ICoTESH
International Conference on Educational Technology and Social Humanities
OPEN ACCESS

Trust in AI, Cognitive Offloading and Learning Outcomes in Higher Education: A Moderated Mediation Model of Critical AI Literacy

A Pilot Study

Guo Qiang Tan
Golden Gate University Worldwide, San Francisco, California , Amerika Serikat
Guoqiang@collaborativeconsultancy.com

ABSTRACT

Purpose – This study develops and pilots a moderated mediation model of how university students' trust in AI-generated knowledge relates to learning outcomes through cognitive offloading, and how critical AI literacy conditions this indirect process. **Design/methodology/approach –** Survey data from students who use generative AI for academic work (N = 27) were analysed using partial least squares structural equation modelling (Smart-PLS 4), with the interaction estimated via the two-stage approach and significance assessed by bootstrapping. **Findings –** Trust positively predicted cognitive offloading. Contrary to expectation, offloading positively predicted self-reported learning; combined with a negative direct trust–learning path, this produced competitive mediation. Critical AI literacy was the strongest predictor of learning and showed a directionally consistent but non-significant buffering of the offloading–learning path, so moderated mediation was supported in direction only. **Research limitations/implications –** The small pilot sample underpowers the interaction; estimates are preliminary and require confirmation with roughly 160 cases and objective learning measures. **Practical implications –** Critical AI literacy merits treatment as a core curricular competency rather than an optional add-on. **Originality/value –** The paper reframes the AI-in-learning debate as a conditional process with a specified mechanism and boundary condition and validates a new instrument for confirmatory testing.

Keywords Trust in AI, Cognitive offloading, Critical AI literacy, Self-regulated learning, Higher education, PLS-SEM, Moderated mediation

INTRODUCTION

The diffusion of generative AI into higher education has shifted debate from tool adoption to more fundamental questions about what students learn and what cognitive work is transferred to the machine when an AI can draft an essay or solve a problem in seconds. This study treats that concern as an empirical, psychological question. Two dynamics sit at its centre. Trust governs how far students rely on AI-generated knowledge whose provenance is often opaque; cognitive offloading is the tendency to use external tools to reduce mental effort. Offloading can free resources for higher-order work, but over-reliance may displace the effortful, generative processing that underpins understanding and retention. We argue that whether trust translates into beneficial or detrimental learning depends on a learner competency: critical AI literacy.

Accordingly, the paper pilots a moderated mediation model in which trust relates to learning outcomes indirectly through cognitive offloading, with critical AI literacy moderating the offloading–learning relationship. As a pilot, it has three aims: to assess the psychometric adequacy of a newly

assembled instrument, to obtain preliminary estimates of the hypothesised relationships, and to generate effect-size and power information for an adequately powered confirmatory study. The contribution is twofold. Theoretically, the model reframes the AI-in-learning debate as a conditional process with a specified mechanism (offloading) and boundary condition (literacy). Empirically, the pilot reveals a more complex pattern than a simple "offloading harms learning" narrative would predict. Because the sample is small, all estimates are preliminary and reported with their inferential statistics and the sample sizes needed to confirm them.

Theoretical Background And Hypotheses

Trust is a willingness to be vulnerable based on positive expectations of another party's competence, reliability and integrity (Mayer et al., 1995); applied to AI, it concerns the degree to which a student believes AI-generated content is accurate and usable. Higher trust lowers the perceived need to verify or generate content independently, making reliance the path of least resistance, so trust is positioned as the antecedent that sets offloading in motion. Cognitive offloading reduces the cognitive demands of a task by delegating recall, reasoning or composition to an external system (Risko and Gilbert, 2016). It is not inherently harmful, but learning depends on effortful, generative processing: when students offload the operations a task is designed to develop, they may achieve the product without acquiring the competence (Grinschgl et al., 2021; Sparrow et al., 2011). Offloading therefore serves as the mediating mechanism linking trust to learning.

Learning outcomes are conceptualised as self-regulated, durable learning – perceived depth of understanding, transfer, retention and capacity for independent learning (Biggs et al., 2001; Zimmerman, 2002) – deliberately distinguishing learning from task completion. Critical AI literacy is the capacity to understand, evaluate, verify and critically interrogate AI systems and their outputs (Carolus et al., 2023; Long and Magerko, 2020; Wang et al., 2023). When literacy is high, students offload selectively and verify what they receive, so the same level of offloading is less damaging; when low, offloading more readily substitutes for genuine learning. This yields a conditional indirect effect whose strength depends on the learner's literacy. The hypotheses are: H1, trust is positively related to offloading; H2, offloading is negatively related to learning; H3, offloading mediates the trust–learning relationship; H4, critical AI literacy moderates the offloading–learning relationship, weakening it when literacy is high; and H5, the indirect effect of trust on learning through offloading is moderated by critical AI literacy (moderated mediation). Figure 1 presents the model.

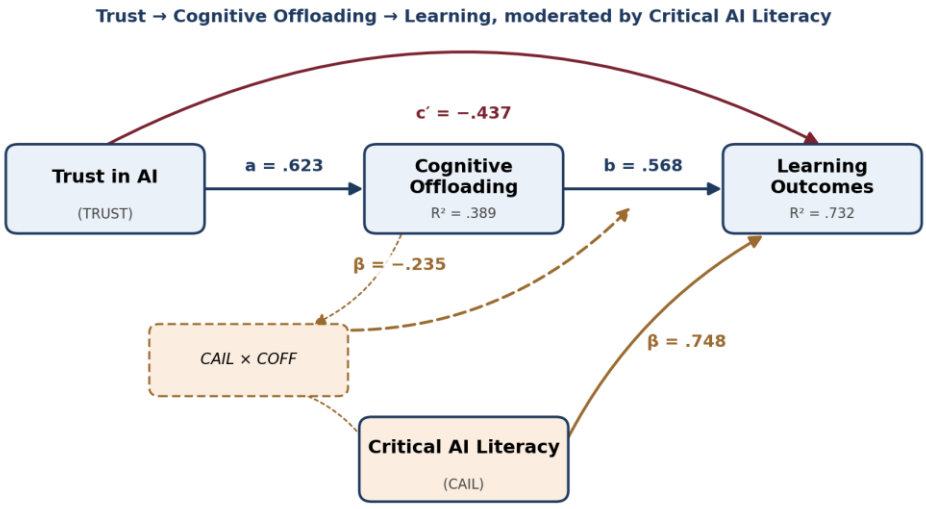

Figure 1. *Estimated moderated mediation model (standardised coefficients). Paths a and b form the indirect effect (H3); critical AI literacy moderates path b via CAIL × COFF (H4); the conditional indirect effect is summarised by the index of moderated mediation (H5). R² values are shown within the endogenous constructs.*

METHOD

The study adopts a quantitative, cross-sectional, survey-based pilot design grounded in a post-positivist philosophy. Because the aim is to predict a target construct and to test a model with a mediator, a moderator and their interaction, variance-based structural equation modelling (PLS-SEM) was used (Hair et al., 2022). Participants were university students who reported using generative AI for academic work, recruited through purposive sampling across disciplines and year levels. After screening and attention-check filtering, the pilot sample comprised 27 complete cases. An a priori power analysis (inverse-square-root method; Kock and Hadaya, 2018) indicates that detecting the interaction at 80% power and $\alpha = 0.05$ requires approximately 112 cases (155–160 for 90% power), whereas the main paths require only 12–33; the achieved N is therefore adequate for the main paths but not the interaction.

All four constructs were modelled reflectively using multi-item, seven-point Likert scales adapted to the AI-in-learning context; the full instrument is provided as supplementary material. Following measurement-model purification, the retained sets were four critical-AI-literacy indicators, three offloading indicators, three learning indicators and five trust indicators, and the moderation term was formed as their product. The model was estimated in SmartPLS 4 (Ringle et al., 2024) with the two-stage interaction approach (Becker et al., 2018). Significance for path coefficients, the specific indirect effect and the conditional effects was assessed by bootstrapping (500 subsamples, one-tailed, $\alpha = 0.05$, fixed seed). Because BCa intervals were numerically unstable at this N, bootstrap t-statistics with one-tailed p-values and stable percentile 90% confidence intervals are reported as the primary inferential statistics; one-tailed borderline effects should be read cautiously. A confirmatory study should use $\geq$ 5,000 subsamples and BCa intervals.

RESULTS AND DISCUSSION

Results

Construct means clustered around the mid-to-upper scale: learning items were highest (5.44–5.78), critical AI literacy moderately high (5.00–5.56), and trust and offloading more moderate (3.85–4.56), indicating cautious rather than wholesale reliance. The reflective measurement model was satisfactory (Table I). With one exception (CAIL1 at 0.526, retained on content grounds), loadings exceeded 0.708 and all retained loadings were significant ($t = 1.77–38.17$). Internal consistency was adequate ($\alpha = 0.724–0.896$; $\rho c = 0.815–0.934$) and convergent validity supported (AVE = 0.532–0.826). Discriminant validity held on both the HTMT ratio (range 0.224–0.761, all below 0.85) and the Fornell–Larcker criterion. Three trust indicators showed outer VIF of 4.27–4.76 (above 3.3 but below 5).

Table I. *Construct reliability and validity (retained items)*

Construct (code)	Items	Loading range	α	ρc	AVE
Critical AI Literacy (CAIL)	4	0.526–0.888	0.724	0.815	0.532
Cognitive Offloading (COFF)	3	0.765–0.860	0.745	0.848	0.652
Learning Outcomes (LEARN)	3	0.852–0.943	0.896	0.934	0.826
Trust in AI (TRUST)	5	0.723–0.923	0.896	0.922	0.705

Note: α = Cronbach's alpha; ρc = composite reliability; AVE = average variance extracted. All AVE ≥ 0.50; all retained loadings significant at $p < 0.05$ (one-tailed bootstrap, 500 subsamples).

Inner VIF values (1.000–2.083) indicated no lateral collinearity. The model explained substantial variance in learning ($R^2 = 0.732$) and a moderate proportion in offloading ($R^2 = 0.389$). Table II reports the structural paths. Supporting H1, trust positively predicted offloading ($\beta = 0.623$, $p = 0.003$, large effect, $f^2 = 0.636$). Contrary to H2, offloading positively – not negatively – predicted learning ($\beta = 0.568$, $p = 0.010$, $f^2 = 0.632$). Critical AI literacy was the strongest path ($\beta = 0.748$, $p < 0.001$, $f^2 = 1.371$), and the direct trust–learning path was negative and only marginally significant ($\beta = −0.437$, $p = 0.048$, 90% CI included zero).

Table II. *Structural model: path coefficients and bootstrap significance (500 subsamples, one-tailed)*

Path	β	t	p	90% CI [LL, UL]	Decision
H1 TRUST → COFF (a)	0.623	2.79	0.003	[0.470, 0.848]	Sig.
COFF → LEARN (b)	0.568	2.33	0.010	[−0.003, 0.766]	Sig.
TRUST → LEARN (c′)	−0.437	1.66	0.048	[−0.748, 0.080]	Marginal
CAIL → LEARN	0.748	5.32	<0.001	[0.570, 1.010]	Sig.

Path	β	t	p	90% CI [LL, UL]	Decision
H4 CAIL × COFF → LEARN	−0.235	1.33	0.092	[−0.348, 0.136]	n.s.

Note: β = standardised path coefficient (original sample). One-tailed at α = 0.05. "Marginal" denotes p just below 0.05 whose 90% CI includes zero (non-significant two-tailed).

The specific indirect effect of trust on learning through offloading was positive and significant (0.354, t = 2.03, p = 0.021), though its 90% CI marginally included zero. Because the direct path was negative while the indirect path was positive, the two offset one another, yielding a small non-significant total effect (−0.083, p = 0.327). This is competitive (inconsistent) mediation: trust relates to learning through a positive route via offloading and a negative route independent of it, so H3 is supported in magnitude but opposite in direction. The CAIL × COFF interaction was negative and directionally consistent with buffering but non-significant (β = −0.235, p = 0.092). As Table III shows, the offloading–learning slope and the conditional indirect effect both weakened as literacy rose, remaining significant at low and mean literacy but not at high literacy; the index of moderated mediation was −0.147. Thus, H4 and H5 are supported in direction only, and – as the a priori analysis anticipated – were underpowered at N = 27. Approximate fit (SRMR = 0.159) exceeded the 0.08 guideline and is reported only for completeness, given the composite-based, prediction-oriented model and small sample.

Table III. *Conditional indirect effects and simple slopes at levels of critical AI literacy*

Level of CAIL	Indirect effect (a×b)	p	Slope (COFF→LEARN)	p
Low (−1 SD)	0.501	0.021	0.803	0.013
Mean	0.354	0.021	0.568	0.010
High (+1 SD)	0.207	0.120	0.333	0.074
Index of moderated mediation	−0.147	—	—	—

Note: Bootstrap (500 subsamples, one-tailed). The index equals a (0.623) × the interaction coefficient (−0.235). The indirect effect is significant at low and mean literacy but not at high literacy, consistent with H5 in direction.

Discussion

Three findings stand out. Trust significantly predicted offloading (H1): students who regard AI output as dependable delegate more cognitive work to it. Contrary to expectation, offloading positively predicted self-reported learning (H2 reversed). And critical AI literacy was by far the strongest predictor of learning, with a directionally consistent but non-significant buffering interaction in which the conditional indirect effect weakened from significant at low and average literacy to non-significant at high literacy (H4, H5 in direction). The interaction effects did not reach significance, exactly as the power analysis predicted, so they are promising leads rather than established effects.

The positive offloading–learning association merits careful interpretation. It is inconsistent with accounts in which delegating effortful processing necessarily undermines learning (Grinschgl et al., 2021; Sparrow et al., 2011), but reconcilable with two explanations. Methodologically, learning was measured by self-report rather than an objective test, so students who lean on AI may feel more capable even if

tested learning differs – a metacognitive illusion of competence the design cannot rule out. Substantively, in a sample reporting only moderate offloading and trust, AI may have been used as a scaffold (clarifying, exemplifying, checking) rather than a wholesale substitute, in which case offloading could plausibly support perceived learning. The competitive mediation pattern fits this dual role, and decomposing the opposing routes is itself a contribution: a focus on the near-zero total effect alone would have masked two countervailing processes.

The result sits within a fast-accumulating and inconsistent literature. Gerlich (2025) found that frequent AI use was associated with greater offloading and weaker critical thinking, and Fan et al. (2025) reported "metacognitive laziness" when learners delegated regulatory work to ChatGPT – both consistent with a negative path. Conversely, Iqbal et al. (2025) modelled offloading as a positive conduit between generative-AI use and achievement, especially with shared metacognition, mirroring the present finding. The coexistence of these results suggests the sign of the offloading–learning relationship is contingent on how, and by whom, AI is used – precisely the premise of the moderated model. The self-report caveat also gains support: Fernandes et al. (2025) found that AI assistance improved objective performance, yet inflated participants' estimates of it, so a positive coefficient on perceived learning may partly reflect inflated confidence, sharpening the case for objective measures in the confirmatory study.

Critical AI literacy is the most informative result. Its large positive main effect indicates that, regardless of offloading, students better able to evaluate and verify AI report stronger learning. Its negative (non-significant) interaction suggests literacy also changes how offloading relates to learning: the slope was steepest for low-literacy students (0.803) and shallowest for high-literacy students (0.333). One reading is that less literate students depend on offloading for whatever gains they report, while more literate students draw on independent competencies that decouple learning from reliance. This is consistent with framing literacy as a protective, self-regulatory competency, and with converging moderation research: Tian and Zhang (2025) found information literacy buffered the effect of AI dependence on critical thinking, while Fernandes et al. (2025) found that mere technical familiarity did not improve calibration – implying it is the critical, evaluative facet of literacy that does the protective work. Taken together, the pilot reframes the debate as conditional rather than uniform and offers a candidate reconciliation of contradictory findings: offloading may erode or enhance learning depending on the learner's evaluative competence.

Preliminary Implications

Implications are offered tentatively. Theoretically, the model integrates trust theory, the offloading literature and self-regulated learning under one testable structure and demonstrates the value of a moderated-mediation lens; the competitive mediation also cautions against inferring "no effect" from a small total effect without decomposing direct and indirect routes. Practically, the prominence of critical AI literacy – the most reliably significant predictor of learning – offers preliminary support for treating it as a core curricular competency rather than an optional add-on. If the buffering pattern is confirmed, instructional design should cultivate evaluation, verification and questioning skills alongside subject content, channelling trust toward selective, verified reliance rather than wholesale delegation – acted on only after confirmation in an adequately powered study.

Limitations And The Confirmatory Study

As a pilot, the study is bounded by design. The sample (N = 27) is small: main paths were adequately powered, but the interaction and moderated-mediation effects were not, and the confirmatory study should target roughly 160 valid cases (recruiting 200–250, or more if the interaction shrinks on replication). Significance rested on 500 one-tailed bootstrap subsamples with percentile intervals because BCa intervals were unstable; the confirmatory study should use ≥ 5,000 subsamples and two-tailed tests for non-directional effects. SRMR exceeded the 0.08 guideline, reinforcing caution. Learning was measured by perceived rather than tested outcomes and all constructs were self-reported, so the positive offloading–learning association may partly reflect a metacognitive illusion of competence; objective and behavioural measures and common-method remedies are needed. The cross-sectional design precludes causal inference, motivating longitudinal and experimental work. Finally, one literacy indicator loaded below threshold and the offloading scale lacks an established benchmark, so scale refinement is warranted, and future work might add motivation, task type or disciplinary context as further moderators.

CONCLUSION

This pilot proposed and preliminarily tested a moderated mediation model linking trust in AI-generated knowledge to learning outcomes through cognitive offloading, conditioned by critical AI literacy. Trust predicted offloading; offloading positively predicted perceived learning; and critical AI literacy was the strongest predictor of learning, with a directionally consistent but non-significant buffering of the offloading–learning pathway (index of moderated mediation = −0.147). The instrument performed well and the analysis surfaced a competitive-mediation pattern in which a near-zero total effect masked two opposing routes. Because the interaction effects were underpowered at N = 27, the pilot's contribution is to validate the measures, generate preliminary evidence and specify the design – a target of roughly 160 valid cases – for confirmation. The key message is not that students use AI, but that critical AI literacy may condition whether that use coincides with learning.

Funding And Conflict Of Interest

The author(s) received no specific external funding for this research and declare no conflicts of interest.

REFERENCES

Becker, J.-M., Ringle, C.M. and Sarstedt, M. (2018), "Estimating moderating effects in PLS-SEM and PLSc-SEM: interaction term generation and effect size assessment", *Journal of Applied Structural Equation Modeling*, Vol. 2 No. 2, pp.1-21.

Biggs, J., Kember, D. and Leung, D.Y.P. (2001), "The revised two-factor Study Process Questionnaire: R-SPQ-2F", *British Journal of Educational Psychology*, Vol. 71 No. 1, pp.133-149.

Carolus, A., Koch, M.J., Straka, S., Latoschik, M.E. and Wienrich, C. (2023), "MAILS – meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies", *Computers in Human Behavior: Artificial Humans*, Vol. 1 No. 2, 100014.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X. and Gašević, D. (2025), "Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance", *British Journal of Educational Technology*, Vol. 56 No. 2, pp.489-530.

Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C. and Welsch, R. (2025), "AI makes you smarter but none the wiser: the disconnect between performance and metacognition", *Computers in Human Behavior*, 108779, doi: 10.1016/j.chb.2025.108779.

Gerlich, M. (2025), "AI tools in society: impacts on cognitive offloading and the future of critical thinking", *Societies*, Vol. 15 No. 1, 6.

Grinschgl, S., Papenmeier, F. and Meyerhoff, H.S. (2021), "Consequences of cognitive offloading: boosting performance but diminishing memory", *Quarterly Journal of Experimental Psychology*, Vol. 74 No. 9, pp.1477-1496.

Hair, J.F., Hult, G.T.M., Ringle, C.M. and Sarstedt, M. (2022), *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 3rd ed., Sage, Thousand Oaks, CA.

Iqbal, J., Hashmi, Z.F., Asghar, M.Z. and Abid, M.N. (2025), "Generative AI tool use enhances academic achievement in sustainable education through shared metacognition and cognitive offloading among preservice teachers", *Scientific Reports*, Vol. 15, doi: 10.1038/s41598-025-01676-x.

Kock, N. and Hadaya, P. (2018), "Minimum sample size estimation in PLS-SEM: the inverse square root and gamma-exponential methods", *Information Systems Journal*, Vol. 28 No. 1, pp.227-261.

Long, D. and Magerko, B. (2020), "What is AI literacy? Competencies and design considerations", *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, pp.1-16.

Mayer, R.C., Davis, J.H. and Schoorman, F.D. (1995), "An integrative model of organizational trust", *Academy of Management Review*, Vol. 20 No. 3, pp.709-734.

Ringle, C.M., Wende, S. and Becker, J.-M. (2024), *SmartPLS 4*, SmartPLS, Bönningstedt, available at: https://www.smartpls.com (accessed 1 June 2025).

Risko, E.F. and Gilbert, S.J. (2016), "Cognitive offloading", *Trends in Cognitive Sciences*, Vol. 20 No. 9, pp.676-688.

Sparrow, B., Liu, J. and Wegner, D.M. (2011), "Google effects on memory: cognitive consequences of having information at our fingertips", *Science*, Vol. 333 No. 6043, pp.776-778.

Tian, J. and Zhang, R. (2025), "Learners' AI dependence and critical thinking: the psychological mechanism of fatigue and the social buffering role of AI literacy", *Acta Psychologica*, Vol. 260, 105725.

Wang, B., Rau, P.-L.P. and Yuan, T. (2023), "Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale", *Behaviour & Information Technology*, Vol. 42 No. 9, pp.1324-1337.

Zimmerman, B.J. (2002), "Becoming a self-regulated learner: an overview", *Theory Into Practice*, Vol. 41 No. 2, pp.64-70.