

The Learning Assurance Frontier: A Capacity-Constrained Theory of Generative AI, Knowledge Verification, and Independent Learning Transfer

Kwan Hong TAN

Singapore University of Social Sciences, Singapore

DOI: <https://doi.org/10.68050/JAMS.2026.762>

Received: 02 October 2026; Accepted: 12 October 2026; Published: 20 October 2026

ABSTRACT

Generative artificial intelligence (GenAI) can improve the speed and polish of assessed work without necessarily strengthening knowledge that learners can independently reproduce. Existing accounts of cognitive offloading and responsible AI identify this risk, but provide limited formal guidance for allocating verification effort when AI output volume rises faster than learner and educator capacity. This article develops the Learning Assurance Frontier (LAF), a capacity-constrained theory of AI-assisted learning that distinguishes production, verification, and delayed independent transfer. Using analytic modelling, constrained optimisation, and transparent numerical scenarios, it derives conditional propositions: (i) the fraction of AI-supported work that can be fully checked declines once verification demand exceeds finite review capacity; (ii) an explicitly specified assurance-adjusted educational objective has a unique, capacity-dependent optimal AI-assistance intensity; and (iii) unequal review resources can create assurance disparities even when students have equal access to AI tools. Additional results establish consequence-weighted allocation of limited review time and a distributionally robust safety margin for uncertain verification demand, both under explicit assumptions. In illustrative, non-empirical scenarios, increasing weekly verification capacity from 1.8 to 3.0 hours shifts the model-optimal assistance intensity from approximately 0.35 to 0.71. The article translates these conditional results into an assessment architecture incorporating evidence checks, self-explanation, delayed unaided transfer, and equity safeguards. It specifies an experimentally testable research programme, while explicitly avoiding any claim that the illustrative parameters are measured educational effects.

Keywords: generative artificial intelligence; independent learning transfer; epistemic verification; educational assessment; cognitive offloading; learning assurance frontier

INTRODUCTION

The educational value of generative artificial intelligence cannot be inferred from the quality of what students submit while an AI system remains available. An assignment can be completed more rapidly, appear more coherent, and receive favourable assessment while leaving its author unable to reconstruct the governing principles, detect errors, or apply the knowledge in an unfamiliar setting. This is a particular challenge for higher education, professional education, and digitally mediated learning where fluent artifacts are readily mistaken for durable competence. The distinction between current performance and durable learning is well established (Soderstrom & Bjork, 2015), but generative tools intensify it by performing substantial parts of the cognitive task themselves.

The empirical picture is neither uniformly optimistic nor uniformly adverse. Noy and Zhang (2023) document measurable productivity benefits for professional writing. In an educational field experiment, Bastani et al. (2025) found that unrestricted chatbot assistance could impair subsequent performance when assistance was withdrawn, whereas appropriately scaffolded assistance mitigated that concern. Chen and Cheung (2025) report positive average GenAI-related learning effects across higher education studies, alongside a non-significant pooled effect for metacognition. The OECD (2026) similarly distinguishes successful AI-assisted

task completion from the learning that pedagogically designed assistance can support. These findings are compatible rather than contradictory if performance gain and transfer gain are separate outcomes.

A second difficulty concerns finite attention. Each generated answer, calculation, citation, recommendation, or worked solution may require inspection. Faster generation raises the volume of items to be checked, but the student and teacher still possess bounded time, domain expertise, and attentional resources. Tan (2026d) formalised a verification bottleneck for AI-accelerated knowledge work; Tan (2026a) separately identified cognitive sovereignty as a human capacity to explain, verify, contest, and retain AI-mediated knowledge. Neither account by itself specifies an educational objective in which the intensity of assistance is optimised against both verification scarcity and subsequent unaided competence. That is the narrower and distinct contribution of the present study.

This article asks three questions. RQ1: Under what conditions does GenAI-assisted educational production exceed a finite capacity to verify its claims? RQ2: How does verification capacity change the theoretically preferred intensity of AI assistance, given an explicit and contestable educational value function? RQ3: How should institutions allocate scarce verification time while measuring learning independently of assisted performance? The analysis produces a formal frontier, exact comparative statics, an auditable priority rule, illustrative scenarios, and a prospective validation design. Its claims remain conditional on transparent assumptions until tested with learner-level observations.

THEORETICAL FOUNDATIONS AND CONTRIBUTION

Learning is not the same construct as assisted performance

Instruction should cultivate knowledge and skills that persist beyond the immediate task and transfer to new problems. Testing and retrieval can improve long-term retention (Roediger & Karpicke, 2006); effective learning strategies include distributed practice and practice testing (Dunlosky et al., 2013). Productive failure research further demonstrates that attempting a problem before instruction can improve later conceptual understanding and transfer (Kapur, 2014). These traditions imply that some difficulties removed by automation are educationally functional. This paper does not assume that more effort is invariably better. Cognitive load theory warns that unnecessary processing can obstruct learning (Sweller, 1988). The design problem is to preserve productive reasoning while eliminating genuinely extraneous work.

An education-specific antecedent is Structured Reflective Dialogue. In an action-research study of 32 part-time adult learners, Tan (2026b) evaluated Socratic questioning, structured academic controversy, and guided reflective journaling as mechanisms for critical reflection. Its reported pre/post improvements motivate a candidate instructional mechanism, but do not identify a causal effect of GenAI or validate the LAF. The present model therefore treats learner reconstruction and delayed unaided transfer as outcomes to be tested, rather than inferred from reflective participation.

Gen AI assistance must be pedagogically differentiated

An AI tool may provide a completed solution, a hint, a misconception diagnosis, an alternative explanation, an adversarial critique, or formative feedback. These activities differ in their likely cognitive consequences. The ICAP framework distinguishes interactive, constructive, active, and passive engagement, thereby providing a conceptual basis for differentiating learning-oriented assistance from answer consumption (Chi & Wylie, 2014). Kasneci et al. (2023) highlight both the potential of educational large language models and the need for critical-literacy and fact-checking skills. AI literacy includes recognition of system limitations and appropriate critical evaluation (Long & Magerko, 2020). Accordingly, the assistance variable introduced below is an exposure-intensity index, not a universal behavioural property of all AI designs.

Cognitive offloading concerns how external actions redistribute information-processing demands (Risko & Gilbert, 2016). Its relevance here is not an assumption that offloading is necessarily harmful; rather, a student may benefit from external support while becoming less able to independently reconstruct selected steps.



Accordingly, any proposed reduction in independent competence remains an empirical hypothesis, not a consequence of the definition of offloading.

What is genuinely new in the present model?

The Learning Assurance Frontier is not a relabelling of earlier notions of AI reliability, cognitive sovereignty, or queueing oversight. It integrates three elements that those contributions do not jointly supply: a student-centred capacity boundary expressed in instructional time; an educational objective whose assistance optimum can be solved and differentiated with respect to that boundary; and an outcome architecture that treats delayed, unaided transfer as the criterion required to interpret apparent performance improvements. It also demonstrates how identical tool access may coexist with unequal checking capacity. The model should be regarded as a falsifiable design theory rather than as a validated psychometric scale or an empirical general law.

To clarify the scope of novelty, Table 1 compares the principal conceptual antecedents with the distinct analytical question addressed here. These frameworks are complementary: the present model does not claim to supersede their psychological or pedagogical explanations.

Table 1. Existing perspectives and the narrower analytical contribution of the Learning Assurance Frontier.

Perspective and source	Primary explanatory focus	Specific addition in LAF
Cognitive load theory (Sweller, 1988)	Processing demands and conditions that support or hinder learning	Finite verification workload measured in hours relative to review capacity, not a measure of mental load.
Cognitive offloading (Risko & Gilbert, 2016)	Moving cognitive work to external resources and metacognitive offloading choices	Modelled boundary between assistance-driven output volume and independently checkable knowledge claims.
ICAP and AI literacy (Chi & Wylie, 2014; Long & Magerko, 2020)	Engagement quality and competence in using and critically evaluating AI	A budget-linked assistance optimum and an auditable check-allocation rule, conditional on a fixed assistance mode.
Cognitive sovereignty threshold (Tan, 2026a)	Human ability to understand, challenge, and retain AI-mediated work	A learner-centred weekly capacity constraint and explicit delayed unaided-transfer endpoint.
Verification bottleneck (Tan, 2026d)	Review congestion and epistemic reliability in AI-accelerated knowledge work	Educational value function, consequence-weighted claim checks, and prediction of unequal assurance under equal tool access.

METHODOLOGY AND MODEL SPECIFICATION

Research design and epistemic status

The study follows analytic theory construction and computational illustration. First, it defines observable work-flow variables. Second, it derives propositions from explicit mathematical assumptions. Third, it evaluates deterministic parameter cases and sensitivity patterns. Fourth, it maps the constructs to a prospective classroom evaluation. No participants were recruited, no educational records were accessed, no intervention was delivered, and no significance tests are represented as completed. The numerical outputs are worked implications of stipulated inputs, not estimates from schools or universities.

Demand, capacity, and unit consistency

Let a denote normalised AI-assistance intensity on the interval from zero (none) to one (highest permitted

intensity). Let λ_0 denote the expected baseline count of verifiable claim units per week, not the number of assignments. Let g represent the assumed proportional increase in claim units between no assistance and maximal assistance; c is the average hours required to check one unit, and B is hours of verification capacity per week. All units of W and B are therefore hours per week. In practice, a claim unit must be operationalised beforehand, for example a checkable factual assertion, cited source, mathematical step, or consequential recommendation. Counts should not be equated mechanically with learning difficulty.

	$W(a) = c\lambda_0(1 + ga), \quad 0 \leq a \leq 1.$	(1)
--	---	-----

Equation (1) is a deliberately simple, testable arrival-demand relationship. It assumes that the relevant claim volume grows linearly with assistance intensity; neither linearity nor the direction of the growth is asserted to hold for every AI-supported task. The assumption is appropriate for transparent sensitivity analysis and must be replaced if field data support saturation or nonlinear effects.

	$\rho(a, B) = \frac{W(a)}{B},$ $v(a, B) = \min\left\{1, \frac{B}{W(a)}\right\}.$	(2)
--	--	-----

For positive B and positive W , ρ measures verification demand relative to available capacity. The quantity v is the theoretical upper bound on the share of equally costly claim units that can receive a complete check, if the entire budget is devoted to verification and there is no duplication, interruption, or rework. It is not a measure of actual truthfulness, a probability that claims are correct, or a guarantee that a completed check detects every error. When ρ exceeds one, complete verification of all units is infeasible within the assumed budget.

	$a_{\text{cap}}(B) = \frac{1}{g} \left(\frac{B}{c\lambda_0} - 1 \right).$	(3)
--	--	-----

The capacity threshold in Equation (3) is the value of assistance intensity at which demand equals capacity. It is meaningful as an interior threshold only when the result lies between zero and one. If it is negative, the course is overloaded even at $a = 0$; if greater than one, full verification remains feasible across the modelled range. This boundary is accounting-based and therefore does not depend on the educational objective introduced next.

The variables, units, and principal scope restrictions are consolidated in Table 2.

Table 2. Definitions, units, and assumptions of the Learning Assurance Frontier model.

Symbol	Interpretation	Unit or bound
a	Normalised assistance intensity	0 to 1
λ_0	Baseline weekly checkable claim units	claim units/week
g	Proportional claim-volume sensitivity to AI assistance	dimensionless, positive
c	Mean full-check time per claim unit	hours/claim unit
B	Learner or learner-and-teacher review budget	hours/week, positive
$W(a)$	Potential verification workload	hours/week
$\rho; v$	Load ratio; maximum fully checked share	dimensionless
r, k	Productivity/scaffolding slope; diminishing-return	value units; value units

	curvature	
θ	Penalty for verification-capacity deficit	value units/(hours/week) ²
n_j, h_j, p_j, s_j, c_j	Stratum size, harm, error risk, detection, check cost	specified per risk stratum
$\alpha, \kappa, \sigma W$	Risk tolerance, workload dispersion, coefficient of variation	α and κ dimensionless; σW h/week

Assurance-adjusted educational objective

A teaching intervention must not maximise verifiable output alone. To obtain a tractable comparison, define a reduced-form assurance-adjusted value function F . Its first term captures the assumed benefit of increasing assistance, such as faster formative feedback or useful scaffolding. Its concave second term captures diminishing marginal educational value as tasks become more heavily AI-mediated. The last term penalises the square of verification demand left uncovered by available review capacity. The objective is a decision-theoretic scenario function; it is not a directly measured learning score. Coefficients r , k , and θ must be calibrated, justified, and tested for each educational setting.

	$F(a; B) = ra - \frac{k}{2}a^2 - \frac{\theta}{2}(\max\{0, W(a) - B\})^2.$	(4)
--	--	-----

Assume $r > 0$, $k > 0$, $\theta \geq 0$, $g > 0$, $c > 0$, $\lambda_0 > 0$, and $B > 0$. The convex deficit penalty is one transparent way to represent the possibility that unchecked claims have nonlinear educational consequences. Alternative loss functions are possible; indeed, future evidence may show that verification deficits have weak, asymmetric, delayed, or context-specific effects. Equation (4) intentionally makes that normative choice visible rather than describing a synthetic outcome as an observed educational fact.

	$d = c\lambda_0 - B, \quad b = c\lambda_0 g.$	(5)
--	---	-----

Operationalisation and measurement safeguards

Assistance intensity a is a prespecified, treatment-specific exposure index, not a universally comparable proportion of cognition performed by AI. For a controlled study, define the eligible task steps and the maximum AI affordances before randomisation; calculate a as the weighted share of steps on which the assigned AI support is made available, using fixed weights and an independently reviewed task blueprint. Record realised uptake separately, including no use, hints, critique, and complete-answer requests. To avoid conflating fundamentally different mechanisms, estimate separate workload-response coefficients g for answer-first, hint-first, and critique-first designs, or hold assistance type constant during the main model test. This design prevents an apparent increase in a from being mistaken for an invariant behavioural effect across AI products.

A checkable claim unit is a task-relevant assertion or reasoning step that two trained annotators can independently classify under an agreed codebook; pretest unit segmentation, report coder agreement, and adjudicate disagreements. Measure c with timed, logged independent checks using primary sources or reproduced calculations, and measure B as actual protected verification hours after excluding tutoring and unaided-practice time. Estimate p_j from blinded expert audit samples and s_j from seeded, independently adjudicated error-detection tasks, with uncertainty intervals. Assess learning through rubric-scored, blinded, delayed near- and far-transfer tasks completed without AI, while controlling baseline competence and task difficulty. These observables are not interchangeable: checking a claim does not demonstrate retention, and

delayed transfer does not by itself validate the quadratic shape of F. Such parameters require prospective calibration before instructional decisions are based on a fitted frontier.

ANALYTIC RESULTS

Proposition 1: finite-capacity assurance

Proposition 1. For fixed B, c, λ_0 , and $g > 0$, the maximum completely checked share v in Equation (2) is non-increasing in a. It is strictly decreasing wherever $W(a) > B$. Proof. Where $W \leq B$, $v = 1$. Where $W > B$, $v = B/[c \lambda_0 (1 + g a)]$ and its derivative with respect to a is negative because the denominator increases strictly with a. At the boundary, v is continuous. Therefore, increased production may coexist with a decreasing fraction of fully checked claim units. This is a bound on coverage, not an assertion that the fraction of true claims declines.

Proposition 2: the model-optimal assistance level

Proposition 2. Under the assumptions stated for Equation (4), F is strictly concave on the entire feasible interval and therefore has a unique maximiser. The unconstrained maximiser can be written piecewise, followed by projection to the feasible interval. Define d and b as in Equation (5).

	$a_{\text{unc}} = \begin{cases} \frac{r}{k}, & d + b \frac{r}{k} \leq 0, \\ \frac{r - \theta b d}{k + \theta b^2}, & d + b \frac{r}{k} > 0. \end{cases}$	(6)
	$a_{\text{opt}}(B) = \min\{1, \max\{0, a_{\text{unc}}\}\}.$	(7)

The proof is constructive. In the no-deficit region, F has derivative $r - k a$ and a stationary point at r/k . If that point lies in the no-deficit region, it is also the global unrestricted maximiser. Otherwise, the optimum lies in the positive-deficit region where the derivative is $r - k a - \theta b(d + ba)$, giving the second branch in Equation (6). The second derivative equals $-k$ in the no-deficit region and $-(k + \theta b^2)$ in the deficit region; the derivative is continuous at the threshold. Thus strict concavity holds, and clipping to $[0,1]$ produces Equation (7). The result is conditional on the chosen objective, rather than a universal recommended AI-use percentage.

Proposition 3: more checking capacity can support more beneficial AI use

Proposition 3. If the optimal assistance intensity is strictly interior and occurs in the positive-deficit region, its response to B is strictly positive and given by Equation (8). At boundaries or in the unconstrained no-deficit regime, the response is weakly non-negative, although it can be zero. The proof follows by differentiating the second branch of Equation (6) using $d = c \lambda_0 - B$; the relevant coefficients are positive. Hence institutional investment in verification capacity can expand the modelled set of beneficial AI-assisted activities without requiring universal avoidance of AI.

	$\frac{\partial a_{\text{opt}}}{\partial B} = \frac{\theta b}{k + \theta b^2} > 0.$	(8)
--	---	-----

This comparative-static claim does not imply that increasing checking hours always produces higher actual learner achievement. Increased B may also be costly, and time spent validating claims can displace retrieval, discussion, or self-explanation. A field evaluation must therefore distinguish reallocated hours from genuinely additional instructional capacity.



Proposition 4: consequence-weighted verification

Uniform checking is often inefficient when some mistakes matter much more than others. Consider J distinct claim strata. Let n_j be the number of claim units in stratum j , and let h_j be the harm weight if a claim in stratum j remains wrong, p_j the prior probability that it is wrong, s_j the probability a complete check detects and corrects such an error, c_j the hours needed for the check, and x_j the proportion of that stratum selected for checking. Under additive expected avoided harm, constant within-stratum parameters, and divisible claim strata, the following objective is maximised, subject to the stated budget and unit bounds:

	$\max_x \sum_{j=1}^J n_j h_j p_j s_j x_j$ $\sum_{j=1}^J n_j c_j x_j \leq B, \quad 0 \leq x_j \leq 1.$	(9)
	$\pi_j = \frac{h_j p_j s_j}{c_j}.$	(10)

Proposition 4. An optimal allocation checks units in descending order of π_j , allowing at most one partially checked stratum unless there are ties. Proof. If a solution funds a lower-index priority unit while leaving a higher-index priority unit unchecked, transferring a small amount of time from the former to the latter increases the objective by the positive difference in avoided harm per hour. Repeated exchanges eliminate all such inversions. This is the fractional-knapsack result. It is not guaranteed for indivisible checks, dependent errors, nonconstant detection probabilities, or checks that themselves teach content; those cases require richer optimisation and potentially a mixed-integer or stochastic programme.

The model also offers a defensible educational qualification: high-error-consequence checking should not consume all time reserved for essential independent practice. The priority rule applies only to the portion of capacity explicitly allocated to verification. A minimum no-AI retrieval reserve is a separate pedagogical constraint, not an optional by-product of risk scoring.

Independent evidence matters as much as checking effort. Repeated agreement among model outputs is not equivalent to independently corroborated evidence when error sources are correlated. Tan (2026c) formalises this common-mode-error problem for multi-agent institutional decisions. In the classroom, that framework motivates a procedural safeguard, not a new assumption in Proposition 4: high-consequence claims should be checked against primary evidence or independently reproduced reasoning rather than against additional outputs from the same model family.

Proposition 5: an uncertainty-aware assurance bound

The deterministic frontier uses expected verification demand. Real assignments differ in claim count, ambiguity, error severity, and the time necessary for source checking. To represent workload uncertainty without assuming normality, let the nonnegative random variable $\tilde{W}(a)$ denote realised weekly verification demand, with mean $\mu\tilde{W}(a) = c \lambda_0(1 + ga)$ and finite standard deviation $\sigma\tilde{W}(a)$. Choose an acceptable upper-tail overload probability α before the activity. A chance-constrained design requires Equation (11). This constraint is optional and distinct from maximising the educational value function: high-consequence curricula may choose a small α , whereas routine low-stakes practice might permit partial checking.

	$\mathbb{P}(\tilde{W}(a) \leq B) \geq 1 - \alpha, \quad 0 < \alpha < 1.$	(11)
--	--	------

Proposition 5. Without specifying a probability distribution, the one-sided Chebyshev-Cantelli inequality implies that the mean-plus-margin condition in Equation (12) is sufficient for Equation (11). The margin is conservative and distributionally robust for the stated mean-variance ambiguity set. Proof. For any $t > 0$, the probability that realised workload exceeds its mean by t is at most $\sigma W^2/(\sigma W^2 + t^2)$. Requiring this bound to be no greater than α and setting $t = B - \mu W$ gives the stated sufficient condition. The guarantee concerns verification workload, not the probability that student knowledge is correct or that every completed check detects an error.

	$\mu_W(a) + \sigma_W(a) \sqrt{\frac{1 - \alpha}{\alpha}} \leq B. \quad (12)$	
--	--	--

Under the special assumption of a constant coefficient of variation $\kappa \geq 0$, so that $\sigma W(a) = \kappa \mu W(a)$, the sufficient bound yields the maximum risk-screened assistance intensity shown in Equation (13). When the expression is negative, even $a = 0$ fails the chosen guarantee; it must not be silently clipped to zero and presented as safe. When it exceeds one, the whole permitted interval satisfies this particular mean-variance screen. Otherwise, the conservative admissible region is $0 \leq a \leq a\text{-rob}$. A risk-sensitive educational designer can maximise Equation (4) over that region instead of the full unit interval. The result is a tractable assurance safeguard, not a distribution-free prediction of actual achievement.

	$a_{\text{rob}}(B, \alpha, \kappa) = \frac{1}{g} \left(\frac{B}{c\lambda_0 \left[1 + \kappa \sqrt{\frac{1 - \alpha}{\alpha}} \right]} - 1 \right). \quad (13)$	
--	--	--

The bound also exposes a non-obvious trade-off: two classrooms can have the same mean review capacity but different workload variability. Even if their deterministic thresholds coincide, the more variable classroom receives a lower risk-screened assistance ceiling. This provides a reason to measure variation in checking time and claim complexity, rather than reporting only class-average AI use. Appendix A provides an independent derivation and specifies the conditions under which the bound ceases to apply.

ILLUSTRATIVE ANALYSIS AND RESULTS

Deterministic benchmark

For numerical transparency, assume eight baseline claim units per week, $g = 1.5$, and a full-check cost of 0.20 hour per claim. Weekly demand is therefore $1.6 + 2.4a$ hours. Let $r = 1.2$, $k = 1.2$ and $\theta = 0.5$. These are dimensioned illustrative values, not fitted coefficients. Compare weekly verification budgets of 1.8, 2.4 and 3.0 hours. All quantities below follow algebraically from Equations (1) to (7), without sampling, random seeds, or an unreported dataset.

Table 3 reports all deterministic scenario calculations; these values are not observed learning outcomes.

Table 3. Model-implied outcomes under three hypothetical weekly verification budgets. Values are analytical illustrations, not empirical observations.

B (h)	Threshold	Optimal a^*	W at a^* (h)	Checkable share	F(a^*)	F(1)
1.8	0.083	0.353	2.447	73.60%	0.244	-0.61
2.4	0.333	0.529	2.871	83.60%	0.412	-0.04

3	0.583	0.706	3.294	91.10%	0.526	0.35
---	-------	-------	-------	--------	-------	------

The results show that the model-optimal assistance intensity increases with B . Under the middle capacity scenario, the threshold beyond which complete checking is infeasible occurs at $a = 0.333$, while the model optimum is approximately 0.529. That optimum permits a theoretical maximum completely checked share below 100%, because the assumed productivity benefit outweighs the convex deficit penalty at the margin. This distinction is important: the frontier does not declare that every unverifiable claim is acceptable. It demonstrates that the answer depends on pedagogical loss weights and independently chosen safety or integrity constraints. For high-stakes domains, a curriculum designer can impose an additional hard requirement of $v = 1$ for critical material.

Figure 1 depicts the intended instructional assurance architecture, while Figure 2 plots the capacity-dependent model optima for all three illustrative scenarios.

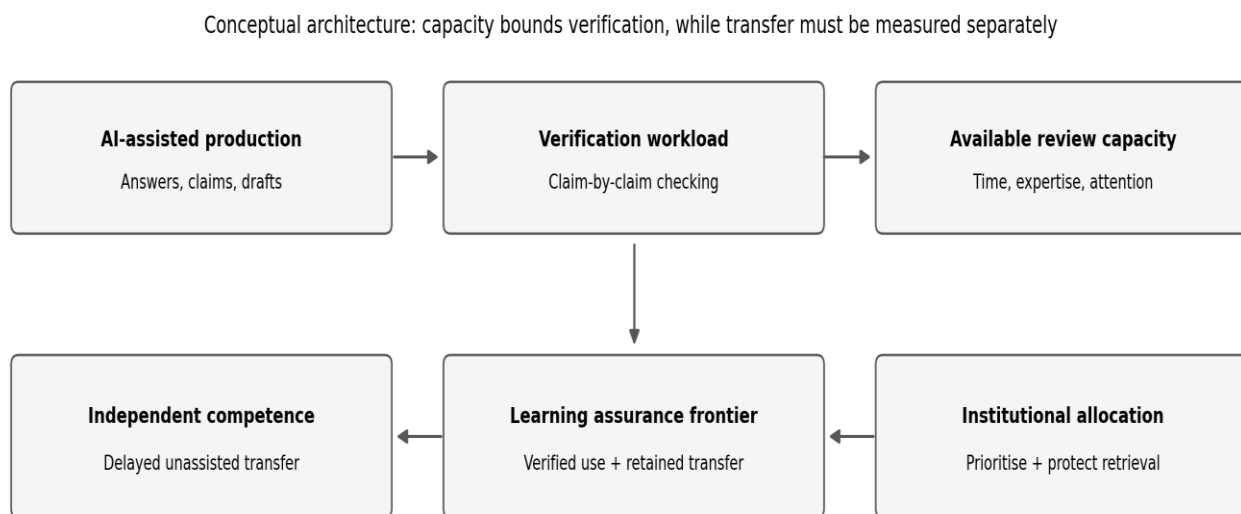


Figure 1. Conceptual structure linking AI-supported task production to finite verification capacity, independent competence, and educational allocation. Arrows indicate a proposed design architecture, not estimated causal effects.

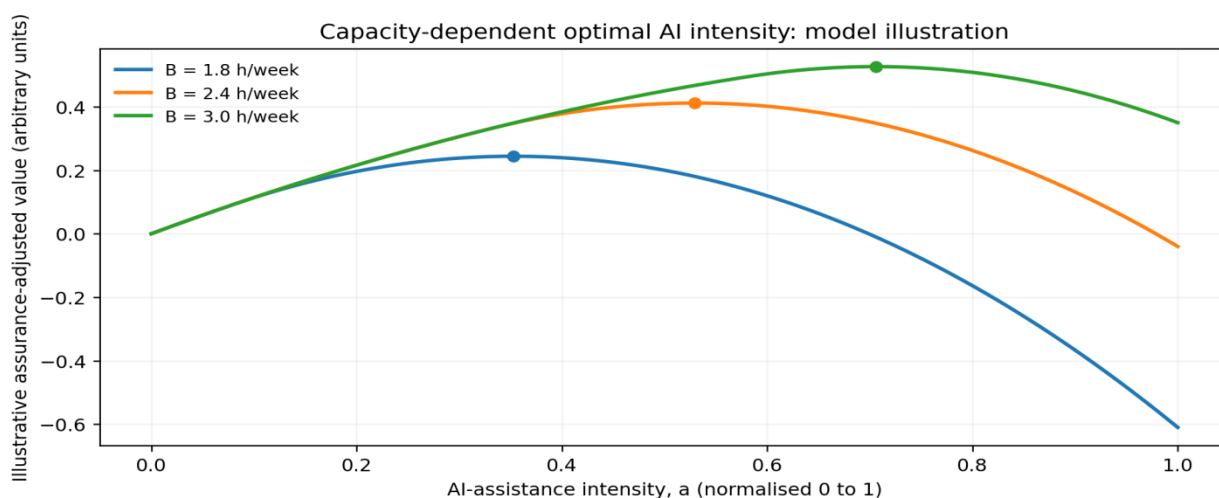


Figure 2. Illustrative assurance-adjusted educational value across AI-assistance intensity. Filled markers indicate the mathematically optimal intensity under the stated hypothetical parameter values. Curves are analytic, not measured student outcomes.

Capacity disparities despite equal tool access

Equal licences, devices, or chatbot access do not imply equal learning assurance. Students may differ in available time, familiarity with sources, background knowledge, language proficiency, disability accommodation, and access to disciplinary mentors. Consider two groups with the same $W(a)$ but with B_L less than B_H . Define $D(a, B)$ as $\max\{0, W(a) - B\}$. Their workload deficit gap is:

	$G(a) = D(a, B_L) - D(a, B_H) \geq 0.$	(14)
--	--	------

The positive-part gap is nonnegative by monotonicity. It increases with a over ranges where the lower-capacity group is overloaded but the higher-capacity group remains below its threshold, and becomes constant at B_H minus B_L once both are overloaded. It does not, on its own, prove unequal test scores. It identifies a mechanistic exposure difference that should be measured and tested. A fairness evaluation should ask whether verification tasks, access to trusted sources, teacher consultations, and time for independent retrieval are distributed according to learning need rather than assuming formal access parity is sufficient.

Sensitivity and boundary checks

The deterministic model has four useful boundary properties. First, with $\theta = 0$, the chosen optimum reduces to the clipped value r/k and no longer responds to checking capacity. Second, with B at or above $W(1)$, the verification-deficit penalty is zero throughout the domain; the frontier ceases to constrain the result. Third, greater θ lowers the interior overloaded optimum, holding other parameters fixed. Fourth, if g approaches zero, verification demand no longer responds to a , and the optimisation collapses to a concave productivity term plus a constant loss. None of these conditions requires assumed causal mediation or a fabricated learning measure.

A more realistic system can allow per-unit checking costs to rise with complexity, different error probabilities across subject domains, shared errors, teacher and student verification to overlap, and the educational benefit of checks themselves. For example, an error-correlation model would make the independent-risk simplification of Equation (9) inappropriate. The analytic frontier should therefore be treated as a transparent baseline against which such extensions can be evaluated, not as a complete description of instruction.

For direct transparency, Table 4 varies one assumed parameter at a time with all others held at the middle-budget benchmark. Larger checking cost c lowers the model-optimal assistance intensity; greater benefit coefficient r increases it; and faster claim production g generally contracts it for these inputs. The r sensitivity concerns an assumed marginal educational-benefit coefficient, not a measured change in delayed transfer. None of the values in Table 4 is an estimate from a classroom.

Table 4. One-factor-at-a-time sensitivity of the illustrative objective, with $B = 2.4$ h/week, $\lambda_0 = 8$, $k = 1.2$, $\theta = 0.5$ and other parameters at $c = 0.20$, $r = 1.2$, $g = 1.5$.

Parameter varied	Value	Model-optimal a^*	Max. checkable share	F at optimum
Check cost c (h/unit)	0.15	0.809	90.40%	0.562
Check cost c (h/unit)	0.2	0.529	83.60%	0.412
Check cost c (h/unit)	0.25	0.316	81.40%	0.244
Benefit coefficient r	0.9	0.456	89.10%	0.264
Benefit coefficient r	1.2	0.529	83.60%	0.412
Benefit coefficient r	1.5	0.603	78.80%	0.582
Claim growth g	1	0.742	86.10%	0.523



Claim growth g	1.5	0.529	83.60%	0.412
Claim growth g	2	0.392	84.00%	0.327

An error-probability sensitivity can be read directly from the allocation rule rather than by inserting p_j into F , where it does not appear. Consider one divisible claim unit in each of two strata. For A, let $h_a = 4$, $s_a = 0.75$, $c_a = 0.15$ hour; for B, let $h_b = 2$, $s_b = 0.90$, $c_b = 0.10$ hour. With $p_a = 0.10$ and $p_b = 0.15$, priority scores are $\pi_a = 2.0$ and $\pi_b = 2.7$ expected harm units per hour, so B is checked first. Raising only p_a to 0.18 changes π_a to 3.6, reversing the order. With a 0.15-hour budget and divisible strata, the first case checks all B and one third of A; the second checks all A. The exercise illustrates why error probabilities affect review allocation but do not automatically identify the assistance optimum in Equation (7).

PEDAGOGICAL AND INSTITUTIONAL IMPLICATIONS

Protect independent learning transfer as the primary outcome

An education provider should report at least four outcomes separately: performance while AI is available; accuracy and evidence quality of the work; ability to explain or challenge the work; and delayed transfer on a comparable but novel task completed without AI. The primary educational outcome should be the latter where the instructional objective is enduring competence. Immediate output quality can be valuable, particularly in professional practice, but it cannot substitute for an unaided transfer measure. This recommendation accords with the learning-performance distinction and the experimental evidence that unrestricted assistance may damage unassisted performance in particular settings (Soderstrom & Bjork, 2015; Bastani et al., 2025).

Replace blanket prohibitions with assurance-aware task design

The LAF supports differentiated assistance: ask students to generate an initial explanation, compare it with an AI solution, identify at least one weakness or assumption, verify selected high-consequence claims against primary sources, and finally complete a delayed unaided application. Where AI provides hints rather than answers, the same sequence may be shortened. Teachers should explicitly teach evidence appraisal and answer contestation, consistent with human-centred guidance for GenAI (Miao & Holmes, 2023). Cognitive forcing can reduce uncritical reliance in some decision settings, although its benefits and effort costs vary by user (Buçinca et al., 2021). Such friction is justified only when it targets the knowledge or reasoning the curriculum actually intends to teach. Low-value formatting and clerical work can often be automated without comparable educational loss.

Structured reflective dialogue provides a candidate mechanism for the independent-reasoning reserve. Tan (2026b) reports an adult-learning intervention incorporating Socratic questioning and reflective journaling; the present proposal adapts those components to AI-mediated classes as a hypothesis to evaluate, not as an intervention whose effects have been established for GenAI.

Allocate review capacity and avoid invisible inequity

Institutions can estimate expected checking workload before assigning AI-intensive activities, then set review budgets and escalation rules. High-stakes calculations, clinical statements, safety claims, and unverified bibliographic references warrant disproportionate attention when their expected harms are large. The decision rule in Equation (10) is a starting point, not a complete equity policy. It should be combined with group-level audits of review time, accessible tools, linguistic fairness, and delayed transfer. Teacher development must cover both AI literacy and domain-specific verification, rather than treating prompt engineering as sufficient professional preparation.

A practical assessment workflow

Table 5 translates the proposed principles into assessable classroom activities without claiming demonstrated intervention effects.

Table 5. Proposed educational assurance workflow and the corresponding auditable measure.

Stage	Required learner activity	Minimum auditable evidence
Pre-AI baseline	Attempt or outline reasoning without assistance	Time-stamped initial response or brief oral explanation
Scaffolded AI use	Request feedback, alternative approaches, or targeted hints	Permitted AI-use declaration and interaction purpose
Verification	Check a risk-stratified sample against sources or worked derivations	Claim ledger, source excerpts, corrected claims
Cognitive ownership	Explain core steps and dispute one AI recommendation	Short defence, counterexample, or reasoning trace
Delayed transfer	Solve novel parallel tasks without AI after a prespecified delay	Blind-scored transfer and retention outcomes
Equity audit	Review differential verification burden and support	Subgroup checking time, uptake, and outcome distributions

AI interaction logs should not be equated with the entire cognitive process. Students may use systems outside institutional platforms, and monitoring can create privacy or surveillance risks. Assurance procedures should use data minimisation, transparent consent, appropriate alternatives, and teacher judgement. The purpose is instructional diagnosis, not punitive inference from a simplistic AI-detection score.

EMPIRICAL VALIDATION RESEARCH AGENDA

Falsifiable hypotheses

H1 (capacity saturation): at comparable task difficulty and check cost, an exogenous increase in claim-output volume reduces the fully verified share only after verification demand crosses the available budget. H2 (transfer divergence): answer-first AI use produces a larger assisted-performance advantage than delayed unassisted transfer advantage, conditional on learner ability and pedagogical context. H3 (capacity moderation): additional structured verification time changes the relationship between assistance intensity and delayed transfer, with a stronger expected benefit when the baseline workload ratio exceeds one. H4 (equity): learners with lower measured verification capacity exhibit larger verification deficits under identical AI-access conditions. These hypotheses are propositions for future testing, not results asserted by this article.

Proposed study design

A confirmatory evaluation could employ a preregistered, three-arm, cluster-randomised study across comparable modules: (A) conventional instruction without GenAI, (B) unguided answer-first GenAI access, and (C) scaffolded GenAI with claim prioritisation, self-explanation, and a protected no-AI transfer phase. To identify the role of capacity separately from interface design, a follow-up factorial experiment could randomise an additional verification-time allocation within AI-using arms. Randomising by tutorial group where spillovers are plausible, and specifying contamination checks, is preferable to assuming that students do not share prompts or outputs. Intervention fidelity, model version, course content, and marking criteria should be preregistered. A qualified statistician should conduct a power analysis using pilot intraclass correlations and a declared smallest educationally important effect; the present study does not invent a sample size or power estimate.

The primary endpoint should be a blinded, delayed, unaided near-and-far transfer composite at a prespecified follow-up, for example two to four weeks after instruction. Important secondary outcomes include immediate task quality, factual checking accuracy, unsupported-claim rate, calibration between confidence and performance, reason-giving quality, learner workload, and time costs. Assessments must include tasks not simply recycled from AI prompts. Because GenAI may benefit different subgroups unequally, measure baseline domain knowledge, language background, accessibility needs, and digital skills only with appropriate consent and protection.

Estimation, threats, and disconfirmation

The intention-to-treat analysis could fit a mixed-effects model with two intervention indicators, a baseline test score, and a classroom random intercept, as indicated in Equation (15). The intervention indicators refer to arms B and C relative to conventional instruction. Additional teacher or institution levels should be added where the sampling design warrants them. Standard errors and confidence intervals should reflect the randomisation unit. Mediation by verified share requires stronger assumptions and should be labelled exploratory unless separately identified. A null or reversed capacity interaction, superior transfer under answer-first assistance, or a weaker-than-assumed demand relation would directly challenge the proposed account.

	$Y_{ij,t} = \beta_0 + \beta_1 T_{1j} + \beta_2 T_{2j} + \beta_3 Y_{ij,0} + u_j + \varepsilon_{ij,t}.$	(15)
--	---	------

Measurement validity is crucial: a completed check may be wrong; a learner can verify accurately but fail to retain; conversational fluency is not evidence of explanation skill; and confidence may be well calibrated for familiar content but not novel transfer. The proposed evaluation should therefore use independent expert scoring, a documented claim taxonomy, inter-rater reliability checks, and independent transfer tasks. Data protection, research ethics review where required, consent, and equitable accommodation must precede any actual intervention.

DISCUSSION

The central theoretical result is an allocation principle. An educational institution may scale AI-supported production faster than it scales learners' opportunity to examine, challenge, and internalise the generated knowledge. The resulting problem is not simply one of AI accuracy. A factually correct answer can still bypass the cognitive activity necessary for retention and transfer, while a partially incorrect answer may sometimes become instructionally useful if the learner has time and support to discover and correct the error. Thus a reliable AI interface, a verification procedure, and a learning design address different aspects of educational quality.

The paper also identifies a policy tension. Restricting assistance everywhere may sacrifice effective tutoring, accessibility support, language assistance, and productive feedback. Maximising assistance everywhere can shift demand toward unseen and underfunded checking work. The appropriate aim is not maximum or minimum AI use, but an explicitly justified balance of immediate support, assurance, and independent competence. The model makes that balance discussable by separating measurable work-flow constraints from value judgements encoded in the objective function.

The contribution is especially pertinent to inclusive education. Learners with limited time, weaker prior knowledge, additional language needs, or reduced access to expert support may receive the same AI tool as peers while bearing higher burdens of verification. This is a hypothesis about differential assurance exposure, not a claim that such learners cannot use AI effectively. In many circumstances a well-designed tutor may reduce those inequities. Equity conclusions should therefore be drawn from properly designed intervention studies, not from a deterministic capacity model alone.

Prior work supplies the conceptual ingredients but does not establish this model's exact conclusion. Tan (2026a) focuses on broader cognitive agency; Tan (2026d) addresses verification queues in organisational knowledge work. This article is differentiated by modelling the classroom-level trade-off between assistance intensity and educational assurance, deriving a capacity-responsive optimum, and specifying delayed unaided transfer as its validation criterion. Nevertheless, novelty is stated as the proposed synthesis and derivation, not as an exhaustive claim that no related educational formalisation exists.

The comparative table clarifies the paper's incremental rather than totalising novelty. Cognitive load and offloading theories explain mechanisms of processing and external assistance, while AI-literacy and engagement frameworks explain the quality of student interaction. The LAF adds an explicit and contestable link between the volume of checkable AI-mediated claims, a finite assurance budget, a model-optimal assistance level, and the delayed unaided outcome that would make that optimum educationally meaningful. Its mathematical optimum should not be interpreted as measured transfer, and the sensitivity exercise establishes model behaviour only under the declared illustrative inputs.

LIMITATIONS

The principal limitation is empirical identification: no field dataset validates the quadratic educational value function, the linear workload growth equation, workload-variance assumptions, or the parameter values. Theoretical concavity is a property of the assumed function, not a discovered universal law of educational psychology. A second limitation is aggregation: both a and B compress heterogeneous tasks, student abilities, teacher support, sources, and error severities into scalar proxies. Third, v is a capacity bound, not a direct assurance probability, since verification competence and error detection vary. Fourth, the priority allocation theorem assumes linear separability and divisible claims. Fifth, the framework does not model strategic behaviour, plagiarism incentives, network spillovers, tutor adaptation, or long-term dynamic skill decay. Finally, unassisted transfer is crucial but not the sole legitimate learning outcome; collaborative practice and supported workplace performance may be educational ends in their own right.

These limits motivate rather than invalidate the proposed empirical programme. A stronger future specification should jointly estimate verification time, error detection, prior knowledge, AI-use patterns, assisted scores, and delayed outcomes. Researchers should compare linear and nonlinear arrival models, test robust alternatives to the quadratic objective, model teacher resource constraints and overload uncertainty explicitly, and examine whether problem type or learner developmental stage changes the frontier's shape.

CONCLUSION

The Learning Assurance Frontier provides a formal answer to a practical dilemma in digital learning innovation: the production of good-looking work can expand faster than learners and teachers can verify and internalise the knowledge it appears to demonstrate. Under clearly stated assumptions, finite capacity imposes a checking boundary; a concave assurance-adjusted objective yields a unique capacity-dependent assistance optimum; uneven verification resources create unequal exposure to unchecked work; and workload uncertainty introduces an explicit, distribution-robust safety margin. These are model-conditional deductions, not measured effects. For educational practice, the actionable principle is to assess independent transfer, allocate checking according to consequence and learning need, and preserve productive opportunities to reason without AI. The next decisive step is prospective classroom validation with preregistered outcomes, ethically collected process measures, and transparent model comparison.

DECLARATIONS

Funding: No external funding was used in preparing this theoretical manuscript, subject to author confirmation.

Competing interests: The author declares no competing interests in relation to the analysis presented, subject to author confirmation.

Ethics approval and consent: Not applicable to the present theoretical and computational study, which involved no human participants, identifiable records, or experimental intervention. Any future human-subject study described here would require the applicable institutional review before recruitment.

Data and code availability: No empirical data were generated or analysed. All illustrative inputs, analytic equations, and tabulated outputs are disclosed in Sections 3 to 5 and Appendix B. The numerical figure is generated deterministically from those inputs.

Appendix A. Derivation and validity conditions

A1. Global concavity and optimum

The positive-part squared function is convex in its scalar argument. Since $W(a)-B$ is affine in a , subtracting a nonnegative multiple of its square leaves a concave contribution. The term $-(k/2)a^2$ is strictly concave for $k > 0$; hence F is strictly concave across the joined regimes. At $W(a) = B$, the first derivatives of both branches coincide because the deficit term and its first derivative both vanish. Equation (6) follows by setting the branchwise first derivative equal to zero. As F is strictly concave, a stationary point, when present, is unique, and its projection to the closed convex interval $[0,1]$ yields Equation (7).

A2. Resource-allocation exchange argument

For any two strata j and l with π_j greater than π_l , an allocation that funds a small positive amount of checking in l while not fully funding j can be improved by shifting an equal amount of checking time to j . If epsilon hours are moved, the change in expected avoided harm equals epsilon times $(\pi_j \text{ minus } \pi_l)$, which is positive. Iteration establishes descending-priority selection. This result holds for divisible units and a fixed budget; it should not be extended without proof to integer tasks, endogenous error dependence, or nonseparable utility.

A3. Proof and interpretation of the distributionally robust margin

Write $Z = W - \mu W$, so that $E[Z] = 0$ and $E[Z^2] = \sigma W^2$. For $t > 0$ and an arbitrary $h \geq 0$, the event $\{Z \geq t\}$ implies $(Z + h)^2 \geq (t + h)^2$. Markov inequality therefore yields $P(Z \geq t) \leq (\sigma W^2 + h^2)/(t + h)^2$. Minimising this bound at $h = \sigma W^2/t$ gives $P(Z \geq t) \leq \sigma W^2/(\sigma W^2 + t^2)$. If $B > \mu W$, setting $t = B - \mu W$ yields the sufficient safety margin in Equation (12). If $\sigma W = 0$, the probability constraint holds exactly whenever $\mu W \leq B$; if the variance is infinite or the mean and variance are incorrectly specified, the asserted distributionally robust implication is unavailable. The strict or weak overload event can be handled by a limiting argument at the threshold. The probability statement describes workload feasibility, not substantive learning validity.

Appendix B. Reproducibility of illustrative scenarios

Inputs: $\lambda_0 = 8$ checkable units/week; $g = 1.5$; $c = 0.20$ hour/unit; $r = 1.2$; $k = 1.2$; $\theta = 0.5$; and B in $\{1.8, 2.4, 3.0\}$ hours/week. For each budget, calculate $d = 1.6 - B$, $b = 2.4$, derive unconstrained assistance value using Equation (6), clip to $[0,1]$, evaluate $W(a^*)$, $v(a^*, B)$, $F(a^*; B)$, and $F(1; B)$. Generate Figure 2 by evaluating F at 300 evenly spaced points spanning $a = 0$ to $a = 1$ for each B ; plot the three curves and mark the calculated maxima. The calculations are deterministic, so no random seed is required.

Illustrative internal validity checks: (i) every optimal a lies in $[0,1]$; (ii) $F(a^*)$ is at least as large as $F(0)$ and $F(1)$; (iii) capacity thresholds from Equation (3) are ordered by B ; (iv) the derivative is negative to the right of an interior optimum and positive immediately to its left; (v) increasing B does not reduce the computed optimum under the specified model. These checks are algebraic and numerical quality controls, not statistical validation.

Sensitivity reproducibility: hold $B = 2.4$, $\lambda_0 = 8$, $k = 1.2$, $\theta = 0.5$ and vary, one at a time, $c \in \{0.15, 0.20, 0.25\}$, $r \in \{0.9, 1.2, 1.5\}$, and $g \in \{1.0, 1.5, 2.0\}$; reset all non-varied coefficients to their Section 5.1 values. Evaluate Equations (1), (2), (4), (6), and (7) without rounding intermediate calculations. Compute π_j directly from Equation (10) for the two-stratum probability example in Section 5.3. Rounded outputs appear in Table 4. These calculations can be reproduced with spreadsheet formulas or fewer than ten lines of numerical code.

REFERENCES

1. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, 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>

2. Bućinca, 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), Article 188. <https://doi.org/10.1145/3449287>
3. Chen, S., & Cheung, A. C. K. (2025). Effect of generative artificial intelligence on university students learning outcomes: A systematic review and meta-analysis. *Educational Research Review*, 49, Article 100737. <https://doi.org/10.1016/j.edurev.2025.100737>
4. Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>
5. Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4–58. <https://doi.org/10.1177/1529100612453266>
6. Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, 38(5), 1008–1022. <https://doi.org/10.1111/cogs.12107>
7. 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>
8. Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3313831.3376727>
9. Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
10. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192. <https://doi.org/10.1126/science.adh2586>
11. OECD. (2026). *OECD Digital Education Outlook 2026: Exploring effective uses of generative AI in education*. OECD Publishing. <https://doi.org/10.1787/062a7394-en>
12. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
13. Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
14. 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>
15. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
16. Tan, K. H. (2026a). 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>
17. Tan, K. H. (2026b). Improving critical thinking skills among adult learners through structured reflective dialogue: An action research study in a university social sciences course in Singapore. *International Journal of Research Publication and Reviews*, 7(6), 5322–5329. <https://doi.org/10.55248/gengpi.07.0626.17a03>
18. Tan, K. H. (2026c). The synthetic consensus trap: Correlated AI errors, verification overload, and the mathematics of institutional epistemic cascades. *ResearchGate*. <https://doi.org/10.5281/zenodo.22009669>
19. Tan, K. H. (2026d). The verification bottleneck in AI-accelerated knowledge work: Queueing epistemics, verification capacity, and adaptive human oversight. *Open Access Journal of Multidisciplinary Research*, 2(9), 1–17. <https://doi.org/10.47760/OAJMR.2026.v02i09.001>