The cognitive offloading trade-off in generative ai-assisted mathematics learning: a model of optimal reliance

ALIKHANOV OLIMJON

Nagangan State University, Namangan, Uzbekistan
e-mail: webmasteruzpre@gmail.com

Abstract. *Generative AI raises measured performance while a student solves problems and may simultaneously reduce what is retained, because the cognitive work that produces durable competence is the work being offloaded. The policy question is therefore not whether to permit AI but how much reliance to permit. This paper models reliance as a continuous control variable entering a differential equation of competence formation through two opposing channels: a linear gain in engagement and coverage, and a convex penalty on retention. The model has an interior optimum at $a^* \approx 0.38$, raising unaided competence from 56.8 % to 61.1 %; reliance beyond $a \approx 0.70$ is worse than using no AI at all, and unrestricted reliance lowers it by 17.7 %. Calibrated independently at two points to a published field experiment, the model also quantifies an assessment illusion: at full reliance, coursework scores exceed unaided competence by 37 percentage points.*

Keywords: *generative artificial intelligence, cognitive offloading, desirable difficulties, retention, optimal reliance, higher mathematics, didactic modelling, assessment validity, effect size.*

1. Introduction

Generative AI has entered student problem-solving faster than institutions could form a position on it, and the debate has settled into two camps: prohibit the tools to protect learning, or embrace them because professional practice already has. Both treat reliance as binary, which is precisely what the evidence suggests it is not.

The empirical picture is genuinely two-sided. In a field experiment with nearly a thousand secondary mathematics students, access to an unmodified GPT-4 interface improved performance during assisted practice by 48 %, yet the same students scored 17 % lower than controls on a subsequent unassisted examination; a version of the same tool with pedagogical safeguards removed the harm [2]. In another controlled trial, a purpose-built tutor designed to withhold answers and elicit reasoning produced roughly double the learning gains of an excellent active-learning class [3]. The two results are usually cited against each other, although they are consistent: what differs is how much of the student's own cognitive work the system absorbs.

The mechanism is well characterised. Cognitive offloading — delegating mental work to an external resource — reliably improves immediate performance while reducing what is encoded [4], a pattern documented for search engines long before generative models [5]. Work on retrieval

practice and desirable difficulties shows that effortful generation, not exposure to correct solutions, produces durable competence [6, 7]. Cognitive load theory explains why an intermediate level of assistance should nonetheless be optimal: removing extraneous load helps, removing germane load does not [8].

What is missing is a quantitative treatment. If some assistance helps and heavy assistance harms, an optimum exists, and the useful questions are where it lies, how wide the safe interval around it is, and how far current practice sits from it — none of which a binary model can answer.

The aim of the study is to model AI reliance as a continuous control variable in a differential equation of competence formation, to derive the reliance that maximises retained competence and the threshold beyond which AI use becomes net-harmful, and to quantify the divergence between assisted and unaided performance that follows.

2. Materials and methods

2.1. The control variable. Let $a \in [0; 1]$ denote AI reliance, operationalised as the share of practice problems for which the student obtains substantive generative-AI assistance. The limiting cases are an unaided course ($a = 0$) and one in which the student consults the model on every problem ($a = 1$).

2.2. The two channels. Reliance acts on learning in opposite directions. It raises engagement and coverage, because a student who would otherwise abandon a problem continues and completes more of the syllabus; this is modelled as a linear gain $u(a) = u_0 + \Delta u \cdot a$. It also displaces the generation and retrieval that consolidate knowledge, which is represented as an increase in the decay coefficient. The penalty is taken to be strongly convex,

$$\beta(a) = \beta_0 \cdot (1 + \kappa a^3), \quad (1)$$

on the grounds that occasional recourse to AI when genuinely stuck removes little retrieval practice, whereas systematic substitution removes nearly all of it. The cubic form is the simplest that encodes this asymmetry.

2.3. Competence dynamics. Competence $y(t) \in [0; 1]$ evolves as

$$dy/dt = \alpha u(a)(K - y) - \beta(a)y, \quad y(0) = 0, \quad (2)$$

whose solution and steady state are

$$y(t) = (\alpha u(a)K/\lambda)(1 - e^{-\lambda t}), \quad y\infty(a) = \alpha u(a)K/(\alpha u(a) + \beta(a)), \quad \lambda = \alpha u(a) + \beta(a). \quad (3)$$

2.4. The optimum. Maximising $y\infty$ is equivalent to minimising $\beta(a)/\alpha u(a)$, so the optimal reliance satisfies

$$2\kappa\Delta u \cdot a^3 + 3\kappa u_0 \cdot a^2 - \Delta u = 0, \quad (4)$$

a cubic with a single root in [0; 1]. The break-even reliance a_{be}, beyond which AI use is worse than none, is the non-trivial solution of $y(T, a) = y(T, 0)$.

2.5. The assessment illusion. Coursework is completed with AI available, so the mark it produces is not a measure of unaided competence. If the model supplies a correct solution with probability η for problems the student could not solve alone, the expected assisted score is

$$S(a) = y + a\eta(1 - y), \quad (5)$$

while the unaided examination measures y itself. The difference $G(a) = S(a) - y(a)$ is the illusion gap.

2.6. Parameters and calibration. The values $\alpha = 0.20$, $\beta_0 = 0.06$, $K = 1$, $u_0 = 0.50$ and $T = 15$ weeks are carried over from the conventional-instruction baseline. The two AI-specific parameters were calibrated against the field experiment reported in [2]: $\Delta u = 0.30$ and $\kappa = 2.0$ reproduce the observed 17 % relative deficit on the unassisted examination at $a = 1$, and $\eta = 0.70$ independently reproduces the observed 48 % improvement in assisted practice

performance. Effectiveness was assessed by Cohen's $d = \Delta/\sigma$ with $\sigma = 0.15$ and by the sample size required at $\alpha = 0.05$ with power 0.80 [10].

3. Results

Table 1 and Figure 1a give the shape of the solution. Retained competence is a concave function of reliance with an interior maximum: moderate AI use helps, heavy AI use harms.

Table 1. Model quantities as a function of AI reliance

AI reliance a	$u(a)$	$\beta(a)$	y^{∞}, %	$y(T)$, %	Assisted, %
0.0 (no AI)	0.500	0.060	62.5	56.8	56.8
0.2	0.560	0.061	64.8	59.9	65.5
0.38 (optimum)	0.614	0.067	64.8	61.1	71.4
0.6	0.680	0.086	61.3	59.1	76.3
0.8	0.740	0.121	54.9	54.0	79.7
1.0 (unrestricted)	0.800	0.180	47.1	46.8	84.0

Note: computed from (1)–(5); "Assisted" is the coursework score S(a) from (5).

The cubic (4) gives an optimal reliance of $a^* = 0.38$ over a fifteen-week horizon, at which unaided competence reaches 61.1 % against 56.8 % without AI — a gain of 4.2 percentage points, $d = 0.28$. The break-even reliance is $a_{be} = 0.70$: a student who consults the model on more than seven problems in ten ends the semester less competent than one who never used it. At unrestricted reliance competence falls to 46.8 %, a loss of 10.1 percentage points or 17.7 % in relative terms, $d = 0.67$.

Figure 1b shows the second result. The assisted coursework score rises monotonically with reliance while retained competence turns down, so the two measures diverge: at $a = 1$ the coursework mark stands at 84.0 % against an unaided competence of 46.8 %, an illusion gap of 37.3 percentage points. Over most of the admissible range the signal an institution routinely collects moves in the opposite direction to the quantity it is meant to track.

A third result concerns detectability. Table 2 gives the sample sizes implied by (5). Demonstrating the benefit of optimal reliance requires 198 students per group, whereas demonstrating the harm of unrestricted reliance requires only 35 — the harm is roughly six times cheaper to detect than the benefit.

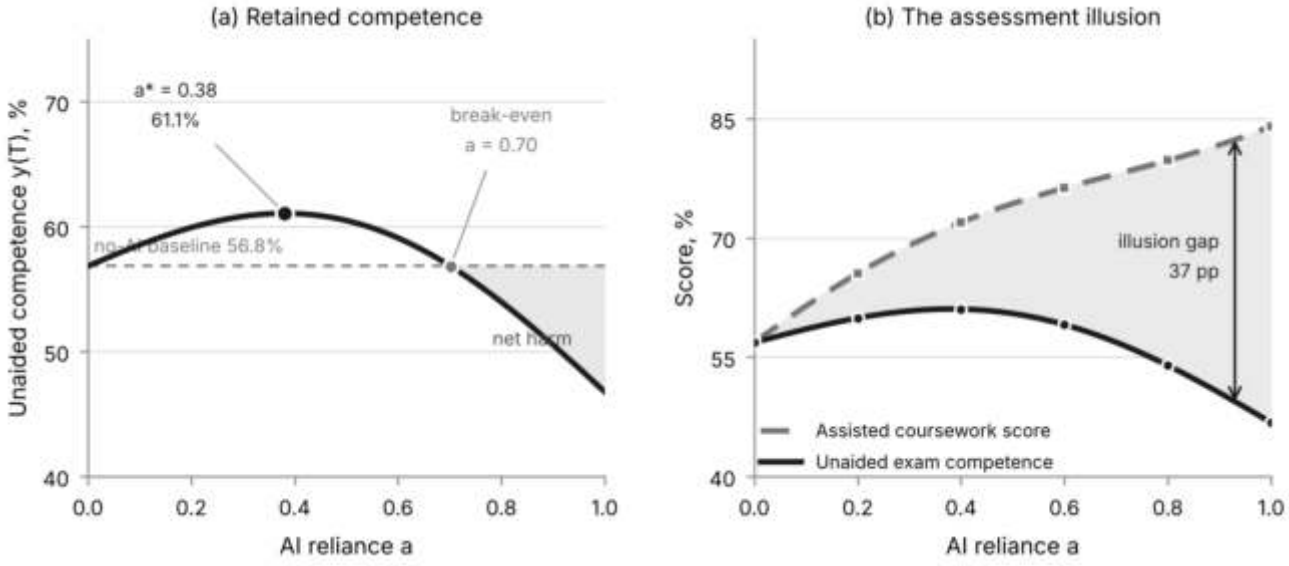


Figure 1. Retained competence (a) and assisted-versus-unaided divergence (b)

Table 2. Regimes of reliance and the sample size needed to detect each

Regime	$y(T)$, %	Δ vs no AI, pp	Effect size d	n per group
No AI (a = 0)	56.8	—	—	—
Optimal (a = 0.38)	61.1	+4.2	+0.28	198
Break-even (a = 0.70)	56.8	0.0	0.00	—
Unrestricted (a = 1)	46.8	−10.1	−0.67	35

Note: effect sizes are computed against the no-AI baseline with σ = 0.15.

4. Discussion

The model is calibrated at two points by a single external study and is not fitted to anything else, which makes the agreement non-trivial: Δu and κ were set from the unassisted deficit, η from the assisted improvement, and the two calibrations are independent of one another [2]. Everything between those endpoints — the location of the optimum, the break-even threshold, the shape of the illusion gap — is a prediction of the functional form rather than a fitted quantity.

The central practical conclusion is that the governing question is a budget rather than a prohibition. Reliance of roughly two problems in five beats either extreme, and the penalty for moderate error is mild: anywhere in $0.2 \leq a \leq 0.55$ the student is within one percentage point of the optimum. What must be avoided is not AI use but unexamined use above the break-even threshold, which an institution that neither measures nor limits reliance cannot detect.

The asymmetry in detectability matters when reading the empirical literature. Because harm at high reliance is six times cheaper to detect than benefit at the optimum, research at conventional sample sizes will accumulate evidence of harm faster than evidence of benefit even if the model is exactly right. The shape of the published evidence is thus partly an artefact of statistical power, and a null result on benefit warrants weaker inference than it appears to.

The model also reconciles the two field experiments that motivated it. A tutor with safeguards, which withholds the solution and requires the student to reason, raises u while suppressing the growth of β, because the student still performs the generation. In the model's terms, guardrails do not reduce AI use — they reduce its offloading content, which is why the guarded arm in [2] escaped the harm and the tutor in [3] produced gains.

A further prediction concerns the time horizon. Because decay acts for longer over a longer course, the optimum shifts downward as the horizon grows: $a^* = 0.38$ over a semester but 0.30 asymptotically. Reliance appropriate for a short intensive module is therefore too high for a year-long course.

Limitations. The cubic penalty in (1) is a structural assumption, not a measurement: its exponent governs where the optimum falls, and only the endpoints are anchored to data. Reliance is treated as a scalar, though which problems are offloaded plausibly matters more than how many. AI assistance is assumed correct, so harm from confidently wrong guidance is absent. Finally, a is treated as a policy variable, whereas students choose it — and those most inclined to offload are those for whom the penalty is steepest.

5. Conclusions

1. AI reliance was modelled as a continuous control variable acting on competence formation through two opposing channels: a linear gain in engagement and coverage and a convex penalty on retention.

2. Retained competence is a concave function of reliance with an interior optimum at $a^* \approx 0.38$, which raises unaided competence from 56.8 % to 61.1 % (d = 0.28) over a semester.

3. Reliance above $a_{be} \approx 0.70$ is worse than

using no AI at all, and unrestricted reliance lowers unaided competence by 17.7 % (d = 0.67).

4. Assisted coursework scores diverge from retained competence, reaching 37 percentage points at full reliance, so marks collected under AI availability are not a valid measure of attainment and assessment must include an unaided component.

5. Harm at high reliance is about six times cheaper to detect than benefit at optimal reliance, which biases the shape of the accumulating empirical literature.

6. The optimum falls as the course horizon lengthens, and safeguards act by reducing the offloading content of AI use rather than its quantity. Testing these predictions requires a trial in which reliance is measured rather than inferred.

References

1. Decree of the President of the Republic of Uzbekistan No. PF-5847 of 8 October 2019 “On approval of the Concept for the Development of the Higher Education System until 2030”. – Tashkent: Lex.uz, 2019.
2. Bastani H., Bastani O., Sungu A., Ge H., Kabakcı Ö., Mariman R. Generative AI without guardrails can harm learning: Evidence from high school mathematics // Proceedings of the National Academy of Sciences. – 2025. – Vol. 122, No. 26. – e2422633122.
3. Kestin G., Miller K., Klales A., Milbourne T., Ponti G. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting // Scientific Reports. – 2025. – Vol. 15. – Art. 97652.
4. Risko E.F., Gilbert S.J. Cognitive offloading // Trends in Cognitive Sciences. – 2016. – Vol. 20, No. 9. – P. 676–688.
5. Sparrow B., Liu J., Wegner D.M. Google effects on memory: Cognitive consequences of having information at our fingertips // Science. – 2011. – Vol. 333, No. 6043. – P. 776–778.
6. Roediger H.L., Karpicke J.D. Test-enhanced learning: Taking memory tests improves long-term retention // Psychological Science. – 2006. – Vol. 17, No. 3. – P. 249–255.
7. Bjork E.L., Bjork R.A. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning // Psychology and the Real World. – New York: Worth Publishers, 2011. – P. 56–64.
8. Sweller J., Ayres P., Kalyuga S. Cognitive Load Theory. – New York: Springer, 2011. – 274 p.
9. Murre J.M.J., Dros J. Replication and analysis of Ebbinghaus’ forgetting curve // PLoS ONE. – 2015. – Vol. 10, No. 7. – e0120644.
10. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. – Hillsdale: Lawrence Erlbaum Associates, 1988. – 567 p.