

Article

Independent Engineering Transfer After Traceable Generative-AI-Assisted Learning: A Six-University Controlled Trial with Deterministic Cluster Allocation in Agricultural Engineering Education

Yurii Syromiatnykov ^{1,2,*} , Farmon Mamatov ³ , Khurshid Chuyanov ⁴, Zafar Batirov ⁵, Dustmurod Chuyanov ⁶, Makhmatmurod Shomirzaev ⁷, Dilrabo Shadieva ⁸ , Khurshid Ilkhomov ⁹, Gulandom Jo'rayeva ⁵ and Mirshohid Egamov ⁴

- ¹ Institute of Soil and Plant Sciences, Latvia University of Life Sciences and Technologies, LV-3001 Jelgava, Latvia
 - ² Department of Fruit and Vegetable Growing and Viticulture, Institute of Agrobiotechnologies and Food Safety, Samarkand State University Named After Sharof Rashidov, Samarkand 140104, Uzbekistan
 - ³ Department of Agricultural Engineering, Faculty of Irrigation Engineering, Karshi State Technical University, Karshi 180100, Uzbekistan
 - ⁴ Department of Higher Mathematics, Faculty of Energy Engineering, Karshi State Technical University, Karshi 180100, Uzbekistan
 - ⁵ Department of Pedagogy and Teaching Methodology, Faculty of Pedagogy, University of Economics and Pedagogy (Non-State Educational Institution), 11A Jayxun Street, Gungon MFY, Karshi 180100, Uzbekistan
 - ⁶ Department of Transport and Civil Engineering, Faculty of Transport and Civil Engineering, Karshi State Technical University, Karshi 180100, Uzbekistan
 - ⁷ Department of Technological Education, Termez State University, Termez 190111, Uzbekistan
 - ⁸ Department of Russian Linguistics, Termez State University, Termez 190111, Uzbekistan
 - ⁹ Department of Physics and Technological Education, Bukhara State Pedagogical Institute, Bukhara 200117, Uzbekistan
- * Correspondence: yuriysyromyatnikov@gmail.com

Abstract

Generative artificial intelligence (GenAI) can support engineering problem solving, but whether AI-assisted practice transfers to independent performance after the tool is removed remains unclear. This multicentre controlled trial evaluated a traceable five-stage GenAI-assisted learning configuration in agricultural engineering education. Twenty-eight second- and third-year classes from six universities in Uzbekistan were assigned within nine teacher blocks by deterministic constrained minimization to GenAI (14 classes) or structured active-control (14 classes) groups. Both groups completed the same 16-week module, tasks, software, contact time, feedback, and verification requirements. They differed in the source and adaptivity of a provisional alternative used after an independent attempt: bounded adaptive GenAI dialogue versus a version-controlled curated alternative. The full assigned cohort included 656 students; likelihood-based available-outcome analyses included 641 immediate and 589 delayed outcomes. Kenward–Roger analyses estimated an adjusted immediate difference of 2.78 points (95% confidence interval (CI) [2.02, 3.55]; $p < 0.001$; model-based $d = 0.72$) and a delayed difference of 2.34 points (95% CI [1.54, 3.14]; $p < 0.001$; $d = 0.51$). The results show a positive adjusted association for the evaluated traceable instructional configuration, but deterministic post-baseline allocation limits causal interpretation.



Academic Editors: Igor Mekterović,
Ivana Bosnić and Ljiljana Brkić

Received: 9 August 2026

Revised: 5 September 2026

Accepted: 8 September 2026

Published: 9 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: generative artificial intelligence; engineering education; human–AI collaboration; deterministic cluster allocation; independent transfer; verification; active control; agricultural engineering

1. Introduction

1.1. *Generative AI, Engineering Judgement, and Independent Transfer*

Large language models can generate technically plausible explanations, calculations, code, and recommendations within seconds. In higher education, this creates an unusual combination of opportunity and risk. Generative artificial intelligence (GenAI) can provide rapid feedback and alternative representations, yet its outputs may be incorrect, weakly justified, or accepted without sufficient scrutiny [1–4]. These concerns are especially important in engineering, where a numerically coherent answer may remain dimensionally inconsistent, operationally infeasible, agronomically inappropriate, or unsafe.

Empirical findings do not support a single conclusion about educational benefit. Controlled studies have reported improved learning gains, higher-order thinking, and feedback effectiveness under structured implementations [5–7], whereas other experiments found that unrestricted or weakly scaffolded access could reduce independent performance or encourage cognitive offloading [8,9]. Carefully generated AI help has also produced gains comparable to human-authored help in some mathematics settings [10]. This variation suggests that the relevant explanatory unit is not GenAI access alone but the configuration of timing, scaffolding, adaptivity, feedback, and subsequent tool-free assessment.

This interpretation is consistent with active-learning, cognitive-engagement, and formative-feedback research. The interactive, constructive, active, and passive (ICAP) framework predicts stronger learning when students construct, explain, test, and revise ideas instead of only receiving finished solutions [11]. Active learning improves outcomes when learners are required to participate in disciplinary reasoning [12], and formative feedback is most useful when it supports self-regulation and closes a specific gap between current and desired performance [13]. Applied to GenAI, these perspectives predict benefits when an independent attempt precedes assistance and a generated output becomes an object for comparison and verification; they predict a weaker transfer when fluent output is substituted for these activities.

1.2. *Traceable GenAI-Assisted Learning as an Applied Engineering Workflow*

AI literacy has been defined broadly as the capacity to understand, use, critically evaluate, and communicate or collaborate with AI while recognizing its ethical implications [14,15]. In professional engineering education, however, these capabilities must become observable technical practices: defining the problem before assistance, documenting the source of an alternative, challenging assumptions, independently checking the result, and taking responsibility for the accepted decision. The intervention therefore treated GenAI literacy as an auditable engineering workflow rather than as interface familiarity.

The five-stage workflow targeted six interconnected capabilities: problem-formulation literacy, prompt and interaction literacy, provenance literacy, critical GenAI literacy, verification literacy, and professional accountability. These labels describe intended processes embedded in the intervention; they were not separately manipulated mechanisms or independently validated outcome scales. The applied endpoint was condition-neutral engineering transfer after the tool had been removed, so the study evaluated the complete instructional configuration rather than any single literacy component. Table 1 summarizes

how the five workflow stages correspond to the intended capabilities and the evidence required from students.

Table 1. Operational new literacy framework embedded in the five-stage human–generative artificial intelligence (GenAI) workflow.

Workflow Stage	Targeted Capability	Required Evidence
1. Independent formulation	Problem-formulation literacy and learner agency	Variables, units, assumptions, constraints, expected range, and initial strategy recorded before GenAI access
2. Provisional alternative	Prompt and interaction literacy	Prompt–response record and identification of the purpose for which GenAI support was requested
3. Critical comparison	Critical GenAI literacy	Comparison of assumptions, units, signs, boundary conditions, omissions, and alternative methods
4. Independent verification	Verification literacy	Analytical, numerical, graphical, or simulation-based check in a separate environment
5. Professional decision	Provenance and professional accountability	Documented acceptance, correction, or rejection of the alternative and a defensible engineering recommendation

1.3. Research Gap, Aim, and Hypotheses

Previous engineering-education research has provided useful survey, qualitative, and controlled course evidence [5,16–18]. Three linked gaps remain. First, many studies measure performance while the tool remains available rather than independent transfer after tool removal. Second, intervention descriptions often combine access, adaptivity, prompting, feedback, and verification without making the instructional configuration auditable. Third, multicentre class-level comparative evidence, with an active comparison condition and delayed assessment, remains limited. These gaps motivate an evaluation of a traceable configuration in which assistance is delayed until after an independent attempt and every provisional alternative must be challenged and verified.

The present study addressed these gaps within a common Engineering Modelling, Verification, and Decision-Making in Agricultural Systems module. Every substantive task began with an independent attempt, followed by exposure to a provisional alternative, critical comparison, independent verification, and a documented professional decision. The active-control condition followed the same sequence but used a version-controlled teacher/textbook alternative. Consequently, the contrast estimates the effect of the complete GenAI-assisted configuration relative to the matched active-control configuration; it does not isolate the GenAI model from adaptivity or other bundled features.

The primary research question (RQ1) was whether classes assigned to the traceable GenAI-assisted configuration would show higher supervised, condition-neutral immediate engineering-transfer performance than classes assigned to the structured active control after covariate adjustment. The secondary research question (RQ2) was whether the adjusted between-condition difference would remain evident at a delayed follow-up. Derived from the engagement and feedback framework above, Hypothesis 1 (H1) predicted a positive adjusted GenAI-minus-control difference on immediate Form A, and Hypothesis 2 (H2) predicted a positive adjusted difference on delayed Form B. These are model-based comparative hypotheses rather than randomization-based causal hypotheses because final class allocation used deterministic post-baseline minimization. Exploratory questions

concerned the study year, baseline performance, attendance, implementation fidelity, and stability to class omissions and missing-data assumptions.

2. Materials and Methods

2.1. Study Design and Setting

A multicentre, two-arm, parallel controlled trial was conducted across 28 intact classes nested within nine teacher blocks at six universities in Uzbekistan: Karshi State Technical University; Samarkand State University of Veterinary Medicine, Livestock and Biotechnologies; Termez State University of Engineering and Agrotechnologies; Tashkent State Agrarian University; Andijan Institute of Agriculture and Agrotechnology; and TIAME National Research University. Universities were purposively selected for an accredited agricultural-mechanization programme, institutional willingness, regional diversity, and capacity to provide computer and internet access. Nine volunteer instructors were eligible if they taught agricultural engineering subjects, had at least three years of teaching experience, and agreed to the standardized workflow training. All eligible intact second- and third-year classes taught by participating instructors were recruited consecutively during September–October 2023 until the planned 28 classes were enrolled. Class formation and participant consent were completed before allocation. The class was the unit of allocation and the student was the unit of observation. The module ran from 3 February to 19 May 2025; Form A was administered on 26 May 2025 and Form B on 8 September 2025.

The intervention description was organized using the Template for Intervention Description and Replication framework [19]. Relevant items from the cluster-trial Consolidated Standards of Reporting Trials (CONSORT) extension [20] and CONSORT 2025 [21] were used as transparency checks for participant flow, clustering, allocation, and analysis; this use does not characterize the deterministic allocation as randomized. Protocol elements were considered against the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) 2025 [22], while the analysis plan followed established guidance on prespecified models, estimands, missing data, and sensitivity analyses [23,24].

2.2. Participants, Consent, Ethical Oversight, and Registration

Eligible participants were consenting students aged 18 years or older who were enrolled in participating second- or third-year classes. Research participation was separated from academic grading, and coded Student_ID values were used in analytical files. The full assigned cohort included 656 eligible and consenting students.

The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Commission of Karshi Engineering-Economics Institute, subsequently reorganized as Karshi State Technical University (Protocol No. 14/2023, 28 August 2023). No formal multicentre ethics-reliance mechanism or single-institutional review board (IRB) arrangement was executed. Ethical governance therefore consisted of the coordinating centre's approval under Protocol No. 14/2023, together with site-specific administrative authorization, rather than six independent ethics approvals. Before study activities, each university provided an official administrative permission letter signed by the rector or vice-rector for research, confirming institutional participation, access to student groups, assessment and interview procedures, and confidentiality requirements. These letters were obtained in August–September 2023 and are retained in the coordinating centre archive for confidential review on request. Written informed consent was obtained before research participation.

The educational study was not prospectively registered in a public trial registry. The date-stamped study protocol (No. 14/2023, 28 August 2023) preceded data collection. Statistical Analysis Plan (SAP) version 1.0 was dated 15 January 2024 and preceded database lock

on 6 August 2026; it specified the endpoints, restricted maximum likelihood (REML) mixed analysis of covariance (ANCOVA), Kenward–Roger inference, the 5% primary-outcome imputation trigger, intention-to-treat and per-protocol definitions, and assessor masking. The historical documents used the terms randomization and intention-to-treat; the manuscript clarifies that final class assignment was a deterministic constrained minimization and therefore uses the full assigned cohort and assigned-group analyses. Signed originals are retained by the coordinating centre and will be provided only upon a substantiated request from the editor or reviewers; they are not part of the public Supplementary Files S1–S8 package. The absence of prospective public registration remains a limitation.

2.3. Sample-Size and Precision Rationale

The design target was fixed at 28 intact classes and balanced at 14 classes per condition. Precision was considered at the class-allocation level rather than from the nominal number of individual students alone. The final full assigned cohort comprised 656 students.

Planning inputs were an individual-level standard deviation of 5.2 points from pilot work, an intraclass correlation of 0.05, a mean class size of 23.4 students, and a coefficient of variation in class sizes of 0.04, together with baseline adjustment in the planned ANCOVA. Because ANCOVA precision depends on baseline–outcome covariance in addition to clustering inputs, the clustering assumptions alone are not used here to reconstruct a standalone expected standard error retrospectively. Achieved model-based precision is therefore reported separately; with 28 allocated classes and 641 observed immediate outcomes, the Kenward–Roger standard error for the primary adjusted contrast was 0.365 points, corresponding to a 95% confidence-interval half-width of approximately 0.77 points. Anticipated loss to delayed follow-up was 10%, and no interim analyses were planned.

2.4. Baseline Assessment, Constrained Minimization, and Allocation Implementation

Before allocation, all students completed the supervised condition-neutral Form P assessment. The admissible set was restricted by hard constraints requiring 14 classes per condition, seven second-year and seven third-year classes per condition, representation of both conditions within each teacher block, and an absolute within-institution class-count difference of no more than one. Enumeration under these constraints yielded 75,288 admissible allocations for the submitted 28-class structure.

An independent statistician implemented allocation in Python after baseline data collection. The algorithm was initialized using the fixed seed 20230828, corresponding to the ethics-protocol date (28 August 2023), and selected from the 75,288 admissible allocations those that minimized imbalance in mean baseline score, sex composition, age, and class schedule. Thus, final assignment was a deterministic constrained minimization after observation of baseline covariates and not random sampling from the admissible set. The selected allocation, inputs, code, seed, and Secure Hash Algorithm 256-bit (SHA-256) hashes are retained in a protected audit log held by the study coordinator and are not included in the public supplement. Assignments were communicated only after baseline assessment and consent; no students were individually allocated.

2.5. Common Instructional Module and Instructor Standardization

All classes completed the 16-week Engineering Modelling, Verification, and Decision-Making in Agricultural Systems module. The planned dose was 64 contact hours and 32 laboratory hours. Weekly topics included problem formulation, dimensional checks, model construction, numerical implementation, machinery performance, power balance, energy use, optimization, sensor data, uncertainty, drone and Internet of Things planning, and an integrated capstone.

Each substantive task followed five stages. First, the student defined the variables, units, assumptions, constraints, expected range, and an initial solution strategy. Second, the assigned condition supplied a provisional alternative. Third, the student compared the initial and provisional approaches using a common checklist. Fourth, the result was verified in an independent analytical or computational environment. Fifth, the student accepted, corrected, or rejected the result and stated a defensible engineering recommendation.

Instructors attended a standardized training workshop covering the five-stage workflow, condition-specific alternative sources, troubleshooting, and fidelity observation. A common instructor manual, lesson plans, tasks, feedback procedures, and assessments were provided exclusively in Uzbek. Although the Microsoft 365 Copilot Chat interface (web service; GPT-4o backbone dated 6 August 2024) supported Uzbek, Russian, and English, the educational process in both conditions was conducted only in Uzbek.

2.6. GenAI and Structured Active-Control Configurations

In the GenAI condition, the provisional alternative was generated using an institution-approved GenAI service after the independent attempt had been recorded. Complete prompt–response exchanges, timestamps, model/version information, and tool-use metadata were retained. The workflow permitted no more than four prompt–response turns for the condition-specific alternative. Generated content was treated as provisional and could not be accepted without critical comparison and a separate verification record.

The GenAI condition used Microsoft 365 Copilot Chat in Microsoft Teams (Microsoft Corporation, Redmond, WA, USA) with GPT-4o. Microsoft did not disclose the provider build; the institutional configuration register identifies the GPT-4o backbone dated as 6 August 2024. Access occurred from 3 February to 19 May 2025 through the Microsoft Teams institutional interface under Microsoft 365 Education A3 licences. The interface supported Uzbek, Russian, and English. Web search, file upload, code execution, cross-session memory, institutional retrieval, and image tools were disabled by administrative policy. Each task began with an empty context and retained only the current four-turn dialogue. Microsoft 365 commercial data protection excluded conversation data from model training, and audit retention followed the institutional 30-day default.

In the active-control condition, students received a matched alternative from a version-controlled teacher/textbook bank. Alternatives were matched to the same task, intended learning outcome, approximate length, complexity, and critique demand. Up to three predefined clarifications were permitted. Tasks, instructional time, disciplinary software, portfolio requirements, instructor feedback, and verification criteria were otherwise matched.

The two conditions were therefore designed to be equivalent in curricular content, assessment demands, contact time, and the obligation to compare and verify a provisional solution. They were not perfectly equivalent in adaptivity: GenAI could respond to a student's prompt within the four-turn limit, whereas the control offered predefined clarifications. Accordingly, the trial estimates the effect of assignment to the complete traceable GenAI configuration compared with the complete structured active-control configuration.

2.7. Privacy, Data Governance, Academic Integrity, and Access Equity

Coded Student_ID values were used in analytical files, and the assessment outcomes were collected under supervised conditions. Prompt–response logs and tool-use metadata were retained separately as intervention-process records. The pedagogical design required disclosure of the prompt, generated alternative, comparison, verification evidence, and final decision; this shifted assessment from possession of a polished answer toward the quality of the documented reasoning chain. The approach also followed a human-centred

principle in which GenAI supported, but did not replace, student agency and professional accountability [25].

Students were instructed not to enter personal identifiers or confidential institutional data into prompts. Prompt–response logs were stored on an institutional encrypted server, with access restricted to the research team. Data were encrypted at rest and in transit. Retention and deletion procedures followed institutional data-protection policies. Cross-border transfer was governed by institutional standard contractual clauses. Hallucinated or unsafe advice was addressed through the mandatory independent-verification stage. Academic integrity rules prohibited direct submission of generated output without documented provenance and verification. Students without paid-tool access were provided institutional accounts, and no students were excluded for connectivity reasons.

2.8. Outcomes and Scoring

The primary outcome was the total score on Form A, a supervised condition-neutral engineering-transfer assessment completed without GenAI, internet access, personal notes, laptops, or communication. The delayed outcome was the parallel Form B assessment. Forms P, A, and B were developed by the author group from the common eight-criterion engineering-transfer rubric (0–4 per criterion; total 0–32) as parallel variants containing unfamiliar integrated agricultural engineering scenarios. The forms and rubric were examined by two independent trained assessors and then evaluated in the crossover pilot described in Section 2.9. This process yielded evidence for score reliability and cross-form comparability, but no separate formal content-validity index or construct-validation study was conducted. The available study records do not document a separate formal subject-matter expert panel or content-validity exercise beyond the two trained assessors and crossover pilot.

Two trained assessors independently scored each criterion from 0 to 4. Before scoring, each student script was assigned a unique code containing no condition or institution information. Condition and institution fields were removed from the scoring files and were re-linked only for statistical analysis after all ratings were complete. The assessors did not have access to class assignments or institutions. The script order was not randomized. Adjudication was required when any criterion differed by more than two points or the total scores differed by more than four points; otherwise, the final criterion score was the arithmetic mean of the two ratings.

Assessors completed training with anchor scripts illustrating scores 0–4 across all eight criteria. Calibration was monitored using quadratic weighted kappa. Instruction, materials, tasks, Forms P/A/B, feedback, and scoring were all conducted in Uzbek; parallel translated assessment versions were not used.

2.9. Parallel Form Pilot and Measurement Interpretation

Forms P, A, and B were evaluated in a 144-participant, six-sequence crossover pilot with 432 administrations. The intraclass correlation coefficient [ICC(A,2)] was 0.992 (95% confidence interval (CI) 0.990–0.993), median quadratic weighted κ was 0.936, and minimum within-one agreement was 99.54%. Form P and Form B satisfied the prespecified ± 1.5 -point cross-form tolerance, whereas P–A and A–B did not. These findings support scoring reliability and identify limits to raw cross-form comparability; because no separate formal content-validity index or construct-validation study was conducted, they do not, by themselves, establish content or construct validity.

The ± 1.5 -point threshold was specified before the pilot as an administrative tolerance for deciding whether raw totals from parallel forms could be treated as approximately interchangeable. It was not derived from an empirical minimal clinically or educationally

important difference and should not be interpreted as a validated pedagogical threshold. Because Form A did not satisfy this tolerance against Forms P and B, the confirmatory analysis compared conditions at each outcome while adjusting for Form P rather than interpreting raw within-person change across forms.

2.10. Attendance, Fidelity, Contamination, Documented Events, and Missing Data

Attendance was recorded for all 16 weeks. Session fidelity was calculated from six required elements: independent attempt, correct alternative source, critical comparison, independent verification, professional interpretation, and the required trace log. Class fidelity was computed as the percentage of required elements completed across all delivered sessions (denominator = 96 elements: 6 elements $\times$ 16 weeks). The per-protocol population required an attendance of at least 80%, a class fidelity of at least 80%, and no participant-level contamination flags. All deviations, contamination events, attendance issues, and data queries were retained in coded event logs.

The full assigned cohort comprised all 656 consented and eligible students allocated with their class. The primary and delayed mixed models preserved assigned condition and used likelihood contributions from participants with the relevant observed outcome (641 immediate; 589 delayed), without single-value imputation. The term intention-to-treat has therefore been removed, as these are assigned-group, available-outcome analyses under missing-at-random interpretations that are conditional on the conditions and prespecified covariates. The 5% trigger for multilevel multiple imputation applied to the primary outcome and was not reached. Because delayed missingness was approximately 10%, a bounded deterministic pattern–mixture stress test shifted imputed outcomes adversely for missing GenAI participants and favourably for missing control participants over differential shifts from 0 to 24 points.

Participant-level contamination flags were recorded from routine weekly fidelity records: eight in GenAI and six in active control. These 14 flags were distinct from the central event register, which summarized five separately adjudicated incidents, including two events classified as major. A participant-level flag could therefore exist without a separate EVT number. The per-protocol definition excluded participants with a contamination flag, attendance below 80%, class fidelity below 80%, or no immediate outcome; the primary assigned-group analysis retained these whenever the outcome was observed.

2.11. Statistical Analysis and Reproducibility

The primary target contrast was the participant-average adjusted mean difference in immediate Form A score—the GenAI-assisted configuration minus the active-control configuration—in the full assigned cohort. The fitted likelihood used observed immediate outcomes while retaining original assignment and prespecified covariates. The prespecified participant-level model is given in Equation (1).

$$Y_{ij} = \beta_0 + \beta_1 \text{Condition}_j + \beta_2 (X_{ij} - \bar{X}_j) + \beta_3 \bar{X}_j + \beta_4 \text{StudyYear}_j + \sum_{k=1}^8 \delta_k \text{Teacher}_{kj} + u_j + \varepsilon_{ij} \quad (1)$$

In this expression, Y_{ij} is the engineering-transfer score for student i in class j ; Condition_j denotes the assigned condition; X_{ij} and $\bar{X}_j$ are the individual baseline score and class baseline mean, respectively; StudyYear_j is the study-year indicator; Teacher_{kj} represents the eight teacher-block indicators; β_0 – β_4 and δ_k are the fixed-effect coefficients; u_j is the class random intercept; and ε_{ij} is the residual error. Individual and class baseline terms were separated to avoid conflating within- and between-class associations [26]. Because allocation was deterministically optimized using observed baseline covariates, model-based

confidence intervals quantify conditional sampling uncertainty under the fitted model, not randomization-based causal uncertainty.

The model was estimated by restricted maximum likelihood. Kenward–Roger covariance adjustment and denominator degrees of freedom were calculated for the condition contrast using the published matrix formulation [27]. A source-aligned Python implementation independently estimated the REML variance components, reconstructed the marginal covariance matrix, applied the adjusted fixed-effect covariance, and calculated the one-degree-of-freedom Kenward–Roger test. The raw adjusted mean difference was the primary outcome, whereas the model-based standardized effects and intraclass correlations were supportive.

The class-level sensitivity model regressed class post-test means on condition, class baseline mean, study year, and teacher-block indicators, weighted by observed class size. Additional analyses included per-protocol institution fixed effects; leave-one-class-out models; a supportive model adding age and sex, as these measured variables entered the allocation criterion; a supportive attendance-adjusted model; and attendance-threshold analyses at 70%, 75%, and 80%. Class schedules also entered allocation but were not present in the locked participant-level analysis file. For delayed missingness, the bounded pattern–mixture stress test imputed from a baseline-and-covariate model and then imposed progressively adverse between-condition shifts between missing outcomes, truncating scores to 0–32. This deterministic analysis is a stress test rather than formal multiple imputation, so completed-data intervals do not incorporate imputation uncertainty. Standardized residual summaries, outlier counts, optimizer messages, and variance components were also examined.

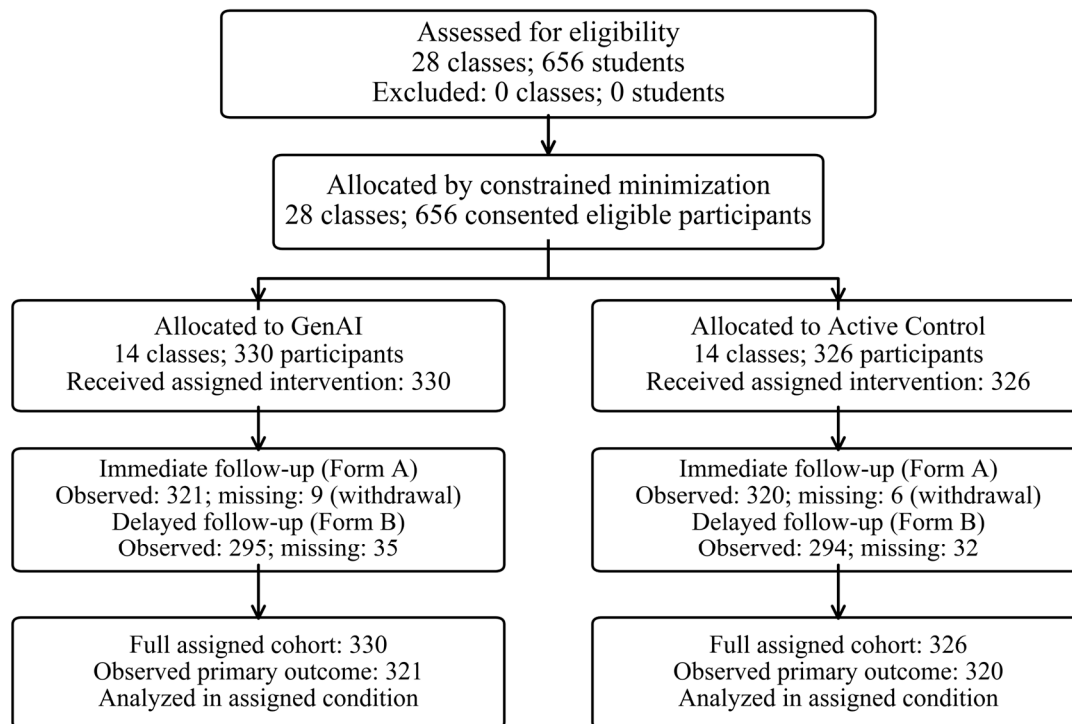
The database was locked on 6 August 2026. The locked CSV SHA-256 was 73f16dd5fa63cf95beac052506822f1cbfdeb2281db0610bd96046d9f0237940. The source-aligned Kenward–Roger implementation (version 1.0.0) was executed in Python 3.13.5 (Python Software Foundation, Wilmington, DE, USA) on Linux using NumPy 2.3.5, pandas 2.3.1, SciPy 1.16.0, and statsmodels 0.14.5. The submitted reproducibility package contains S3_KR_v4.py, S4_Analysis_v4.py, S5_Confirm_v4.R, the corrected S2 workbook, the locked database, and a SHA-256 manifest.

3. Results

3.1. Cluster and Participant Flow

Figure 1 presents cluster and participant flow. All students assessed for eligibility who provided written informed consent and met the age and enrolment criteria were retained. Deterministic constrained minimization allocated 28 classes (14 GenAI, 14 active controls), comprising 656 consented eligible participants in the full assigned cohort. Immediate missingness was 2.3% overall (2.7% in GenAI and 1.8% in active controls; withdrawal only). Delayed missingness was 10.2% (absence, medical reasons, transfer, and withdrawal).

The two conditions had similar class counts, study-year compositions, baseline transfer scores, ages, grade point averages (GPAs), and sex distributions, as summarized in Table 2. These similarities partly reflect the allocation criterion and are descriptive; they should not be interpreted as evidence of successful randomization.



Overall: full assigned cohort $n = 656$; baseline $n = 656$; immediate outcome $n = 641$; delayed outcome $n = 589$

Figure 1. Cluster and participant flow through deterministic constrained allocation. GenAI = generative artificial intelligence.

Table 2. Baseline characteristics by assigned conditions.

Characteristic	GenAI	Active Control	Overall
Students, n	330	326	656
Classes, n	14	14	28
Baseline score, mean (SD)	16.26 (4.98)	15.85 (5.43)	16.06 (5.21)
Age, years, mean (SD)	20.97 (1.48)	20.84 (1.43)	20.91 (1.46)
Baseline GPA, mean (SD)	3.29 (0.44)	3.21 (0.43)	3.25 (0.44)
Female, n (%)	163 (49.4%)	176 (54.0%)	339 (51.7%)
Male, n (%)	161 (48.8%)	146 (44.8%)	307 (46.8%)
Prefer not to say, n (%)	6 (1.8%)	4 (1.2%)	10 (1.5%)
Study year 2, n (%)	165 (50.0%)	166 (50.9%)	331 (50.5%)
Study year 3, n (%)	165 (50.0%)	160 (49.1%)	325 (49.5%)

Note. GPA = grade point average; SD = standard deviation.

3.2. Immediate and Delayed Engineering Transfer

The adjusted immediate mean was 18.77 in GenAI and 15.98 in active controls. The Kenward–Roger adjusted difference was 2.78 points (adjusted SE = 0.365; 95% CI [2.02, 3.55]; $df = 19.56$; $p < 0.001$). The model-based standardized effect was 0.72, and the estimated ICC was 0.014. On the original metric, 2.78 points equals 8.7% of the 32-point scale; if distributed evenly across all eight rubric criteria, it corresponds to approximately 0.35 point per 0–4 criterion. Criterion-specific effects were not modelled, so the latter is only a scale translation rather than evidence that every criterion improved equally.

At delayed follow-up, the corresponding adjusted means were 18.61 and 16.27. The Kenward–Roger adjusted difference was 2.34 points (adjusted SE = 0.395; 95% CI [1.54, 3.14]; $df = 36.23$; $p < 0.001$; model-based $d = 0.51$). The estimated class variance was effectively zero. The boundary estimate and the optimizer precision-loss message are

reported transparently; the class-level ANCOVA gave a closely similar estimate and interval, supporting the stability of the fixed condition contrast without implying meaningful delayed between-class variance. Table 3 brings together the adjusted means, between-condition contrasts, confidence intervals, standardized effects, and ICC estimates for the immediate and delayed assessments.

Table 3. Adjusted immediate and delayed engineering-transfer outcomes.

Outcome	Adjusted GenAI Mean	Adjusted Control Mean	Difference	Kenward–Roger 95% CI	Model-Based <i>d</i>	ICC
Immediate transfer (primary)	18.77	15.98	2.78	[2.02, 3.55]; <i>df</i> = 19.56	0.72	0.0143
Delayed transfer	18.61	16.27	2.34	[1.54, 3.14]; <i>df</i> = 36.23	0.51	<0.0001

Note. CI = confidence interval; ICC = intraclass correlation coefficient.

Figure 2 summarizes the primary, delayed, class-level, per-protocol, institution-adjusted, allocation-covariate-adjusted, attendance-adjusted, and leave-one-class-out estimates in a single compact forest display.

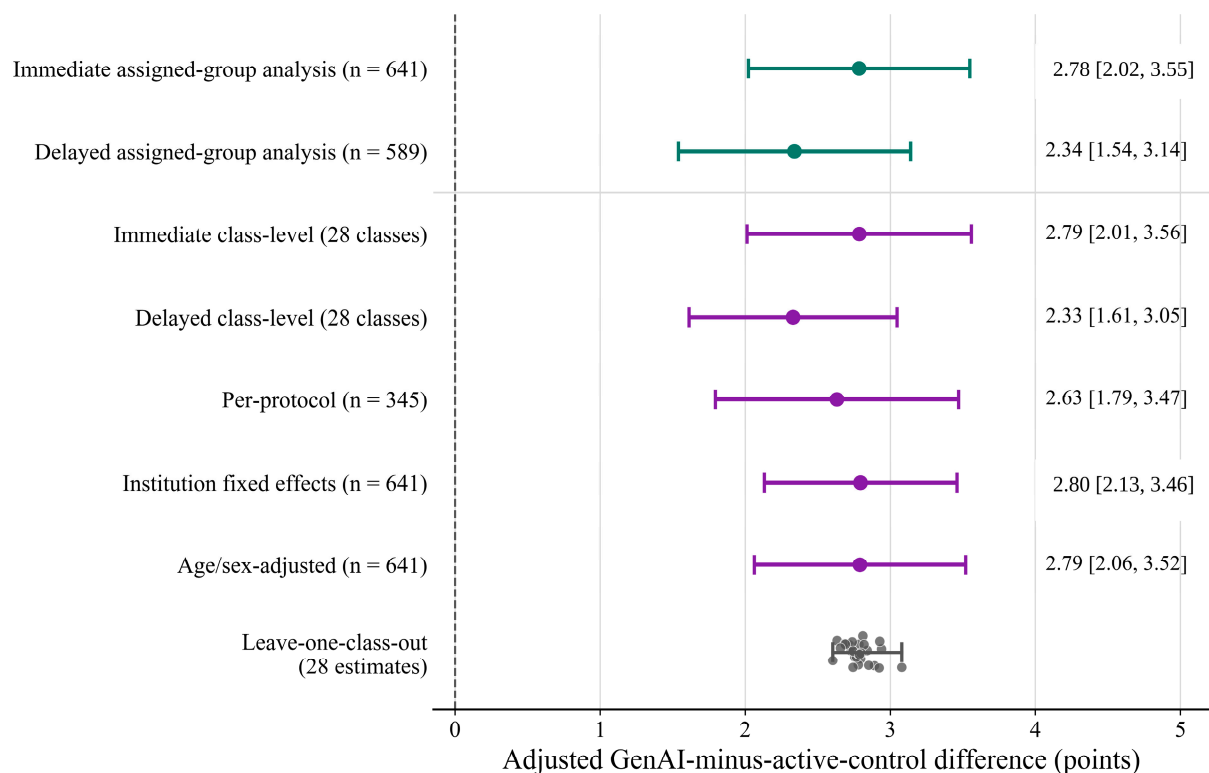


Figure 2. Adjusted GenAI-minus-active-control contrasts and robustness analyses. Points and horizontal bars show adjusted mean differences and capped model-based 95% confidence intervals; teal points show the immediate and delayed assigned-group estimates; purple points show class-level and supportive estimates; grey points show all 28 leave-one-class-out estimates. The dashed vertical line marks zero.

3.3. Robustness, Time, and Exploratory Effect Modification

The weighted class-level ANCOVA closely reproduced the primary estimate: 2.79 points (95% CI [2.01, 3.56]; $p < 0.001$). The delayed class-level estimate was 2.33 points (95% CI [1.61, 3.05]; $p < 0.001$). Adding age and sex—the measured individual characteristics used in the allocation criterion—to the immediate model gave 2.79 points (95% CI [2.06, 3.52]); the class schedule could not be added because it was not included in the

locked analysis file. The per-protocol analysis included 345 students with an observed immediate outcome after applying the prespecified criteria; its estimate was 2.63 points (95% CI [1.79, 3.47]). The reduction from 641 to 345 was driven principally by the attendance criterion: 287 students had an attendance below 80%. Attendance-threshold analyses remained similar at 70% and 75% ($n = 417$; 2.68 points, 95% CI [1.85, 3.52]) and 80% ($n = 345$; 2.63 points, 95% CI [1.79, 3.47]). A supportive model adjusting for continuous mean attendance estimated 2.76 points (95% CI [2.05, 3.46]). These analyses support robustness but do not eliminate confounding from deterministic allocation or selection in the per-protocol subset.

Replacing teacher-block indicators with institution fixed effects produced an estimate of 2.80 points (95% CI [2.13, 3.46]). Across 28 leave-one-class-out analyses, the condition estimate ranged from 2.60 to 3.08 points and remained positive in every model. The condition-by-time interaction was -0.38 points ($p = 0.426$), providing no statistical evidence that the between-condition contrast changed materially between the immediate and delayed follow-up. This interaction was supportive because different parallel forms were used at the two occasions.

The study-year interaction was not statistically evident ($p = 0.619$). Interactions for the middle and high baseline tertiles were also not statistically evident ($p = 0.768$ and $p = 0.977$, respectively). These subgroup analyses were exploratory and were not interpreted as proof of equal effects. Table 4 summarizes the prespecified sensitivity analyses and the additional supportive models used to examine the stability of the main result.

Table 4. Prespecified supportive and sensitivity analyses.

Analysis	Students	Classes	Difference	95% CI	Role
Primary class-level ANCOVA	641	28	2.79	[2.01, 3.56]	Prespecified sensitivity
Delayed class-level ANCOVA	589	28	2.33	[1.61, 3.05]	Prespecified sensitivity
Per-protocol mixed model	345	28	2.63	[1.79, 3.47]	Supportive only
Institution fixed effects	641	28	2.80	[2.13, 3.46]	Exploratory
Leave-one-class-out	—	27 per model	2.60–3.08	Point-estimate range	Influence analysis
Attendance-adjusted mixed model	641	28	2.76	[2.05, 3.46]	Supportive
Age/sex-adjusted mixed model	641	28	2.79	[2.06, 3.52]	Supportive
Attendance threshold 70% and 75%	417	28	2.68	[1.85, 3.52]	Supportive

Figure 3 shows participant-level score distributions at baseline, immediate assessment, and delayed follow-up. The largest between-condition separation emerged at the immediate assessment and remained visible at delayed follow-up, while baseline distributions overlapped closely.

Figure 4 presents the association between class-level baseline and immediate outcome means separately by their assigned conditions. The revised display uses condition-specific descriptive fitted trends and does not include a $y = x$ no-change reference, as Forms P and A did not satisfy the prespecified cross-form comparability tolerance.

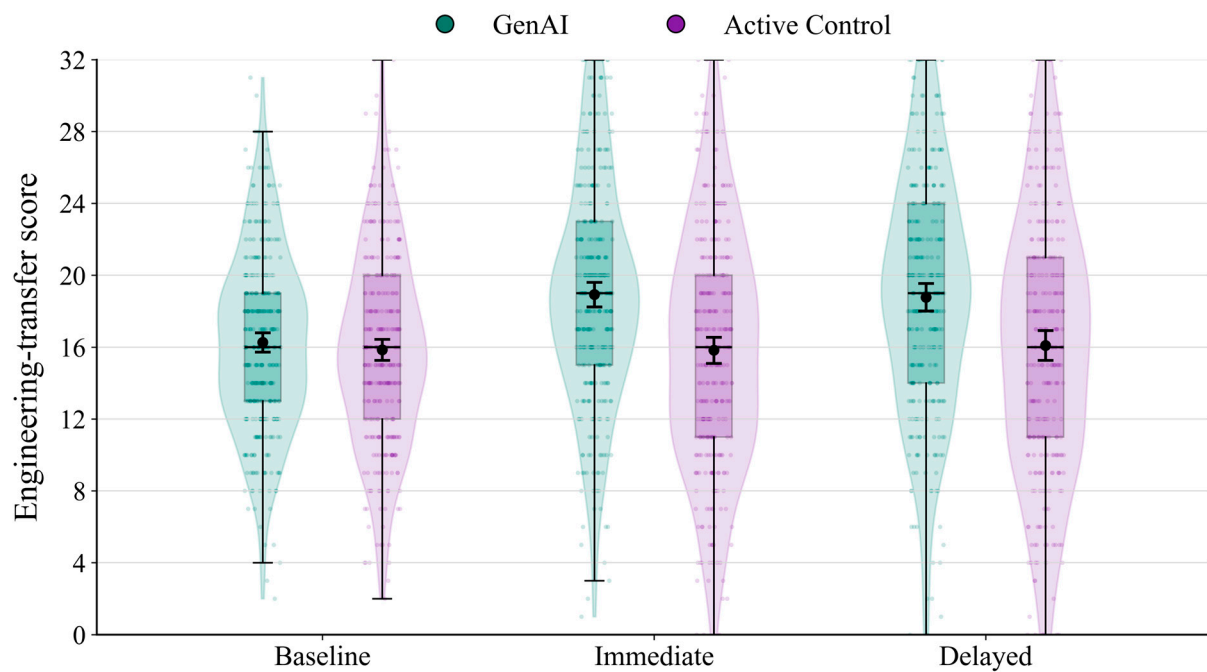


Figure 3. Participant-level engineering-transfer score distributions at baseline, immediate assessment, and delayed follow-up by their assigned conditions. Violin envelopes show score density; boxplots show medians and interquartile ranges; and black points and error bars denote group means with 95% confidence intervals.

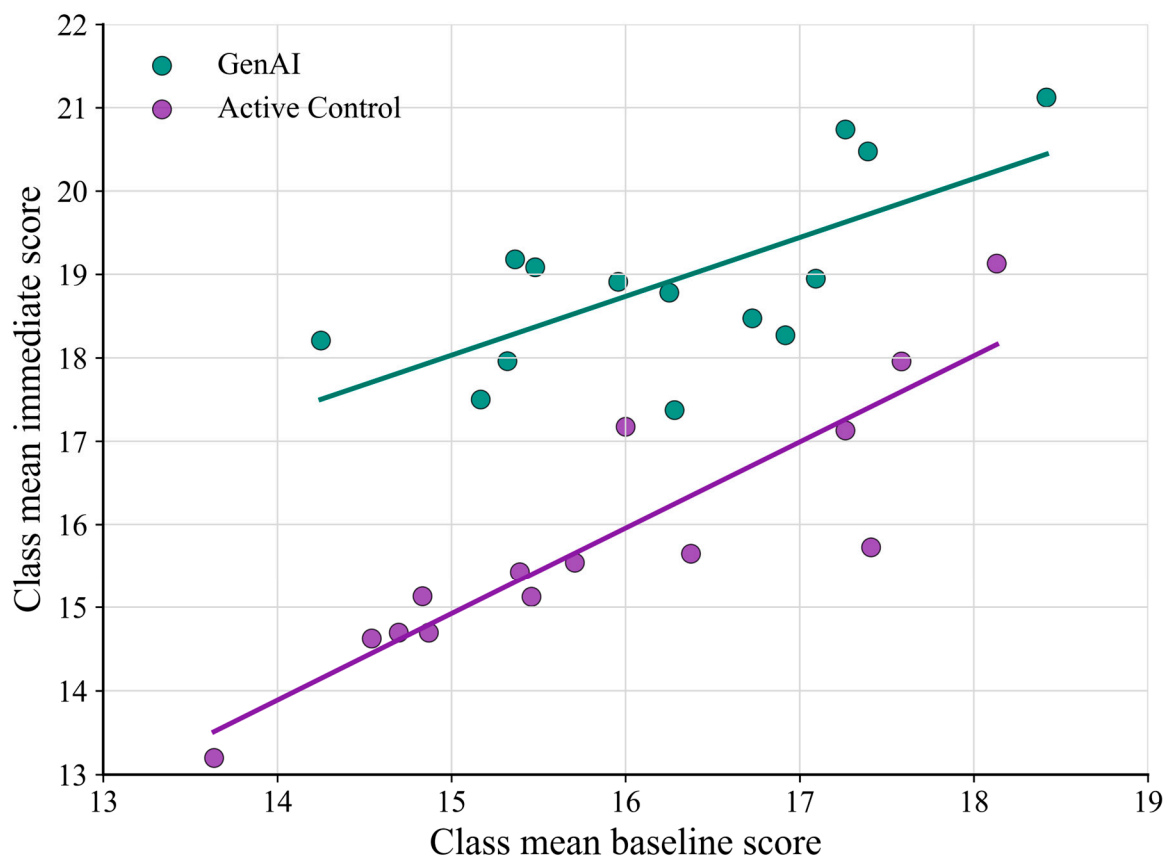


Figure 4. Class-level baseline and immediate engineering-transfer scores by their assigned conditions. Points represent classes; solid lines show within-condition descriptive fitted trends.

3.4. Attendance, Fidelity, Contamination, Missing Outcomes, and Events

The mean attendance was 78.48% for GenAI and 76.82% for the active controls. The corresponding mean class fidelities were 84.88% and 85.03%. Attendance of at least 80% was achieved by 179 GenAI students and 183 active-control students. Thus, approximately 45% of participants fell below the prespecified attendance threshold, explaining most of the difference between the available-outcome cohort and the per-protocol analysis. Participant-level contamination flags were recorded for eight GenAI and six active-control participants; these flags were used for the per-protocol definition but did not exclude observed outcomes from the primary assigned-group analysis. Table 5 summarizes attendance, implementation fidelity, and contamination by their assigned conditions.

Table 5. Attendance, implementation fidelity, and participant-level contamination flags by their assigned conditions.

Condition	Students	Classes	Attendance, Mean (SD)	Attendance $\geq$ 80%, <i>n</i> (%)	Class Fidelity, Mean (%)	Participant-Level Contamination Flag, <i>n</i> (%)
GenAI	330	14	78.48 (16.44)	179 (54.2%)	84.88	8 (2.4%)
Active Control	326	14	76.82 (17.74)	183 (56.1%)	85.03	6 (1.8%)

Immediate outcome missingness was 2.73% for GenAI (nine out of 330; all recorded as withdrawal) and 1.84% for the active controls (six out of 326; all recorded as withdrawal), below the prespecified 5% trigger for primary-outcome multiple imputation. Delayed missingness was 10.61% for GenAI (35 of 330) and 9.82% for the active controls (32 of 326). In the delayed pattern–mixture stress test, the adjusted difference was 2.35 points under model-based prediction, 1.93 with a four-point adverse differential shift, 1.52 with an eight-point shift, and 1.12 with a 12-point shift. The completed-data 95% interval first included zero at a 16-point adverse differential shift (estimate 0.73; interval −0.10 to 1.55). Because this was a deterministic single imputation, the intervals do not include imputation uncertainty and are interpreted as sensitivity bounds rather than confirmatory estimates.

Five centrally adjudicated events were documented and closed before database lock. The event register was not intended to enumerate every participant-level weekly contamination flag. EVT-001 (active control, minor) concerned mobile-phone use during the baseline; EVT-002 (active control, major) concerned external GenAI access during an independent attempt; EVT-003 (GenAI, minor) concerned an incorrect handout version; EVT-004 (active control, minor) concerned temporary family-related absence; and EVT-005 (GenAI, major) concerned omissions of independent verification. The two major events were included among the 14 participant-level flags, while the remaining flags originated from routine fidelity records and did not receive separate EVT identifiers. The resulting per-protocol exclusions and missing-outcome counts are summarized by assigned condition in Table 6.

Table 6. Per-protocol exclusions and missing outcomes by their assigned conditions.

Category	GenAI, <i>n</i>	Active Control, <i>n</i>	Total, <i>n</i>
Observed immediate outcome	321	320	641
Attendance < 80%	146	141	287
Class fidelity < 80%	0	0	0
Participant-level contamination flag	8	6	14
Multiple exclusion reasons	2	3	5

Table 6. *Cont.*

Category	GenAI, <i>n</i>	Active Control, <i>n</i>	Total, <i>n</i>
Final per-protocol population with observed immediate outcome	169	176	345
Delayed absence	8	6	14
Delayed medical reason	9	9	18
Delayed transfer or withdrawal	18	17	35
Per-protocol eligible (all criteria met)	173	180	353
Of whom, observed immediate outcome	169	176	345
Of whom, missing immediate outcome	4	4	8
Total delayed missing	35	32	67

3.5. Process Evidence and Mechanistic Scope

Trace logs and portfolios were retained for implementation monitoring, but the locked database did not contain a scored analysis of prompt quality, detected GenAI errors, verification strategies, or acceptance and rejection decisions. The trial therefore supports inference about the complete instructional configuration and the independent-transfer endpoint. The six process capabilities in Table 1 remain design targets rather than independently tested mediators.

4. Discussion

4.1. Principal Findings

Classes assigned by deterministic constrained minimization to the traceable GenAI-assisted configuration had a 2.78-point higher adjusted immediate transfer score on a 32-point scale than active-control classes after accounting for individual and class baseline performances, study year, and teacher block. The model-based standardized difference was 0.72. Because allocation was not probabilistically randomized and the intervention bundled bounded adaptive dialogue with delayed access, trace capture, critical comparison, and independent verification, this is a positive adjusted comparative association for the evaluated configuration rather than a definitive causal effect of GenAI.

The principal applied contribution is the measurement of tool-free engineering transfer rather than performance while AI assistance remained available. Students completed the primary assessment under supervision without GenAI or internet access. The observed contrast therefore concerns independent engineering reasoning after a semester of differently configured learning support, which is more directly relevant to curriculum and assessment design than assignment quality or satisfaction measured during AI access.

4.2. Interpretation in Relation to Previous Evidence

The findings are consistent with experimental work in which structured ChatGPT use was combined with peer assessment, flipped learning, or explicit higher-order thinking activities [6,7]. They also help explain why controlled studies have reported weaker outcomes under less structured access [9]. GenAI can reduce effort by supplying a finished response, but the educational effect depends on whether the learner must still formulate the engineering problem, expose assumptions, test the alternative, and defend a final decision.

The mathematics trial reported by Bastani et al. [8] is particularly informative, because it showed that guardrails can determine whether AI-supported practice transfers to unaided performance. The present intervention similarly treated generated content as provisional evidence rather than as an answer. Students produced an initial human attempt before access, compared the alternative against a common checklist, and verified the result in a separate analytical or computational environment.

Pardos and Bhandari [10] showed that carefully generated AI help can produce learning gains comparable to human-authored help in some mathematics settings. The present active controls extended that comparison by matching curricular tasks, intended outcomes, instructional time, feedback expectations, and verification demands. Nevertheless, GenAI offered bounded adaptive dialogue whereas control clarifications were predefined. The observed contrast should therefore be attributed to the assignment to the complete instructional configurations and not solely to the provisional alternatives that were machine-generated or curated.

4.3. Plausible Learning Mechanisms and Human–AI Accountability

Three features are plausible contributors to independent transfer. First, the initial human attempt preserved retrieval, planning, and productive struggle before external support. Second, critical comparison converted the generated alternative into an object for evaluation rather than passive consumption. Third, independent verification required students to reconstruct the reasoning in an environment that had not generated the alternative. These features correspond to constructive and interactive engagement and to feedback practices that support self-regulation [11,13].

Viewed through an AI-literacy lens, the workflow operationalized collaboration as a set of accountable practices: formulating the problem, asking for a bounded form of support, preserving provenance, challenging the response, checking it independently, and taking responsibility for the accepted recommendation [14,15]. The study did not separately manipulate these components or provide a validated mediator for each literacy dimension. Prompt quality, novelty, variation in alternative solutions, motivation, and the adaptivity of the GenAI dialogue may also have contributed; separate factorial, process, or mediation studies are needed to distinguish these mechanisms.

4.4. Delayed Transfer and Missing-Data Caution

The delayed point estimate was smaller than the immediate estimate but remained substantial. The condition-by-time interaction was not statistically evident, but Form A was not interchangeable with Forms P and B, so this supportive longitudinal contrast should not be read as a precise decay estimate. The delayed analysis also used 589 observed outcomes and estimated essentially zero class variance. The adverse pattern–mixture stress test showed that the direction and conventional interval exclusion of zero persisted through a 12-point differential shift among missing outcomes and weakened at 16 points. This analysis reduces, but cannot eliminate, uncertainty from unobserved delayed outcomes.

4.5. Implementation and Practical Significance

Implementation fidelity was averaged to be approximately 85% in both conditions, reducing concern about a large class-level dose imbalance. The mean attendance was below 80% in both groups, and the 80% rule excluded many participants from the per-protocol set; the continuous attendance-adjusted analysis and the 70% and 75% threshold analyses nevertheless produced estimates close to the primary result. The 2.78-point adjusted difference represents 8.7% of the complete scale and approximately 0.35 point per criterion if spread evenly. This provides a practical scale translation, although the study did not estimate criterion-specific effects or establish a minimal educationally important difference. At the same time, the attendance pattern indicates incomplete uptake of the intended 16-week instructional dose and should be considered when assessing the feasibility of sustained implementation under routine university conditions.

The workflow can be integrated into existing engineering courses without replacing disciplinary content with a separate generic AI-literacy module. In this trial, verification was embedded in field-capacity calculation, power balance, optimization, sensing, uncertainty,

and other ordinary engineering tasks. The minimum transferable implementation consists of an initial unaided formulation, a bounded alternative, explicit comparison, independent checking, a provenance record, and an accountable professional conclusion. A curated alternative remains a practical control or instructional option where GenAI access, privacy, language support, or institutional policy is restrictive.

4.6. Privacy, Integrity, Equity, and Institutional Governance

Traceability creates pedagogical value but also produces potentially sensitive process data. Prompt–response records should therefore be collected only for a stated educational purpose, minimized, deidentified, protected, and retained for a defined period. Students should not be required to disclose personal or confidential information to a commercial model, and access conditions should not make a paid account or high-bandwidth device an implicit prerequisite. The human-centred approach recommended by UNESCO [25] is directly relevant: institutional validation, data protection, inclusion, and preservation of human agency must accompany pedagogical experimentation.

Academic integrity should also be addressed through assessment designs instead of relying on AI detectors. In the present workflow, disclosure and verification were assessed as part of the reasoning process, while the primary outcome was completed without GenAI. This combination provides a more direct basis for evaluating learning than attempting to infer authorship from stylistic characteristics alone.

4.7. Strengths and Limitations

The principal strengths were multicentre participation, 28 intact allocated classes, explicit enumeration of 75,288 admissible assignments, a structured active control, a common 16-week module, supervised condition-neutral outcomes, delayed follow-up, coded fidelity records, and a reproducible small-sample analysis. Baseline assessment preceded disclosure of allocation, and analysis retained original class assignment.

Several limitations require emphasis. The final class assignment was produced by deterministic minimization after the baseline covariates were observed. The main model adjusted for baseline score and stratification variables; a supportive model additionally adjusted for age and sex, but class schedules were unavailable in the locked analysis file. Residual confounding and uncertainty introduced by the data-dependent allocation rule therefore remain possible, and the confidence intervals are model-based rather than randomization-based. Students and instructors could not be blinded to the source and interaction format of the alternative. Bounded GenAI dialogue was adaptive, whereas the control clarifications were predefined, so the study cannot isolate the model, adaptivity, dialogue, or traceability components. Scoring files concealed condition and institution, but script order was not randomized. The optimizer reported precision loss, and delayed class variance was on the boundary, although residual summaries and class-level analyses did not materially change the contrast.

The forms were developed by the author group and reviewed through independent trained scoring and crossover piloting, but reliability alone does not establish validity. No separate formal content-validity index or construct-validation study was conducted. Form A did not meet the prespecified cross-form tolerance with Forms P and B; raw longitudinal differences and $y = x$ no-change interpretations are therefore inappropriate. Delayed missingness was approximately 10%, and the deterministic stress test cannot replace formal multiple imputation or recover unobserved outcomes. The per-protocol cohort was substantially smaller because the 80% attendance criterion was strict, and those supportive estimates remain vulnerable to selection.

Finally, the study was conducted in Uzbek in six universities within agricultural engineering. Although use of one instructional and assessment language removes cross-language form variation within this study, prior GenAI experience, device access, institutional policy, and regional curriculum features may influence implementation and transferability. No other instructions, GenAI practices, or scheduled learning activities related to the study module occurred between its completion on 19 May and Form A on 26 May 2025. The assessment was administered in dedicated time outside module teaching. The seven-day interval reduced immediate recall of the final session and allowed consolidation, although unrecorded independent study cannot be excluded.

4.8. Implications and Future Research

Engineering programmes should distinguish unrestricted answer generation from traceable collaboration. A defensible policy can permit GenAI after an independent attempt, require complete provenance, make verification evidence part of assessed work, and retain a tool-free condition-neutral outcome for selected high-stakes competencies. Instructor preparation should include model failure recognition, code and calculation checking, privacy protection, design of verification tasks, and fair assessment of process traces.

Future trials should compare individual workflow components, standardize or experimentally vary dialogue adaptivity, test other engineering disciplines and countries, and extend follow-up. Research should validate separate measures of prompt literacy, provenance literacy, critical GenAI literacy, verification literacy, and professional accountability. Process analyses should determine when students seek assistance, what model errors they detect, whether checking is substantive, and how often generated advice changes the final engineering model. Longer-term studies should evaluate authentic design tasks involving safety, cost, and operational consequences.

5. Conclusions

Across six universities, classes assigned by deterministic constrained minimization to a traceable GenAI-assisted learning configuration had a 2.78-point higher adjusted immediate engineering-transfer score than structured active control (model-based Kenward–Roger 95% CI [2.02, 3.55]), equivalent to 8.7% of the 32-point scale. The delayed adjusted difference was 2.34 points, and the direction remained positive under substantial adverse missing-data shifts. A matched active control, supervised tool-free outcomes, and consistent sensitivity estimates strengthen the comparative evidence. Nevertheless, post-baseline deterministic allocation prevents a fully randomization-based causal interpretation; the contrast belongs to the complete configuration rather than GenAI alone, and instrument validity, attendance selection, delayed missingness, cross-form non-equivalence, and context also bound interpretation. Future multicentre randomized trials should preregister allocation and analysis materials, validate the transfer construct, isolate workflow components, and extend the follow-up.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16188949/s1>. Supplementary File S1: S1_Supp_v4.docx, containing Tables S1–S10 on completeness, class-level summaries, the module, psychometrics, events, robustness, Kenward–Roger results, delayed missing-data sensitivity, diagnostics, and the restricted-allocation audit. Supplementary File S2: S2_Results_v4.xlsx; Supplementary File S3: S3_KR_v4.py; Supplementary File S4: S4_Analysis_v4.py; Supplementary File S5: S5_Confirm_v4.R; Supplementary File S6: S6_Stats_v4.docx; Supplementary File S7: S7_Data_v4.csv; and Supplementary File S8: S8_SHA256_v4.txt; Analysis. The signed protocol and SAP are retained by the coordinating centre and will only be provided upon a substantiated request. Local permission letters and the protected

allocation log are likewise retained for confidential review when specifically requested. None of these documents are included in the public package.

Author Contributions: Conceptualization, Y.S., F.M., K.C. and D.C.; methodology, Y.S., F.M., K.C. and D.C.; software, Y.S., F.M., K.C., Z.B., D.C., M.S., D.S., K.I., G.J. and M.E.; validation, Y.S., F.M., K.C. and D.C.; formal analysis, Y.S., F.M., K.C., Z.B., D.C., M.S., D.S., K.I., G.J. and M.E.; investigation, Y.S., F.M., K.C. and D.C.; resources, Y.S., F.M., K.C. and D.C.; data curation, Y.S., F.M., K.C., Z.B., D.C., M.S., D.S., K.I., G.J. and M.E.; writing—original draft preparation, Y.S., F.M., K.C. and D.C.; writing—review and editing, Y.S., F.M., K.C. and D.C.; visualization, Y.S., F.M., K.C. and D.C.; supervision, Y.S., F.M., K.C. and D.C.; project administration, Y.S., F.M., K.C. and D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Commission of Karshi Engineering-Economics Institute, subsequently reorganized as Karshi State Technical University (Protocol No. 14/2023, 28 August 2023). No formal multicentre ethics-reliance mechanism or single-IRB arrangement was executed. Ethical governance consisted of the coordinating centre’s approval under Protocol No. 14/2023 together with site-specific administrative authorization rather than six independent ethics approvals. Each university issued a rector- or vice-rector-signed administrative permission letter in August–September 2023 before study activities. The approval and permission letters are retained for confidential review on request.

Informed Consent Statement: Written informed consent was obtained from all participants before inclusion in research-related data collection. Participation was voluntary, research participation was separated from academic grading, coded identifiers protected confidentiality, and participants could decline participation or withdraw without adverse academic consequences.

Data Availability Statement: The deidentified locked participant-level dataset supporting the primary, delayed, and sensitivity analyses is provided as Supplementary File S7 (S7_Data_v4.csv). Materials containing potentially identifiable prompt–response records, named teacher records, detailed implementation logs, or restricted assessment keys are not included in the supplementary package. Access to such restricted materials may be considered upon reasonable request to the corresponding author, subject to institutional authorization, ethics approval, participant-consent restrictions, and applicable data-protection requirements. Provision of the deidentified dataset is subject to Protocol No. 14/2023, the approved consent materials, institutional authorization, and applicable data-protection requirements.

Acknowledgments: The authors acknowledge the participating students, instructors, institutional coordinators, assessors, and administrative staff who supported the delivery of the educational module, supervised assessment, implementation-fidelity monitoring, data collection, scoring, quality control, and deidentification. After completion of data collection, database locking, and the principal statistical analyses, the authors used Gemini 3.6 Flash (Google DeepMind, London, UK; accessed July–August 2026) for language refinement and document preparation. The system was not used to generate or alter participant data, allocate classes, score outcomes, estimate the reported contrasts, or determine the study conclusions. The authors accept full responsibility for the content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Farrokhnia, M.; Banihashem, S.K.; Noroozi, O.; Wals, A. A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innov. Educ. Teach. Int.* **2024**, *61*, 460–474. [[CrossRef](#)]
2. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [[CrossRef](#)]

3. Lo, C.K. What is the impact of ChatGPT on education? A rapid review of the literature. *Educ. Sci.* **2023**, *13*, 410. [CrossRef]
4. Tlili, A.; Shehata, B.; Adarkwah, M.A.; Bozkurt, A.; Hickey, D.T.; Huang, R.; Agyemang, B. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ.* **2023**, *10*, 15. [CrossRef]
5. Bravo, F.A.; Cruz-Bohorquez, J.M. Engineering education in the age of AI: Analysis of the impact of chatbots on learning in engineering. *Educ. Sci.* **2024**, *14*, 484. [CrossRef]
6. Huesca, G.; Martínez-Treviño, Y.; Molina-Espinosa, J.M.; López-Islas, J.R.; Ramírez-Montoya, M.S. Effectiveness of using ChatGPT as a tool to strengthen benefits of the flipped learning strategy. *Educ. Sci.* **2024**, *14*, 660. [CrossRef]
7. Wu, T.-T.; Lee, H.-Y.; Chen, P.-H.; Lin, C.-J.; Huang, Y.-M. Integrating peer assessment cycle into ChatGPT for STEM education: A randomised controlled trial on knowledge, skills, and attitudes enhancement. *J. Comput. Assist. Learn.* **2025**, *41*, e13085. [CrossRef]
8. Bastani, H.; Bastani, O.; Sungu, A.; Ge, H.; Kabakçı, Ö.; Mariman, R. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2422633122. [CrossRef] [PubMed]
9. Essel, H.B.; Vlachopoulos, D.; Essuman, A.B.; Amankwa, J.O. ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100198. [CrossRef]
10. Pardos, Z.A.; Bhandari, S. ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS ONE* **2024**, *19*, e0304013. [CrossRef] [PubMed]
11. Chi, M.T.H.; Wylie, R. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educ. Psychol.* **2014**, *49*, 219–243. [CrossRef]
12. Freeman, S.; Eddy, S.L.; McDonough, M.; Smith, M.K.; Okoroafor, N.; Jordt, H.; Wenderoth, M.P. Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 8410–8415. [CrossRef] [PubMed]
13. Nicol, D.J.; Macfarlane-Dick, D. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Stud. High. Educ.* **2006**, *31*, 199–218. [CrossRef]
14. Long, D.; Magerko, B. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–16. [CrossRef]
15. Ng, D.T.K.; Leung, J.K.L.; Chu, S.K.W.; Qiao, M.S. Conceptualizing AI literacy: An exploratory review. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100041. [CrossRef]
16. Kosar, T.; Ostojić, D.; Liu, Y.D.; Mernik, M. Computer science education in the ChatGPT era: Experiences from an experiment in a programming course for novice programmers. *Mathematics* **2024**, *12*, 629. [CrossRef]
17. Li, R.; Li, M.; Qiao, W. Engineering students' use of large language model tools: An empirical study based on a survey of students from 12 universities. *Educ. Sci.* **2025**, *15*, 280. [CrossRef]
18. Sánchez-Ruiz, L.M.; Moll-López, S.; Nuñez-Pérez, A.; Morano-Fernández, J.A.; Vega-Fleitas, E. ChatGPT challenges blended learning methodologies in engineering education: A case study in mathematics. *Appl. Sci.* **2023**, *13*, 6039. [CrossRef]
19. Hoffmann, T.C.; Glasziou, P.P.; Boutron, I.; Milne, R.; Perera, R.; Moher, D.; Altman, D.G.; Barbour, V.; Macdonald, H.; Johnston, M.; et al. Better reporting of interventions: Template for Intervention Description and Replication (TIDieR) checklist and guide. *BMJ* **2014**, *348*, g1687. [CrossRef] [PubMed]
20. Campbell, M.K.; Piaggio, G.; Elbourne, D.R.; Altman, D.G. CONSORT 2010 statement: Extension to cluster randomised trials. *BMJ* **2012**, *345*, e5661. [CrossRef] [PubMed]
21. Hopewell, S.; Chan, A.-W.; Collins, G.S.; Hróbjartsson, A.; Moher, D.; Schulz, K.F.; Tunn, R.; Aggarwal, R.; Berkwits, M.; Berlin, J.A.; et al. CONSORT 2025 statement: Updated guideline for reporting randomised trials. *BMJ* **2025**, *389*, e081123. [CrossRef] [PubMed]
22. Chan, A.-W.; Boutron, I.; Hopewell, S.; Moher, D.; Schulz, K.F.; Collins, G.S.; Tunn, R.; Aggarwal, R.; Berkwits, M.; Berlin, J.A.; et al. SPIRIT 2025 statement: Updated guideline for protocols of randomised trials. *BMJ* **2025**, *389*, e081477. [CrossRef] [PubMed]
23. Gamble, C.; Krishan, A.; Stocken, D.; Lewis, S.; Juszczak, E.; Doré, C.; Williamson, P.R.; Altman, D.G.; Montgomery, A.; Lim, P.; et al. Guidelines for the content of statistical analysis plans in clinical trials. *JAMA* **2017**, *318*, 2337–2343. [CrossRef] [PubMed]
24. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. *ICH E9(R1): Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials*; ICH: Geneva, Switzerland, 2019. Available online: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf (accessed on 9 August 2026).
25. UNESCO. *Guidance for Generative AI in Education and Research*; UNESCO: Paris, France, 2023. Available online: <https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research> (accessed on 9 August 2026).

26. Enders, C.K.; Tofighi, D. Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychol. Methods* **2007**, *12*, 121–138. [[CrossRef](#)] [[PubMed](#)]
27. Kenward, M.G.; Roger, J.H. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **1997**, *53*, 983–997. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.