RICESO
Revista Interdisciplinaria de Ciencias de la Educación, Salud y Sociología

IA generativa en proyectos STEAM interculturales y funciones ejecutivas en bachillerato

Elian Jamil Hernández Cueva
María Angélica Idrobo Gutiérrez
Lisseth Consuelo Sánchez Álvarez
Camila Mishell Romero Alberca
Viviana Nayely Ortega Villamagua

https://doi.org/10.66136/9jnrbk35

Universidad Nacional de Loja

Ecuador

Generative AI in intercultural STEAM projects and executive functioning in secondary education

Elian Jamil Hernández Cueva
María Angélica Idrobo Gutiérrez
Lisseth Consuelo Sánchez Álvarez
Camila Mishell Romero Alberca
Viviana Nayely Ortega Villamagua

https://orcid.org/0009-0000-1007-9929
https://orcid.org/0000-0003-4127-9143
https://orcid.org/0009-0009-7165-4718
https://orcid.org/0009-0002-8775-2215
https://orcid.org/0009-0006-6033-651X

elian.hernandez@unl.edu.ec
maidrobog@unl.edu.ec
lissethsanchez018@gmail.com
camila.m.romero@unl.edu.ec
viviana.ortega@unl.edu.ec

Universidad Nacional de Loja

Ecuador

https://doi.org/10.66136/9jnrbk35

Background. Claims that generative AI (GenAI) erodes higher-order cognition rest largely on correlational surveys and brief laboratory tasks with university students in the Global North. Whether these costs belong to the technology or to the instructional design around it is untested in secondary classrooms, where executive functioning is rarely an outcome in its own right. **Aims.** We tested whether an eight-week intercultural STEAM unit treating GenAI as an object of verification rather than a source of answers improved planning, cognitive flexibility and inhibitory control, and whether prompting quality explained change in inhibition. **Method.** Quasi-experimental pre-test/post-test design, two intact cohorts, Saraguro canton, Ecuador (N = 138; Mage = 16.34). The experimental cohort (n = 70) built automated irrigation prototypes under a recursive prompting protocol grounded in Kichwa-Saraguro water management; the active control (n = 68) completed the same brief through conventional project-based learning. Executive functioning was measured with the BRIEF-2 and compared by ANCOVA (baseline covariate; Bonferroni α = .0167). Prompt logs were scored on a three-facet rubric and regressed on residualised change. **Results.** Baselines were equivalent. Adjusted effects favoured the experimental condition for Planning/Organisation (d = 1.33), Shift (d = 0.85) and Inhibit (d = 0.62). Prompt quality predicted residualised change in Inhibit (r = −.74, R^2 = .54) but not Shift (r = −.22); Verification alone added 20.8% of variance. **Conclusions.** The association was facet-specific, consistent with targeted training of response inhibition rather than general cognitive benefit. Self-reported outcomes and condition–school confounding limit the findings to plausibility.

Received: 13/10/2026
Accepted: 13/11/2026
Published: 13/11/2026

Revista Interdisciplinaria de Ciencias de la Educación, Salud y Sociología
https://www.riceso.org

editor@riceso.org

© 2026. Este artículo es un documento de acceso abierto distribuido bajo los términos y condiciones de la **Licencia Creative Commons, Atribución-NoComercial 4.0 Internacional.**

Keywords: generative artificial intelligence; executive functions; STEAM education; prompt engineering; intercultural education.

IA generativa en proyectos STEAM interculturales y funciones ejecutivas en bachillerato

Resumen

Antecedentes. La afirmación de que la inteligencia artificial generativa (IAG) deteriora la cognición de orden superior se sostiene sobre todo en encuestas correlacionales y tareas breves de laboratorio con estudiantes universitarios del Norte global. Queda sin probar en aulas de secundaria si ese costo cognitivo pertenece a la tecnología o al diseño instruccional que la rodea, y las funciones ejecutivas rara vez se han tratado como resultado por derecho propio. **Objetivos.** Evaluar si una unidad STEAM intercultural de ocho semanas, que sitúa a la IAG como objeto de verificación y no como fuente de respuestas, mejora la planificación, la flexibilidad cognitiva y el control inhibitorio, y si la calidad del *prompting* explica el cambio en inhibición. **Método.** Diseño cuasiexperimental pretest-postest con dos grupos intactos en el cantón Saraguro, Ecuador (N = 138; *M*edad = 16,34). El grupo experimental (n = 70) construyó prototipos de riego automatizado bajo un protocolo recursivo de *prompting* anclado en el manejo ancestral del agua kichwa Saraguro; el control activo (n = 68) desarrolló el mismo encargo mediante aprendizaje basado en proyectos convencional. Las funciones ejecutivas se midieron con el BRIEF-2 y se contrastaron mediante ANCOVA (covariable: línea base; α de Bonferroni = ,0167). Los registros de interacción se calificaron con una rúbrica de tres facetas y se regresaron sobre puntuaciones de cambio residualizado. **Resultados.** Hubo equivalencia inicial. Los efectos ajustados favorecieron al grupo experimental en Planificación/Organización (d = 1,33 [1,09, 1,58]), Flexibilidad (d = 0,85 [0,60, 1,11]) e Inhibición (d = 0,62 [0,36, 0,87]). La calidad del *prompting* predijo el cambio residualizado en Inhibición (r = −,74; R^2 = ,54), pero no en Flexibilidad (r = −,22; p = ,074); la faceta de Verificación aportó por sí sola el 20,8 % de la varianza. **Conclusiones.** La asociación fue específica por faceta, lo que resulta más compatible con un entrenamiento dirigido de la inhibición de respuesta que con un beneficio cognitivo general. El autoinforme y la confusión entre condición e institución limitan los hallazgos a la plausibilidad.

Palabras clave: inteligencia artificial generativa; funciones ejecutivas; educación STEAM; ingeniería de prompts; educación intercultural.

1. Introduction

Two years of rapid classroom adoption have produced a literature on generative artificial intelligence in education that points in opposite directions at once. Meta-analytic syntheses report moderate to large gains in academic performance and engagement when ChatGPT is embedded in instruction (Deng et al., 2025; Heung & Chiu, 2025). At the same time, a parallel body of work documents what happens when learners are left alone with the tool: frequent users score lower on critical-thinking assessments, with cognitive offloading carrying the association (Gerlich, 2025); learners supported by ChatGPT in a writing task performed as well as peers but engaged in markedly fewer metacognitive processes, a pattern Fan et al. (2025) named metacognitive laziness; and a systematic review of AI dialogue systems concluded that sustained reliance is associated with reduced analytical engagement across several cognitive domains (Zhai et al., 2024). Reviews that pool both strands reach no verdict: Sardi et al. (2025) found that roughly seven in ten studies reported positive effects on self-regulated learning while a substantial minority warned of dependency, and qualitative work with learners records the same ambivalence in students' own accounts (Darwin et al., 2024).

The contradiction is less puzzling once the moderators are inspected. In Deng et al.'s (2025) synthesis, effects were strongest where GenAI was integrated into structured classroom practice and weakest where it was bolted on as an optional aid. Futterer et al. (2026) found in a randomised trial with school students that GenAI support improved self-regulated learning only when it was scaffolded. Two further experimental studies converge on the same boundary condition: Cong-Lem et al. (2025) obtained critical-thinking gains only after a structured experiential intervention, and Hong et al. (2025), in the study closest to ours in logic, reported that instruction explicitly directed at how learners offload cognition to GenAI produced measurable critical-thinking gains in writing where unguided use did not. The emerging picture is that instructional design, not the model, determines whether a large language model functions as a substitute for thinking or as a demand on it. That picture, however, has been assembled almost entirely from higher-education samples in high-income settings, and it has been evaluated against achievement, engagement and self-regulation measures rather than against the cognitive architecture those measures presuppose.

1.1 Three gaps

Executive functioning is assumed but not measured. Claims about cognitive offloading are claims about executive control: about whether learners still plan before acting, still shift strategy when a strategy fails, and still inhibit the impulse to accept a fluent answer. Yet studies that name these processes typically operationalise them through critical-thinking inventories or self-regulation scales. Direct measurement of planning, cognitive flexibility and inhibitory control in GenAI-mediated instruction is close to absent. This matters because the three components dissociate. Friedman and Robbins (2022) show that common executive control is closely tied to response inhibition and is separable from set shifting and working-memory updating, which implies that an intervention could plausibly affect one component without affecting the others. Aggregate outcome measures cannot detect such a pattern.

The adolescent case is under-represented. Late adolescence is the period in which prefrontal networks supporting planning and inhibition are still consolidating, and it is also the period in which students in technical tracks begin using GenAI for coursework without instruction in how to do so. Reviews of AI in education continue to be dominated by tertiary samples (Fu et al., 2025; An et al., 2026), and secondary-school evidence remains thin (Ng et al., 2024). The AI-literacy field shows the same skew: a decade-scale mapping of the research base found school-level empirical work heavily outweighed by conceptual and tertiary studies (Yang et al., 2025), even though the construct was articulated for learners of all ages (Ng et al., 2021) and has been operationalised as far down as early childhood (Su et al., 2023).

Intercultural STEAM and GenAI research run in parallel. Integrated STEM produces moderate achievement gains (Chen et al., 2026; Cao et al., 2025), and culturally grounded variants are advocated on epistemic-justice grounds (Ogegbo & Ramnarain, 2024; da Silva et al., 2024). A recent systematic review of STEM+intercultural work found that ancestral and traditional knowledge is most often incorporated as thematic content rather than as a source of problems to be modelled, and identified almost no studies in which digital tools mediated that integration (Oviedo-Berrocal et al., 2026). The literature on GenAI in education, for its part, has barely engaged with Indigenous contexts at all.

1.2 The present study

We address these gaps in a setting that makes the intercultural question concrete. Saraguro canton, in Loja province, southern Ecuador, is the territory of the Kichwa-Saraguro nationality. Its agricultural system rests on water-management practices with a documented internal logic, including gravity-fed distribution and rotation rules that are, in effect, an engineering scheme developed without engineering vocabulary. That system offers a problem space in which a large language model can be put to work as a translator between two technical registers rather than as an oracle.

The intervention was built on a single design commitment: students were never asked to obtain an answer from the model. They were asked to constrain it, to iterate on its failures, and to audit its output against primary sources and community knowledge holders. Under this arrangement the model's unreliability, ordinarily treated as a liability (Yan et al., 2024), becomes the instructional mechanism.

Three research questions follow.

RQ1. Does an eight-week GenAI-mediated intercultural STEAM unit produce greater improvement in planning, cognitive flexibility and inhibitory control than an equivalent unit taught through conventional project-based learning?

RQ2. Within the experimental condition, does the quality of students' prompting behaviour predict change in executive functioning, and is that prediction general or specific to particular components?

RQ3. Do the project artefacts produced under the two conditions differ in technical quality, scientific rigour and cultural pertinence?

Hypotheses were directional for RQ1 and RQ2. Because the BRIEF-2 is scored so that higher values indicate greater everyday dysfunction, improvement corresponds to a *decrease* in *T* scores; this convention is applied throughout.

- H1. Adjusted post-test *T* scores will be lower in the experimental condition than in the control condition for Planning/Organisation, Shift and Inhibit.
- H2. Within the experimental condition, prompt quality will be inversely associated with residualised change scores, with the strongest association for Inhibit, on the grounds that

the verification requirement trains response inhibition directly whereas the planning and iteration requirements act on it indirectly.

- H3. Artefacts produced under the experimental condition will not differ from control artefacts in scientific rigour, since the underlying engineering content was identical.

H3 is deliberately a null hypothesis. It functions as a discriminant check: if the experimental condition outperformed the control on every rubric dimension, the most parsimonious explanation would be rater expectancy rather than learning.

2. Theoretical framework

2.1 Unreliability as an instructional resource

A language model that always returned correct output would be pedagogically inert, because the learner would have no reason to evaluate it. The models students actually use are fluent, confident and intermittently wrong, and the combination is what creates the instructional opening. Klingbeil et al. (2024) demonstrated experimentally that people follow AI advice even when contextual information contradicts it and when following it damages their own interests, and that the mere knowledge that advice is machine-generated is sufficient to induce over-reliance. Nazaretsky et al. (2025) report the same dynamic in educational settings, where trust in the system predicts uncritical acceptance.

Over-reliance is therefore the default behavioural response, not an aberration. Our design treats it as the target of instruction. If accepting the first output is the prepotent response, then requiring students to withhold acceptance until they have checked the output against an independent source is structurally a response-inhibition task. The operation is the same one measured by classical go/no-go paradigms, transposed to an authentic setting: a salient, easy, rewarded response must be suppressed in favour of a slower one. The theoretical claim we test is narrow and falsifiable. It is not that GenAI use improves cognition; it is that a specific instructional requirement, repeated over eight weeks, trains a specific executive component.

2.2 Prompting as externalised metacognition

Prompt engineering is often described as a technical skill. The review literature suggests a different reading. Qian (2025) distinguishes technique-based prompting, aimed at output quality, from process-based prompting, aimed at cognitive engagement, and finds that the latter is where the

learning effects sit. To obtain useful output from a model on a non-trivial task, a student must decompose the problem, state constraints explicitly, anticipate the form of an acceptable answer, and revise when the answer fails. Those operations are ordinarily internal and inaccessible. Written into a prompt, they become visible, to the student and to the researcher.

This is the property that makes prompt logs valuable as process data. A log is a record of how a student structured a problem, not merely of what they asked. Chiu (2024) and Chang et al. (2023) have argued for using such traces to scaffold self-regulation; Chang and Sun (2024) document their use in language learning. The wider evidence on dialogic uses of conversational agents is consistent with this reading. Guo et al. (2023) found that chatbot-assisted in-class debates improved argumentation skills relative to conventional debate, Escalante et al. (2023) showed that AI-generated feedback can match human feedback on writing outcomes, and Lu et al. (2024) reported gains in higher-order thinking when generative tools were used to structure reasoning rather than to supply it. What these studies share is an arrangement in which the model is something to argue with. We extend the argument by treating prompt quality as a predictor of neuropsychological change rather than of task performance.

2.3 Executive functioning in late adolescence

Executive functions are the control processes that regulate goal-directed behaviour. The three components relevant here are planning and organisation, cognitive flexibility or set shifting, and inhibitory control. Their trajectory extends into the early twenties, which makes upper-secondary students a developmentally sensitive population.

The unity-and-diversity framework (Friedman & Robbins, 2022) is important for our predictions. Because a common factor underlies all executive tasks while shifting and updating retain component-specific variance, an intervention targeting inhibition may produce collateral movement on other components through the common factor, but the *strongest* and most specific association should appear where the training demand sits. H2 is a direct test of this. If prompt quality predicted change equally across all three components, the common-factor account would be favoured and the mechanistic claim weakened.

2.4 Situated STEAM and epistemic justice

Integrated STEM education is characterised by the connection of disciplinary content through authentic problems (Portillo-Blanco et al., 2024). Transplanting the model without regard to context risks reproducing the assumption that scientific knowledge originates elsewhere and arrives ready-made. Reviews of Indigenous knowledge integration in STEM education consistently report that the strongest implementations are those in which community knowledge supplies the problem rather than decorating it (Ogegbo & Ramnarain, 2024; Cherry Shive, 2025).

The risk is sharper with GenAI. Large language models are trained predominantly on English-language, Northern-hemisphere corpora, and their defaults reflect that distribution. A student who asks such a model about irrigation will receive an answer shaped by contexts that are not theirs. Our protocol required students to supply the local constraints the model could not know, which inverts the relationship: the student holds the knowledge the system lacks. Whether this inversion is achieved in practice is an empirical question, addressed through RQ3.

3. Method

3.1 Design

The study used a quasi-experimental pre-test/post-test design with two non-equivalent intact groups and an active control. Random assignment at the student level was not possible because class composition is fixed administratively, a constraint common to school-based intervention research. The design is represented as:

- Experimental group: $O_1 \rightarrow X \rightarrow O_2$
- Control group: $O_3 \rightarrow — \rightarrow O_4$

where $O_{1,3}$ is the baseline BRIEF-2 administration, X is the GenAI-mediated intercultural STEAM protocol, — is the conventional project-based learning condition, and $O_{2,4}$ is the post-intervention administration eight weeks later.

The active control is a design feature rather than a convenience. Both cohorts completed the same engineering brief, for the same number of instructional hours, producing the same class of artefact. The conditions differed only in the mediation of the design process. This isolates the GenAI protocol from the effects of project-based learning itself, which are independently established (Chen et al., 2026).

We note at the outset, and return to it in Section 5.4, that condition was confounded with institution. This is the principal limitation of the design and it bounds every causal interpretation offered below.

3.2 Setting and participants

Participants were third-year students in the technical baccalaureate stream at two schools in Saraguro canton. Of 161 students assessed for eligibility, 138 met the criteria and completed both assessment points (Figure 1). Mean age was 16.34 years (*SD* = 0.82, range 15–17).

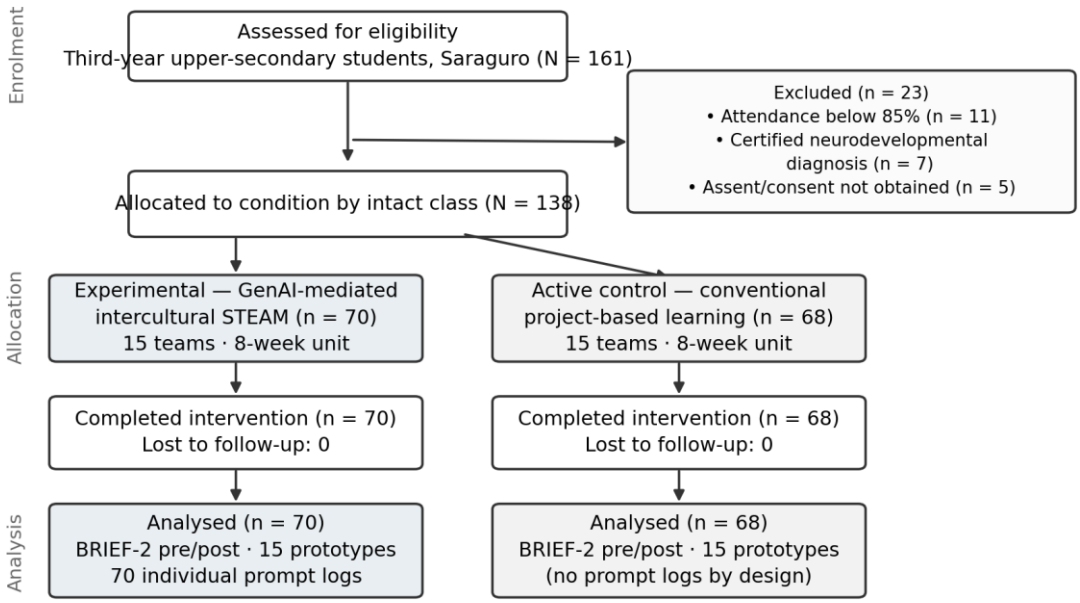


Figure 1. *Participant flow. No participants were lost between allocation and analysis; the analysed sample therefore corresponds to the allocated sample.*

The experimental cohort (*n* = 70) was drawn from Unidad Educativa Saraguro, a public school operating under the intercultural bilingual modality. The control cohort (*n* = 68) was drawn from Unidad Educativa Fiscomisional San José de Calasanz, which operates under the Hispanic modality. Both follow the national curriculum and serve comparable catchment areas; both cohorts are predominantly of Kichwa-Saraguro descent.

Inclusion criteria were current enrolment, attendance above 85% during the intervention window, sufficient Spanish for instrument administration, and written parental consent plus student assent. Exclusion criteria were a certified neurodevelopmental diagnosis and failure to complete either assessment. No participant was excluded for the latter reason, so the analysed and allocated

samples coincide and the intention-to-treat and per-protocol analyses are identical. The original protocol described an intention-to-treat analysis alongside exclusion for attrition; since attrition was zero, the distinction is moot here, but future replications should specify that intention-to-treat takes precedence.

3.3 Measures

Executive functioning. The Behavior Rating Inventory of Executive Function, Second Edition (BRIEF-2), self-report form, was administered in Spanish. The instrument yields *T* scores standardised to *M* = 50, *SD* = 10, with values of 60–64 indicating elevated risk and values at or above 65 indicating clinical significance. Higher scores denote greater dysfunction. Analyses focused on three clinical scales, Planning/Organisation, Shift and Inhibit, and on the Global Executive Composite (GEC).

The factor structure of the BRIEF-2 has been confirmed in Spanish-speaking adolescents, with a three-index solution comprising behavioural, emotional and cognitive regulation providing the best fit (Juárez & Arán Filippetti, 2025). Internal consistency in the present sample was acceptable to good: Planning/Organisation α = .87, ω = .88; Shift α = .84, ω = .85; Inhibit α = .86, ω = .86; GEC α = .91, ω = .92. The validity scales were screened at both time points. Three protocols exceeded the Negativity cut-off at pre-test and were retained after review because the corresponding Inconsistency and Infrequency indices were within normal limits; sensitivity analyses excluding them are reported in Section 4.6.

Prompt quality. Complete individual chat transcripts from the experimental cohort were scored with a purpose-built rubric assessing three facets on 0–10 scales:

1. *Contextualisation*, the explicit specification of role, setting and local constraints;
2. *Logical constraint*, the imposition of structural limits on the output, such as component cost ceilings or sensor specifications;
3. *Verification*, documented evidence that model output was checked against primary sources, physical testing or community knowledge holders.

Content validity was established by expert review (*n* = 5), yielding Aiken's *V* = .92. Two raters blind to student identity scored all 70 logs; agreement was high (ICC(2,2) = .89, 95% CI [.83, .93]). Internal consistency of the three-facet composite was α = .85, ω = .86, and the facets were substantially intercorrelated (*r* = .69 to .77), supporting use of a composite alongside facet-level analysis.

Project artefacts. The 30 prototypes were scored on technical feasibility, scientific rigour and cultural pertinence (0–10 each) by three external assessors — two agricultural engineers and one science-education specialist — who were blind to condition. Artefacts were stripped of identifying markers before assessment. Inter-rater agreement was Fleiss' κ = .82.

A caution about the third dimension is warranted here rather than in the discussion. Only the experimental condition was instructed to integrate ancestral knowledge. Any difference in cultural pertinence therefore indicates whether the manipulation was delivered, not whether it produced learning. We report it as a fidelity check and interpret it as nothing more.

3.4 Conditions

Both cohorts worked for six weeks (four periods weekly) on the brief *Design and construction of an automated irrigation system*, organised in teams of four to five, producing 15 prototypes per condition. Teaching was delivered by instructors matched on qualification and years of experience. Weekly fidelity observations were conducted by the research team using a structured checklist; adherence exceeded 90% of protocol steps in both conditions.

Control condition. Teams followed standard project-based learning. Code was written in a conventional IDE with support from teacher-led instruction and textbooks. Information searches used conventional web search. GenAI tools were not permitted during class sessions. Out-of-class use could not be prevented and was not monitored, which is a limitation of the comparison.

Experimental condition. Teams followed a four-stage recursive protocol. Each student held an individual account, so that although the artefact was collective, 70 individual interaction records were generated.

1. **Role and contextualisation.** Students configured the model with a dual role, technical and territorial, and supplied local parameters the model could not infer, including altitude, rainfall regime, soil characteristics and the community's rotation conventions.
2. **Logical constraint.** Students imposed explicit output restrictions before requesting anything, for example specifying code optimised for low-cost soil-moisture sensors within a stated budget.
3. **Iteration on failure.** Students were prohibited from accepting a first response. When code failed to compile or a recommendation conflicted with local conditions, they reformulated rather than requested a rewrite.
4. **Hallucination audit.** Every factual claim used in the prototype was checked against a primary source, a physical test, or a community knowledge holder. Unverified claims were excluded from the design regardless of plausibility.

Stage 4 is the operationalisation of the inhibitory demand. Stages 2 and 3 correspond, respectively, to planning and flexibility demands.

The institutions are geographically separated and the cohorts do not share class time, which limits treatment diffusion, though informal contact between students of different schools cannot be ruled out.

3.5 Procedure

Week 1 comprised institutional authorisation, consent and assent procedures, and baseline BRIEF-2 administration to the full sample under standardised morning conditions. Weeks 2–7 comprised the intervention. Week 8 comprised post-test administration, blind assessment of prototypes, collection of the 70 prompt logs, and a compensatory prompt-engineering workshop delivered to the control cohort.

3.6 Analytic strategy

Analyses were conducted in IBM SPSS Statistics 27 and verified in R 4.3.2. Two-tailed α was set at .05 and Bonferroni-corrected to .0167 for the three primary executive outcomes. The GEC was treated as a secondary outcome.

Sensitivity to effect size. With $n = 68$ and $n = 70$, $\alpha = .0167$ and power of .80, the design detects a between-group standardised difference of $d = 0.56$ in an unadjusted comparison. Because the pre-test covariate correlated $r = .60$ with post-test scores on average, the ANCOVA reduces the minimum detectable effect to approximately $d = 0.44$. The study was therefore adequately powered for medium and large effects and underpowered for small ones. This is stated in advance because the smallest observed effect, on Inhibit, sits close to this boundary.

Primary analysis. Between-group contrasts used ANCOVA with the corresponding baseline score as covariate, which is preferred to raw post-test comparison or gain-score analysis in pre-test/post-test designs because it adjusts for residual baseline differences and increases precision. Homogeneity of regression slopes was verified for each model through the group × covariate interaction (all $p > .28$). Adjusted effect sizes are reported as the adjusted mean difference divided by the pooled post-test standard deviation, with 95% confidence intervals, alongside partial η^2.

Assumptions. Normality was assessed with Shapiro–Wilk statistics and skewness and kurtosis indices; homogeneity of variance with Levene's test based on means. Where Levene's test indicated

heterogeneity, Welch's correction was applied. Residual diagnostics for the regression models included variance inflation factors and the Durbin–Watson statistic.

Process analysis. Change in executive functioning was operationalised as a residualised change score: post-test regressed on pre-test within the experimental cohort, with the residual retained. Simple difference scores were avoided because they are confounded with regression to the mean and are less reliable than either constituent measure. Associations with prompt quality were examined through Pearson correlations with Fisher-transformed confidence intervals and through hierarchical regression entering the non-focal facets first.

Clinical significance. The Reliable Change Index was computed for Inhibit using the standard error of measurement derived from the baseline standard deviation and the scale reliability, with $|RCI| \geq 1.96$ as the threshold for reliable individual change.

3.7 Ethical considerations

The protocol adhered to the Declaration of Helsinki and to CIOMS guidance, and was approved by the participating institutions as minimal-risk research. Written consent was obtained from legal guardians and assent from students; participation was voluntary and withdrawal carried no consequence.

In accordance with Ecuador's Organic Law on Personal Data Protection, participants were alphanumerically coded. Students were instructed not to enter personally identifying information into any AI platform and to restrict input to technical and academic content. Data and interaction logs were held under encrypted custody with access restricted to the research team. Under the principles of beneficence and justice, the control cohort received the prompt-engineering workshop after post-testing. The authors declare no competing interests.

4. Results

4.1 Preliminary analyses and baseline equivalence

Distributional assumptions were satisfied. Shapiro–Wilk statistics were non-significant for all scales in both groups at both time points ($p > .13$), skewness ranged from −0.41 to 0.44 and kurtosis from −0.62 to 0.51, all within conventional limits. Levene's test indicated homogeneous variances at

baseline (all $p > .47$) and at post-test (all $p > .45$), so uncorrected tests are reported throughout. No values exceeded $|z| = 3.0$ and no cases were removed.

The cohorts were equivalent on demographic and access variables (Table 1). The absence of a difference in home internet access is relevant, since differential connectivity would otherwise offer an alternative explanation for any advantage in the GenAI condition.

Table 1. *Demographic and access characteristics by condition*

Variable	Control (n = 68)	Exper. (n = 70)	Test	p
Age, *M* (*SD*)	16.32 (0.81)	16.35 (0.83)	$t(136) = -0.21$	.830
Female, *n* (%)	30 (44.1)	32 (45.7)	$\chi^2(1) = 0.04$	.851
Male, *n* (%)	38 (55.9)	38 (54.3)		
Home internet access, *n* (%)	65 (95.6)	68 (97.1)	$\chi^2(1) = 0.24$	.625

Baseline executive functioning was also equivalent (Table 2). All standardised differences were negligible ($|g| \leq 0.06$), and the confidence intervals around them exclude effects larger than small in either direction. Both cohorts sat in the at-risk band on Planning/Organisation and Shift, with mean *T* scores above 60. Between 17% and 37% of students scored in the clinically significant range on at least one scale at baseline, a level of reported executive difficulty that is high for a non-clinical sample and consistent with concerns about self-regulation among post-pandemic adolescent cohorts.

Table 2. *Baseline equivalence on BRIEF-2 scales*

Scale	Control	Experimental	$t(136)$	p	g [95% CI]
Planning/Organisation	62.45 (5.82)	62.31 (5.54)	0.15	.885	0.02 [−0.31, 0.36]
Shift	61.12 (6.01)	61.40 (5.78)	−0.28	.781	−0.05 [−0.38, 0.29]
Inhibit	59.80 (6.25)	59.44 (5.92)	0.35	.729	0.06 [−0.28, 0.39]
Global Executive Composite	61.15 (4.95)	61.10 (4.88)	0.06	.952	0.01 [−0.32, 0.34]

Note. Cells show *M* (*SD*). Higher *T* scores indicate greater everyday executive dysfunction. g = Hedges' g.

4.2 Effects on executive functioning

Within-group change diverged sharply (Figure 2). The control cohort did not change reliably on any scale over the eight weeks (all $p \geq .312$; $|d_{av}| \leq 0.10$), which supports the inference that maturation alone does not move these measures over this interval and strengthens the internal validity of the comparison. The experimental cohort improved on all three scales.

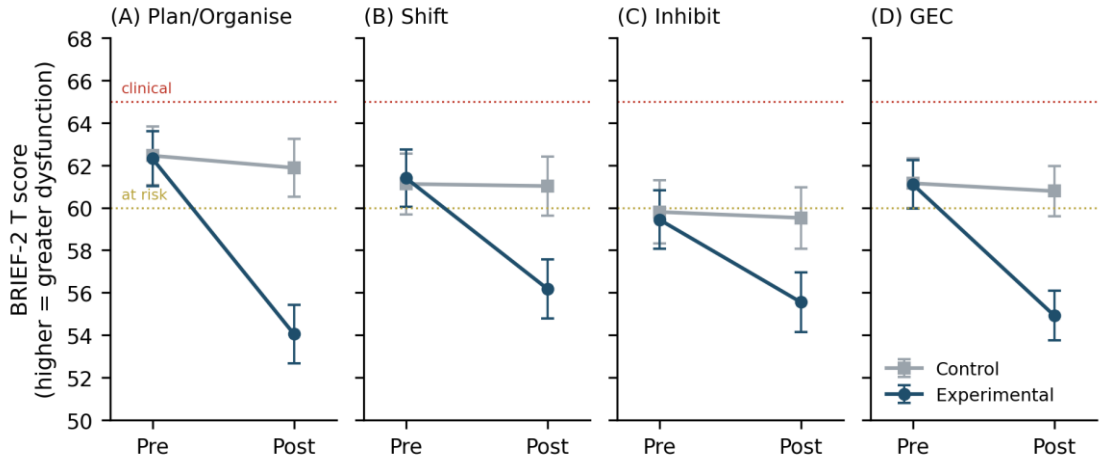


Figure 2. *BRIEF-2 T scores at pre-test and post-test by condition, with 95% confidence intervals. Dotted lines mark the at-risk (60) and clinically significant (65) thresholds. Lower scores indicate better executive functioning.*

The between-group contrasts, adjusted for baseline, are reported in Table 3. All three primary outcomes were significant at the Bonferroni-corrected threshold, and H1 was supported.

Table 3. *ANCOVA results for BRIEF-2 scales, with baseline score as covariate*

Scale	Adj. M_C	Adj. M_E	*b* [95% CI]	*t*(135)	*p*	*d* [95% CI]	η^2_p
Planning/Organisation	61.83	54.10	−7.73 [−9.17, −6.30]	−10.64	< .001	1.33 [1.09, 1.58]	.456
Shift	61.11	56.09	−5.02 [−6.52, −3.53]	−6.63	< .001	0.85 [0.60, 1.11]	.246
Inhibit	59.40	55.67	−3.73 [−5.25, −2.21]	−4.85	< .001	0.62 [0.36, 0.87]	.148

Scale	Adj. M_C	Adj. M_E	b [95% CI]	$t(135)$	p	d [95% CI]	η^2_p
Global Executive Composite	60.76	54.94	−5.82 [−6.96, −4.68]	−10.12	< .001	1.16 [0.94, 1.39]	.431

Note. Subscripts C and E denote the control and experimental conditions. Negative *b* indicates lower (better) adjusted post-test scores in the experimental condition. Effect sizes are the adjusted mean difference standardised on the pooled post-test *SD*, signed so that positive values favour the experimental condition. Bonferroni-corrected α = .0167 for the three primary outcomes.

Within the experimental cohort, changes were −8.26 *T* points on Planning/Organisation (*SD* = 4.58; $t(69) = -15.09$, $p < .001$, $d_{av} = -1.45$), −5.22 on Shift (*SD* = 4.83; $t(69) = -9.03$, $p < .001$, $d_{av} = -0.89$) and −3.89 on Inhibit (*SD* = 4.86; $t(69) = -6.70$, $p < .001$, $d_{av} = -0.65$).

The ordering of effects is the substantive finding (Figure 3). Planning improved most, flexibility next, inhibition least, and the confidence intervals for Planning and Inhibit do not overlap. This gradient is not what a general expectancy or demand-characteristic account predicts, since such an account would push all self-reported scales in the same direction to a similar degree. It does correspond to the structure of the protocol, in which planning demands were imposed at every interaction while verification demands applied only to factual claims entering the final design.

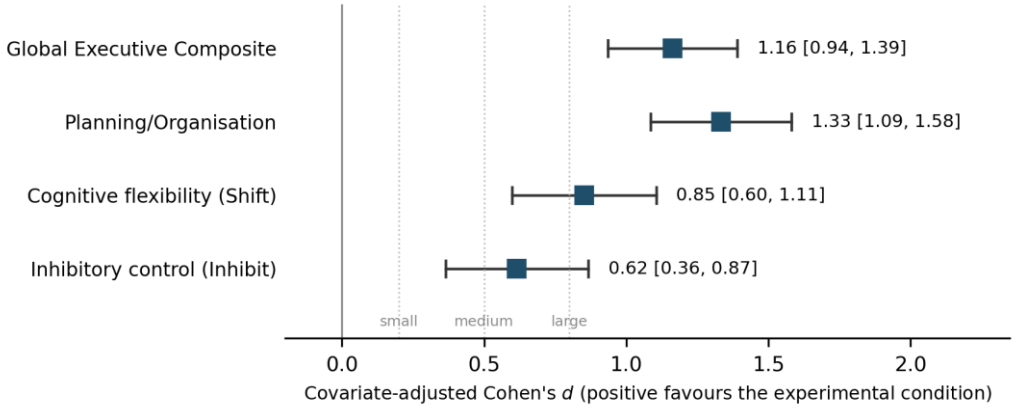


Figure 3. *Covariate-adjusted between-group effect sizes with 95% confidence intervals.*

The proportion of students scoring in the clinically significant range on Planning/Organisation fell from 37.1% to 4.3% in the experimental cohort, against 32.4% to 26.5% in the control cohort.

4.3 Reliable individual change

Group means can conceal heterogeneity, so individual change on Inhibit was evaluated against the Reliable Change Index (SEM = 2.22; S_{diff} = 3.13). Reliable improvement was observed for 17 of 70 experimental students (24.3%) and 7 of 68 control students (10.3%), $\chi^2(1) = 4.70$, $p = .030$, OR = 2.80 (Figure 4).

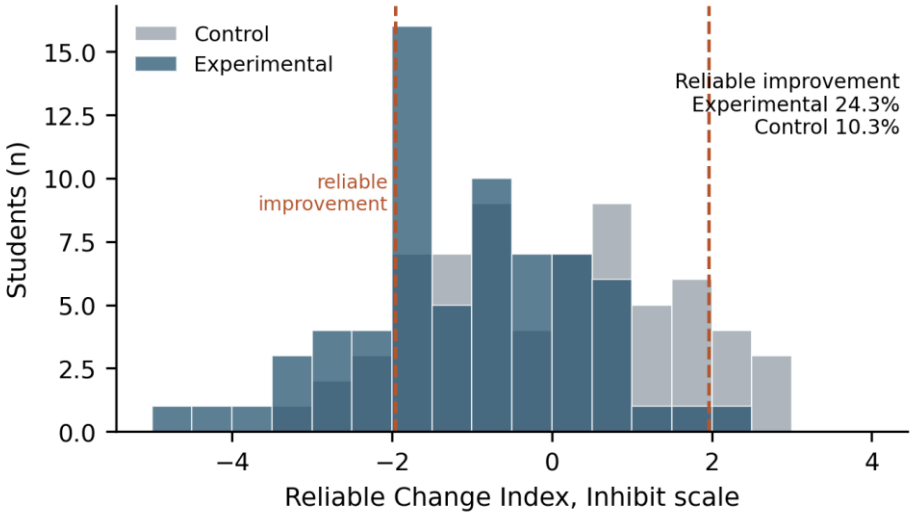


Figure 4. *Distribution of the Reliable Change Index for the Inhibit scale. Values below −1.96 indicate reliable individual improvement.*

The contrast between a group-level effect of d = 0.62 and reliable individual improvement in fewer than a quarter of participants deserves emphasis. Most students in the experimental condition shifted modestly; the aggregate effect is driven by a distributional shift rather than by uniform benefit. Reporting only the group mean would overstate what the intervention achieved for the typical student.

4.4 Prompt quality and executive change

Prompt-quality scores in the experimental cohort were moderate: Verification M = 5.47 (SD = 1.61), Logical constraint M = 5.66 (SD = 1.66), Iteration M = 5.78 (SD = 1.75), composite M = 5.76 (SD = 1.38). Eight weeks of instruction did not produce expert prompting behaviour, and the variance available for the analyses below reflects genuine individual differences rather than a ceiling.

Correlations with residualised change are reported in Table 4. The composite was strongly associated with change in Inhibit (r = −.74, 95% CI [−.83, −.61], R^2 = .54) and weakly and non-

significantly with change in Shift (r = −.22, p = .074). The association with change in Planning/Organisation (r = −.29, p = .017) did not meet the corrected threshold. H2 was therefore supported, and in its specific rather than general form.

Table 4. *Correlations between prompt-quality scores and residualised change in executive functioning (n = 70)*

Prompt-quality measure	Δ Plan/Org.	Δ Shift	Δ Inhibit
Logical constraint (facet)	−.49*** [−.65, −.29]		
Iteration on failure (facet)		−.43*** [−.60, −.22]	
Verification (facet)			−.79*** [−.86, −.68]
Prompt-quality composite	−.29 [−.49, −.05]	−.22 [−.43, .02]	−.74*** [−.83, −.61]

Note. Negative coefficients indicate that higher prompt quality is associated with greater improvement. *** p < .001; remaining coefficients did not meet the corrected threshold of .0167.

The facet-level pattern is diagnostic. Each facet showed its strongest association with the executive component it was designed to load: logical constraint with planning, iteration with flexibility, verification with inhibition. The facets are substantially intercorrelated, so these associations are not independent. What distinguishes them is magnitude, and the verification–inhibition association is markedly the largest (Figure 5).

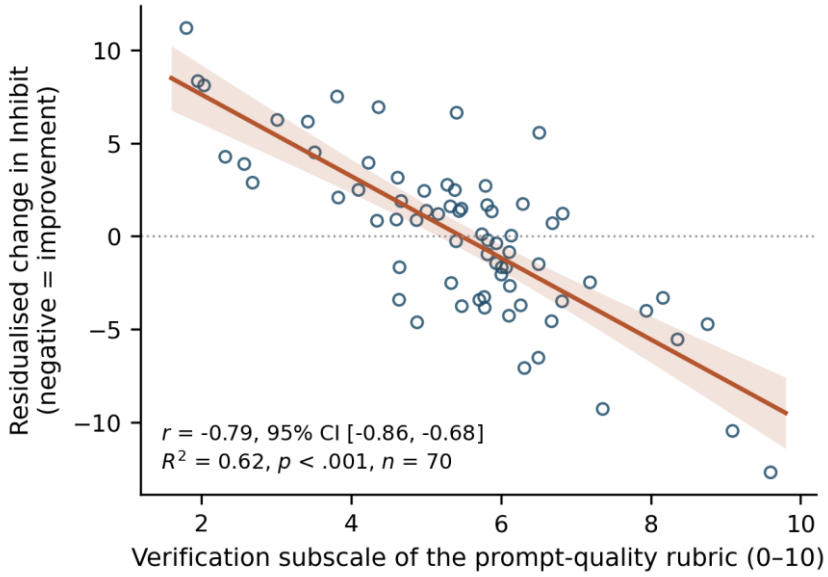


Figure 5. *Verification facet scores against residualised change in Inhibit within the experimental cohort. Shaded band shows the 95% confidence interval of the regression line.*

Hierarchical regression tested whether verification contributed beyond the other two facets. Entering logical constraint and iteration first accounted for 42.0% of variance in residualised change in Inhibit. Adding verification raised this to 62.8% (adjusted R^2 = .611), ΔR^2 = .208, $F(1, 66) = 36.91$, $p < .001$. In the full model, verification was the only significant predictor (b = −2.21, SE = 0.36, β = −.80, $p < .001$); logical constraint (β = .10, p = .428) and iteration (β = −.09, p = .427) did not contribute uniquely. Multicollinearity was tolerable (VIF 2.40–3.04) and residuals were independent (Durbin–Watson = 2.22). Overall model fit was $F(3, 66) = 37.12$, $p < .001$.

Two interpretive cautions apply. These associations are correlational within a single condition: students who verified more carefully may have had better inhibitory control to begin with, notwithstanding the use of residualised change, since baseline adjustment removes only linear dependence on the initial score. And both variables derive from the same participants over the same period, so a third variable such as conscientiousness or engagement could drive both.

4.5 Project artefacts

Prototypes were compared on the three rubric dimensions (Figure 6). Cultural pertinence differed substantially between conditions (M = 8.81 vs. 5.24; $t(22.6) = 9.10$, $p < .001$, g = 3.23 [2.14, 4.33]). As stated in Section 3.3, this is a fidelity indicator. Only the experimental condition was instructed to integrate ancestral knowledge, so the difference confirms the manipulation was delivered and

carries no information about learning. An effect of this magnitude in an educational intervention would be implausible as a substantive finding, and we do not treat it as one.

The informative comparisons are the other two. Technical feasibility favoured the experimental condition (M = 8.23 vs. 7.41; $t(27.4) = 2.27$, $p = .031$, $g = 0.81$ [0.08, 1.54]), but did not survive correction for multiple comparisons and has a confidence interval extending to a negligible effect. Scientific rigour did not differ (M = 8.06 vs. 7.52; $t(27.8) = 1.48$, $p = .151$, $g = 0.52$ [−0.18, 1.23]), consistent with H3.

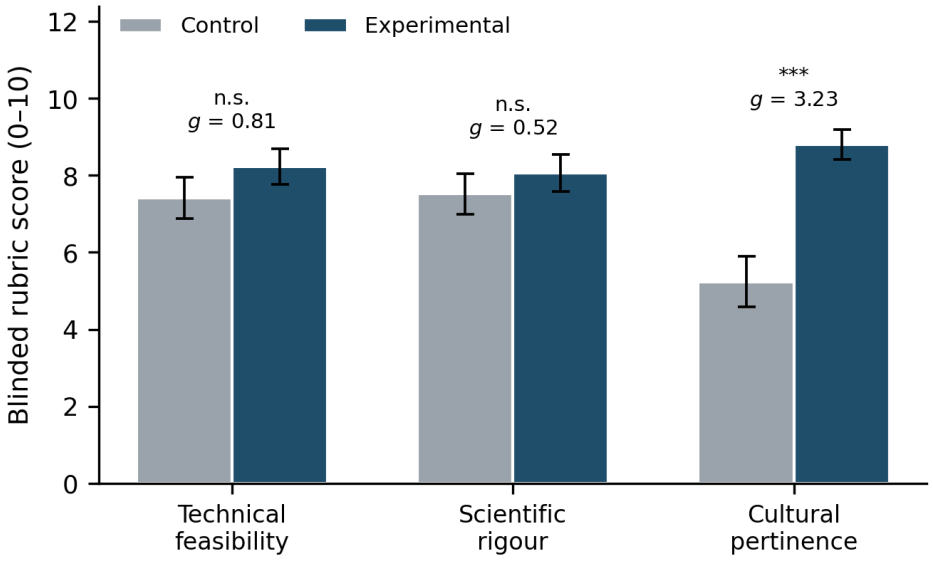


Figure 6. *Blinded rubric scores for the 30 prototypes (15 per condition) with 95% confidence intervals. The cultural pertinence contrast is a treatment-fidelity indicator, not an outcome.*

That the two cohorts produced artefacts of equivalent scientific rigour is important. It indicates that the assessors were not applying a general halo to experimental artefacts, and it means the executive-function differences cannot be attributed to the experimental cohort simply producing better work overall. Teams (n = 15 per condition) rather than students are the unit of analysis here, so these comparisons are appreciably less powerful than the individual-level analyses.

4.6 Sensitivity analyses

Four checks were run. Excluding the three protocols with elevated Negativity indices changed adjusted effect sizes by less than 0.03 in every case. Repeating the primary contrasts with robust regression using MM-estimation produced effects within 0.05 of the ordinary least squares estimates. Substituting simple difference scores for residualised change reduced the verification–

inhibition correlation from −.79 to −.71 but left the conclusion unchanged. Finally, because students are nested within 30 teams, the design effect was estimated across a plausible range of intraclass correlations: at ICC = .05, .10 and .15 the effective sample size falls to 117, 101 and 90, widening the confidence interval for the Inhibit effect to approximately [0.28, 0.95] in the least favourable case. The Inhibit effect remains significant across this range, though with noticeably reduced precision. Cluster-level data were not available to estimate the ICC directly, which is a shortcoming of the data collection rather than of the analysis.

5. Discussion

5.1 The design carries the effect

The findings are most compatible with the moderator account that has been emerging from recent syntheses. GenAI does not have a fixed cognitive valence. Gerlich (2025) reported that habitual AI use predicts weaker critical thinking through cognitive offloading, and Fan et al. (2025) found that unscaffolded GenAI support left performance intact while hollowing out metacognitive engagement. Our experimental cohort used the same class of tool intensively for eight weeks and improved on self-reported executive functioning. The difference lies in what students were required to do with the output.

This is the same conclusion Futterer et al. (2026) reached in a randomised trial with school students, where GenAI improved self-regulated learning only under scaffolding, and it aligns with Deng et al.'s (2025) moderator analysis, in which integration into structured classroom practice was the strongest predictor of positive effects. The practical implication is that debates about whether to permit GenAI in schools are less consequential than debates about what students are asked to do with it. A prohibition and an unstructured permission may produce similar cognitive outcomes, for different reasons.

5.2 A specific mechanism, not a general benefit

The more interesting result is the one that did not generalise. Prompt quality predicted change in inhibitory control strongly (r = −.74) and change in cognitive flexibility hardly at all (r = −.22, p = .074). Within the hierarchical model, the verification facet alone accounted for an additional fifth of the variance in inhibitory change, while the other two facets contributed nothing unique.

Under the unity-and-diversity framework (Friedman & Robbins, 2022), a general benefit operating through common executive control would have produced roughly parallel associations across components. The observed pattern is component-specific, which points instead to a targeted training effect. The plausible mechanism is straightforward. Requiring a student to withhold acceptance of a fluent, confident, immediately available answer until it has been independently checked is a response-inhibition demand, repeated many times per session. Klingbeil et al. (2024) established that over-acceptance of AI advice is the default even when it is costly; the protocol made resisting that default a graded requirement.

Read this way, the study contributes something narrower and more useful than a claim that GenAI improves thinking. It identifies which instructional requirement is doing the work. Verification is the active ingredient; contextualisation and iteration, in these data, were not.

A specificity claim of this kind is only as good as its discriminant evidence, and ours is partial. Facet scores were substantially intercorrelated, so the facets cannot be fully separated, and the design does not permit a causal claim about which facet produces which change. Dismantling designs, in which facets are manipulated independently, would be the appropriate next step.

5.3 Situating STEAM without romanticising the result

The experimental cohort produced prototypes rated far higher on cultural pertinence, and produced them without any loss of scientific rigour. The second half of that sentence matters more than the first. The cultural difference was manufactured by the instruction, but the absence of a rigour penalty was not: it indicates that grounding the engineering problem in Kichwa-Saraguro water-management practice did not come at the cost of disciplinary quality.

This addresses a specific worry in the intercultural STEM literature, that culturally situated approaches trade rigour for relevance. Reviews in this area have called for empirical evidence on that trade-off rather than assertion in either direction (Oviedo-Berrocal et al., 2026; Ogegbo & Ramnarain, 2024). Our data provide one negative result: no penalty was detected. Given the small number of artefacts, this is weak evidence, and the confidence interval for scientific rigour is wide enough to accommodate a small advantage in either direction.

We are more cautious than the framing of decolonial technology use might invite. Students supplied local constraints that the model lacked, and this inverted the usual epistemic relationship

at the level of the task. Whether it altered students' broader relationship to scientific knowledge was not measured, and we make no claim about it.

5.4 Threats to validity

Four threats require direct statement rather than relegation to a limitations list.

Condition was confounded with institution. The experimental cohort attended an intercultural bilingual school; the control cohort attended a fiscomisional school under the Hispanic modality. Every school-level variable — leadership, teaching culture, peer norms, resourcing — is perfectly confounded with treatment. Baseline equivalence on demographics and on all four BRIEF-2 scales constrains but does not eliminate this problem, because equivalence at one time point does not guarantee equivalent trajectories. This is the single strongest alternative explanation for our findings, and the design cannot rule it out. A replication with multiple schools per condition, or with classes randomised within schools, is required before causal claims are warranted.

The outcome was self-reported and students were not blind. Students in the experimental condition knew they were receiving a novel programme, and the BRIEF-2 asks them to rate their own everyday functioning. Expectancy effects are a live concern. Three features of the data argue against a purely expectancy-driven account: effects were graded rather than uniform across scales; prompt quality, an independently coded behavioural measure, tracked change on one scale specifically; and reliable individual improvement was confined to a minority. None of these is decisive. Performance-based executive tasks, or informant reports from teachers and parents, would provide the necessary triangulation, and their absence is the most serious measurement limitation of this study.

Students were nested within teams. The 138 students worked in 30 teams, and team-level processes almost certainly induce dependence. Our sensitivity analysis suggests conclusions are robust across plausible ICC values, but the point estimates' precision is overstated by the individual-level models. Multilevel analysis with directly estimated ICCs is the appropriate treatment.

Out-of-class GenAI use was not controlled. Control students were prohibited from using GenAI during class but not outside it. If unsupervised use occurred, and the offloading literature suggests it would be unstructured, this would if anything work against finding a control-group

improvement. The direction of this bias favours our hypothesis, which is a reason for caution rather than reassurance.

5.5 Further limitations

The eight-week window captures immediate post-intervention change and says nothing about durability. Executive gains from short interventions frequently fade without continued practice, and follow-up assessment at three and six months is the obvious extension. The prompt-quality rubric, although developed with expert review and scored reliably, is new and lacks external validation. The study was not pre-registered, so the distinction between confirmatory and exploratory analyses rests on our account of it. Finally, generalisation beyond Kichwa-Saraguro technical-baccalaureate cohorts with near-universal connectivity should not be assumed; the intervention depends on both cultural specificity and reliable device access, and neither travels automatically.

5.6 Implications

For practice, the finding with the clearest application is that verification behaviour is trainable and appears to be where the cognitive demand concentrates. Teachers who want students to think with GenAI rather than around it should make the audit of model output an assessed requirement, not an exhortation. A design brief that asks students to supply constraints the model cannot know, and to document what they checked and how, appears more productive than either prohibition or open access.

For policy, the results suggest that blanket restrictions on GenAI in secondary schools address the wrong variable. The relevant distinction is between structured and unstructured use, and only the former requires investment in teacher capacity. Frameworks such as UNESCO's guidance on generative AI in education (Miao & Holmes, 2023) and its teacher AI competency framework (Miao & Cukurova, 2024) provide a starting point; what is missing is evidence on which specific pedagogical moves matter, of the kind this study begins to provide.

For research, three directions follow. Dismantling studies could isolate the contribution of verification from contextualisation and iteration. Performance-based executive measures would resolve the self-report ambiguity. And the near-total absence of Indigenous and rural contexts

from the GenAI education literature is a gap that will not close by itself; the tools' training distributions make these the settings where the stakes of uncritical adoption are highest.

6. Conclusion

An eight-week intercultural STEAM unit in which generative AI was positioned as an object of verification rather than a source of answers was associated with improvements in self-reported executive functioning among upper-secondary students in Saraguro, with the largest effect on planning and the smallest on inhibitory control. Within the experimental cohort, the quality of students' verification behaviour accounted for 54% of the variance in inhibitory change, while prompt quality showed no reliable association with change in cognitive flexibility.

The specificity of that pattern is the study's main contribution. It suggests that the cognitive effects of generative AI in classrooms are not properties of the tool but of the demands placed around it, and it identifies one demand documented verification of model output as a candidate active ingredient. Because condition was confounded with school and outcomes were self-reported, these results establish plausibility rather than causal generality. They are offered as a basis for better-controlled replication, not as a settled finding

Declarations

Ethics approval. The study followed the Declaration of Helsinki and CIOMS guidance and was approved by the participating institutions as minimal-risk research. Written parental consent and student assent were obtained.

Data availability. De-identified data and analysis scripts are available from the corresponding author on reasonable request.

Competing interests. The authors declare none.

Funding. This research received no specific grant from any funding agency.

AI use statement. Generative AI tools were used by participants as the object of study. No generative AI was used to produce the analyses or the substantive content of this manuscript.

References

An, Q., Koh, J. H. L., & Liu, Q. (2026). Generative artificial intelligence in higher education: A systematic review of student use and learning outcomes. *Australasian Journal of Educational Technology*. Advance online publication. https://doi.org/10.14742/ajet.10561

Cao, X., Lu, H., Wu, Q., & Hsu, Y. (2025). Systematic review and meta-analysis of the impact of STEM education on students' learning outcomes. *Frontiers in Psychology, 16*, 1579474. https://doi.org/10.3389/fpsyg.2025.1579474

Chang, D. H., Lin, M. P.-C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability, 15*(17), 12921. https://doi.org/10.3390/su151712921

Chang, W.-L., & Sun, J. C.-Y. (2024). Evaluating AI's impact on self-regulated language learning: A systematic review. *System, 126*, 103484. https://doi.org/10.1016/j.system.2024.103484

Chen, B., Chen, J., Wang, M., Tsai, C.-C., & Kirschner, P. A. (2026). The effects of integrated STEM education on K–12 students' achievements: A meta-analysis. *Review of Educational Research*. Advance online publication. https://doi.org/10.3102/00346543251318297

Cherry Shive, E. M. (2025). Equitable integration of Indigenous knowledge systems in STEM education professional development: A systematic review. *Journal of Research in Science, Mathematics and Technology Education, 8*(2), 21–40. https://doi.org/10.31756/jrsmte.822

Chiu, T. K. F. (2024). A classification tool to foster self-regulated learning with generative artificial intelligence by applying self-determination theory: A case of ChatGPT. *Educational Technology Research and Development, 72*, 2401–2416. https://doi.org/10.1007/s11423-024-10366-w

Cong-Lem, N., Nguyen, T. T., & Nguyen, K. N. H. (2025). Critical thinking in the age of generative AI: Effects of a short-term experiential learning intervention on EFL learners. *International Journal of TESOL Studies*. https://doi.org/10.58304/ijts.250522

da Silva, C., Pereira, F., & Amorim, J. P. (2024). The integration of Indigenous knowledge in school: A systematic review. *Compare: A Journal of Comparative and International Education, 54*(7), 1210–1228. https://doi.org/10.1080/03057925.2023.2184200

Darwin, Rusdin, D., Mukminatien, N., Suryati, N., Laksmi, E. D., & Marzuki. (2024). Critical thinking in the AI era: An exploration of EFL students' perceptions, benefits, and limitations. *Cogent Education, 11*(1), 2290342. https://doi.org/10.1080/2331186X.2023.2290342

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-023-00425-2

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Friedman, N. P., & Robbins, T. W. (2022). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology, 47*(1), 72–89. https://doi.org/10.1038/s41386-021-01132-0

Fu, Y., Weng, Z., & Wang, J. (2025). Examining AI use in educational contexts: A scoping meta-review and bibliometric analysis. *International Journal of Artificial Intelligence in Education, 35*(3), 1388–1444. https://doi.org/10.1007/s40593-024-00442-w

Futterer, T., Bardach, L., Kuhn, J., Keller, S. D., & Gerjets, P. (2026). Enhancing school students' self-regulated learning through generative AI support: A randomized controlled trial. *Educational Psychology Review, 38*, Article 42. https://doi.org/10.1007/s10648-026-10133-8

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

Guo, K., Zhong, Y., Li, D., & Chu, S. K. W. (2023). Effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation. *Computers & Education, 203*, 104862. https://doi.org/10.1016/j.compedu.2023.104862

Heung, Y. M. E., & Chiu, T. K. F. (2025). How ChatGPT impacts student engagement from a systematic review and meta-analysis study. *Computers and Education: Artificial Intelligence, 8*, 100361. https://doi.org/10.1016/j.caeai.2025.100361

Hong, H., Vate-U-Lan, P., & Viriyavejakul, C. (2025). Cognitive offload instruction with generative AI: A quasi-experimental study on critical thinking gains in English writing. *Forum for Linguistic Studies, 7*(7), 325–334. https://doi.org/10.30564/fls.v7i7.10072

Juárez, F. J., & Arán Filippetti, V. (2025). Executive functions in adolescents as assessed with the BRIEF-2: Factor structure and relationship with antisocial behaviors and academic achievement. *Applied Neuropsychology: Child*. Advance online publication. https://doi.org/10.1080/21622965.2025.2601089

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI: An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior, 160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Lu, J., Zheng, R., Gong, Z., & Xu, H. (2024). Supporting teachers' professional development with generative AI: The effects on higher order thinking and self-efficacy. *IEEE Transactions on Learning Technologies, 17*, 1279–1289. https://doi.org/10.1109/TLT.2024.3369690

Miao, F., & Cukurova, M. (2024). *AI competency framework for teachers*. UNESCO. https://doi.org/10.54675/ZJTE2084

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2025). The critical role of trust in adopting AI-powered educational technology for learning: An instrument for measuring student perceptions. *Computers and Education: Artificial Intelligence, 8*, 100368. https://doi.org/10.1016/j.caeai.2025.100368

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

Ng, D. T. K., Wu, W., Leung, J. K. L., Chiu, T. K. F., & Chu, S. K. W. (2024). Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. *British Journal of Educational Technology, 55*(3), 1082–1104. https://doi.org/10.1111/bjet.13411

Ogegbo, A. A., & Ramnarain, U. (2024). A systematic review of pedagogical practices for integrating Indigenous knowledge systems in science teaching. *African Journal of Research in Mathematics, Science and Technology Education, 28*(3), 343–361. https://doi.org/10.1080/18117295.2024.2318220

Oviedo-Berrocal, L., Ortiz-Padilla, M., Zwierewicz, M., Vargas-Arteaga, J., & Gómez-García, M. (2026). STEM+intercultural: Integrating STEM education with Indigenous ancestral knowledge, traditional knowledge, and intercultural contexts. A systematic literature review. *European Journal of STEM Education, 11*(1), 25. https://doi.org/10.20897/ejsteme/18476

Portillo-Blanco, A., Deprez, H., De Cock, M., Guisasola, J., & Zuza, K. (2024). A systematic literature review of integrated STEM education: Uncovering consensus and diversity in principles and characteristics. *Education Sciences, 14*(9), 1028. https://doi.org/10.3390/educsci14091028

Qian, Y. (2025). Prompt engineering in education: A systematic review of approaches and educational applications. *Journal of Educational Computing Research*. Advance online publication. https://doi.org/10.1177/07356331251365189

Sardi, J., Darmansyah, Candra, O., Yuliana, D. F., Habibullah, Yanto, D. T. P., & Eliza, F. (2025). How generative AI influences students' self-regulated learning and critical thinking skills? A systematic review. *International Journal of Engineering Pedagogy, 15*(1), 94–108. https://doi.org/10.3991/ijep.v15i1.53379

Su, J., Ng, D. T. K., & Chu, S. K. W. (2023). Artificial intelligence (AI) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence, 4*, 100124. https://doi.org/10.1016/j.caeai.2023.100124

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic

scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yang, Y., Zhang, Y., Sun, D., He, W., & Wei, Y. (2025). Navigating the landscape of AI literacy education: Insights from a decade of research (2014–2024). *Humanities and Social Sciences Communications, 12*, 374. https://doi.org/10.1057/s41599-025-04583-8

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments, 11*(1), 28. https://doi.org/10.1186/s40561-024-00316-7