 requires revoked states not to remain active and verifies that revoke on a submitted attempt deactivates the record. The checker also requires approved states to have eligible delayed transfer. These checks are stronger than merely counting the transition labels, but still assume correct event semantics.



Table 7. Complete event transition guards and effects in the executable model

Event	Guard required	Effect when accepted
enrol	not enrolled	enrolled = 1
submit	enrolled and not submitted	submitted = 1
disclose	submitted and not disclosed	disclosed = 1
open_challenge	disclosed; not open, invalidated or revoked	challenge_open = 1
pass_challenge	challenge_open; nonce unused; not invalidated/revoked	challenge_passed = 1; nonce_used = 1
pass_delayed_transfer	challenge_passed; not yet passed; not invalidated/revoked	delayed_passed = 1
approve	delayed_passed; not invalidated/revoked	approved = 1
issue	all 7 prerequisite flags; not invalidated/revoked	active = 1
replay_nonce / forge_approval	none; adversarial labels intentionally rejected	no state change by definition
tamper	submitted	invalidated = 1; approved = 0; active = 0
expire	open challenge and not passed	invalidated = 1; approved = 0; active = 0
revoke	submitted	revoked = 1; approved = 0; active = 0

Appendix B. Residual threat register

Table 8 identifies credible attacks and governance failures explicitly omitted from the Boolean controller. Proposed controls are design recommendations, not evaluated efficacy claims. Institutional deployment requires evidence about both attack prevention and fair, proportionate redress.

Table 8. Extended residual-risk register and external validation priorities

Failure mode	Residual risk despite state guards	Required independent test or governance control
Compromised identity provider / impersonation	Authenticated session can be falsely attributed	Risk-tiered proofing, account recovery audit, session substitution tests
Assessor collusion or bribery	Authentic approval may encode fabricated performance	Role separation, blinded audit, conflicted-case escalation
Coercion / remote assistance	Genuine learner may not act independently	Proportionate permitted-aid checks, interview and protected appeal
Challenge item leakage	Fresh nonce attached to memorized solution	Exposure monitoring, task rotation, quarantined tests
Accessibility false failure	Assessment modality changes what is measured	Equivalence study, accommodation-aware calibration, subgroup monitoring
False or retaliatory revocation	Issuer status tool used abusively	Dual authorization, immutable reason audit, independent appeal
Compromised signing key or stale verifier	Credential forged or revoked claim appears valid	Key rotation drills, status-freshness and outage tests
Data vault breach or deanonymization	Sensitive evidence exposed despite public hash	Role audit, retention limits, privacy incident simulation

Appendix C. Challenge operations

The proposed challenge workflow begins before individual issuance. Domain experts define a construct map, authentic and alternate tasks, pass/fail thresholds, assessor competencies and a versioned policy of allowed aids. The testing team pilots tasks, estimates item difficulty and scoring agreement, and retires exposed

content. At an attempt, a service authorizes the session, creates a fresh nonce, selects an eligible variation from a secured bank, signs a task commitment, and prevents reuse through a transactional one-time-consumption check. It then stores the assessed result and passes only authorized summaries to the issuance service. After a prespecified delay, a materially altered task is evaluated against an independently documented rubric. A second assessor or appeal channel resolves disputes before approval. No such pipeline is executed in the supplied abstract controller; this specification is a deployment requirement and empirical research agenda.

Comparable scores require comparable interpretations, not identical materials. Mathematics may use new problem instances with the same reasoning demand; programming may require debugging an unseen program; humanities may use oral defence of unfamiliar counterarguments; and vocational fields may use simulated choices under altered constraints. For each discipline, institutions should report construct coverage, rater agreement, task exposure and modality differences, then test whether a CAP credential forecasts subsequent unaided transfer better than the comparator assessment. Failure to demonstrate validity, acceptable burden and subgroup fairness would argue against deployment, even if every computer-security guard remained intact.

Table 9. Proposed minimum challenge and scoring lifecycle

Stage	Required evidence	Release / escalation rule
Blueprint and versioning	Construct, tasks, aids, rubric, content owner	Approval before live use; freeze versions per attempt
Selection and session	CSPRNG variant, fresh nonce, attempt binding, expiry	Abort or retry on failure; no reused token
Independent assessment	Task response, rubric scores, assessor authorization	Human review of anomalous or borderline case
Delayed altered assessment	Predefined lag, altered family, independent rubric score	No credential on unresolved transfer failure
Issue and subsequent review	Moderated approval, commitment, issuer scope, status	Sign once; audit, dispute, revoke or reinstate with due process

REFERENCES

- 1EdTech Consortium. (2024). Open Badges specification, version 3.0. <https://standards.1edtech.org/open-badges/specifications/standards/v3p0>
- 1EdTech Consortium. (n.d.). *Learning Tools Interoperability (LTI)*. <https://www.1edtech.org/standards/lti>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association. <https://www.testingstandards.net/>
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1). <https://doi.org/10.1145/3449287>
- Coalition for Content Provenance and Authenticity. (2025). *C2PA technical specification, version 2.2*. https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968. <https://doi.org/10.1126/science.1152408>

10. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
11. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. <https://doi.org/10.1016/j.patter.2023.100779>
12. Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://doi.org/10.54675/EWZM9535>
13. National Institute of Standards and Technology. (2025). Digital identity guidelines (NIST Special Publication 800-63-4). <https://doi.org/10.6028/NIST.SP.800-63-4>
14. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
15. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-207>
16. Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. <https://doi.org/10.1177/1745691615569000>
17. Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
18. Tan, K. H. (2026a). The adaptive opportunity floor: Distributionally robust safeguards for inclusive AI tutoring. *Journal of Advanced Multidisciplinary Studies, 1*–13. <https://doi.org/10.68050/JAMS.2026.768>
19. Tan, K. H. (2026b). The cognitive sovereignty threshold: An interdisciplinary humanities and sciences framework for human agency in generative AI-assisted knowledge work. *International Journal of Research Publication and Reviews, 7*(7), 2445–2454. <https://doi.org/10.55248/gengpi.07.0726.2117>
20. Tan, K. H. (2026c). The learning assurance frontier: A capacity-constrained theory of generative AI, knowledge verification, and independent learning transfer. *Journal of Advanced Multidisciplinary Studies, 16*–31. <https://doi.org/10.68050/JAMS.2026.762>
21. W3C. (2013). *PROV-O: The PROV ontology*. <https://www.w3.org/TR/prov-o/>
22. W3C. (2025). *Verifiable credentials data model v2.0*. <https://www.w3.org/TR/vc-data-model-2.0/>
23. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*, Article 26. <https://doi.org/10.1007/s40979-023-00146-z>