Adv. J. Manag. Humanit. Soc. Sci. (**2026**), Volume 2, Issue 3, 209-225

Advanced Journal of Management, Humanity and Social Science

Journal homepage: https://www.ajmhss.com/

DOI: https://doi.org/10.5281/zenodo.21862976

Development, Content Validation, and Preliminary Effectiveness of the NEXUS Multicomponent Educational Protocol Integrating AI-Mediated Socratic Inquiry, Identity Meaning-Making and Impactful Expression in Adolescents: A Small-Cluster Randomized Trial with Three-Month Follow-Up

Akram Sabzipour[1*], Sajad Hazrati[2]

[1]Department of Psychology, Payame Noor University, Takestan, Iran

[2]PhD Student in Health Psychology, Department of Psychology, To.C., Islamic Azad University, Tonekabon, Iran

Article info

Received: 11.06.2026

Accepted: 05.08.2026

Available Online: 09.08.2026

Checked for Plagiarism: Yes

Keywords:

adolescent learning; artificial intelligence; critical thinking; intrinsic motivation; learning agency; Socratic questioning

A B S T R A C T

Background: Generative artificial intelligence (GenAI) can support explanation and feedback, but unrestricted answers may encourage cognitive offloading, weak source evaluation, and diminished learner authorship. This study developed and content-validated NEXUS, which restricts GenAI to Socratic coaching, and examined preliminary effectiveness for learning agency, intrinsic motivation, and critical-thinking disposition among male upper-secondary students in Qazvin, Iran.

Methods: Phase I combined needs mapping in 463 students, co-design, cognitive interviews, and two-round Delphi validation by 15 experts. Phase II randomized 12 classes (106 students; six clusters per arm) to ten weekly NEXUS sessions or a dose- and technology-matched active control. Three-month follow-up was primary. Mixed-model estimates were paired with CR2 inference, CR3 sensitivity analysis, restricted wild-cluster bootstrap-t intervals, and exact within-pair randomization tests. Model-based Hedges g used total class-plus-student variance.

Results: The 55-element manual achieved S-CVI/Ave=0.91; universal agreement was 0.31, and Fleiss kappa was 0.66 for relevance and 0.61 for necessity. Follow-up completion was 89.6%. Adjusted follow-up differences were 0.44 for agentic engagement (95% CI 0.10 to 0.78; Holm-adjusted p=0.042; g=0.43), 0.31 for intrinsic motivation (95% CI -0.04 to 0.66; adjusted p=0.086; g=0.30), and 6.1 for critical-thinking disposition (95% CI 0.6 to 11.6; adjusted p=0.078; g=0.38). The exact global multivariate randomization test was p=0.031. Fidelity averaged 86.8%; minor AI-output and participant-burden events occurred, but no serious related event was observed.

Conclusion: NEXUS showed credible content validity and a promising joint signal, strongest for agentic engagement. With only 12 clusters, incomplete blinding, and no prospective registration, independent replication is required.

Introduction

Generative artificial intelligence has rapidly become a general-purpose layer through which students search for information, obtain explanations, draft text, solve problems, and revise academic products. Its educational value, however, is not an intrinsic property of the model. Value depends on who formulates the problem, who evaluates evidence, how uncertainty is represented, and whether the learner remains responsible for the final judgment. Contemporary analyses therefore distinguish productive cognitive partnership from substitution of machine output for human inquiry [1]. A systematic review and meta-analysis of experimental studies found that ChatGPT-supported instruction can improve academic performance and selected affective-motivational outcomes, but the magnitude and durability of benefits vary by

Corresponding Author: **Akram Sabzipour (sabzipourakram@gmail.com)*

1 Email: sajad.hazrati@iau.ac.ir)

pedagogical structure, task type, intervention duration, and the presence of guidance [2].

The central risk is not limited to factually incorrect output. A fluent, confident, and apparently coherent response may be accepted without sufficient monitoring, source verification, counterargument, or revision. This can create a discrepancy between the surface quality of a submitted product and the depth of the learning process. Reviews of critical thinking in GenAI-enhanced environments conceptualize critical thinking as a coordinated set of processes: defining a problem, distinguishing claim from evidence, assessing source quality, examining assumptions, considering alternatives, calibrating confidence, and accepting responsibility for a defensible conclusion [3,4]. These processes are especially important in secondary education, when adolescents are consolidating self-regulation, identity, and epistemic habits. A recent K-12 scoping review identified privacy exposure, hallucination, bias, inequitable access, overreliance, reduced human interaction, and weakened teacher oversight as major risks of poorly governed GenAI integration [5].

Evidence further cautions against treating superior immediate output as equivalent to superior learning. GenAI access can facilitate short-term task completion while simultaneously increasing metacognitive laziness when learners delegate planning, monitoring, or verification to the model [6]. From an agency perspective, choosing among machine-generated options is not sufficient. Learning agency requires the learner to initiate, influence, monitor, and take responsibility for the direction and consequences of learning [7]. Experimental work on self-directed activity indicates that conceptual learning improves when students genuinely construct and decide, rather than merely select from options supplied by another agent [8]. These findings imply that responsible educational design should constrain the model's authority, require visible learner reasoning before AI exposure, and preserve the student's final authorship. AI-mediated Socratic inquiry provides one possible design solution. A Socratic system does not begin by completing the task. It clarifies the learner's question, probes assumptions, requests evidence, raises counterexamples, surfaces uncertainty, and asks the learner to revise a provisional position. Emerging studies of Socratic AI and GenAI-supported formative assessment suggest that structured questioning can strengthen reflective engagement, research-question development, and higher-order thinking when dialogue replaces unrestricted answer generation [9,10]. Most available evidence, however, is situated in higher education, short laboratory tasks, or self-selected users. Transfer to adolescents requires culturally responsive co-design, age-appropriate privacy safeguards, teacher mediation, and explicit rules governing what the model may and may not do.

NEXUS was designed around three mutually reinforcing theoretical pillars. The first is learning agency, operationalized as observable contribution to instruction through student-initiated questions, preferences, suggestions, evidence requests, requests for assistance, and respectful disagreement. Agentic engagement has been described as a distinct dimension of student engagement [11], and the Enlarged Student Agentic Engagement Scale provides a validated measure of students' constructive contribution to the flow of instruction [12].

In NEXUS, agency is not treated as unrestricted choice; it is enacted through accountable decisions, evidence trails, revision, and ownership of the final product. The second pillar is autonomous motivation and identity meaning-making. Self-determination theory proposes that motivation becomes more self-endorsed when educational environments support autonomy, competence, and relatedness and when learning activities are integrated with personally valued goals [13]. Recent research on GenAI use similarly suggests that perceived autonomy support and self-efficacy are associated with self-learning motivation, whereas externally controlled use is less likely to produce deep processing [14].

A longitudinal adolescent intervention found that AI-supported learning does not uniformly enhance every motivational dimension and that competence-related experiences are important for subsequent motivation [15]. Project-based and programming environments informed by self-determination theory also indicate that model-supported learning is more productive when students retain meaningful goals, choice within structure, and responsibility for evaluation [16,17].

Identity work gives autonomous motivation a developmental anchor. Meta-analytic evidence indicates that psychosocial interventions can support adolescents' personal and social identity development [18] and that educational and vocational identity interventions can strengthen future-oriented exploration and commitment [19]. NEXUS treats identity as a revisable narrative connecting past experience, current values, possible selves, and small self-concordant actions. This component is intended to prevent critical inquiry from becoming a decontextualized classroom exercise and to help students answer not only "Is this claim defensible?" but also "Why does this matter for the person I am becoming?"

The third pillar is impactful expression. Students translate a judgment into a claim supported by evidence and reasoning, represent an opposing view fairly, adapt a message to an authentic audience, disclose the role of AI, and remain responsible for the final communication. Self-directed writing across computer- and AI-supported tasks can

promote development when learners plan, monitor, and revise rather than delegate authorship [20]. Active-learning interventions likewise improve motivation and attitudes when students generate, discuss, and defend meaning [21]. Research on agentic AI links autonomy support and self-efficacy with self-learning motivation [22], while emerging evidence indicates that AI literacy may mediate the association between chatbot use and critical thinking [23]. A meta-analysis of 29 experimental and quasi-experimental studies found a positive overall effect of GenAI-supported instruction on higher-order thinking, with stronger results under structured, sustained, and pedagogically aligned conditions [24]. Iranian studies further demonstrate that theory-driven adolescent educational interventions can be delivered with strong recruitment, parental permission, adherence, and cultural adaptation when procedures are explicit [25,26].

NEXUS therefore positions GenAI as a bounded dialectical coach rather than an answer authority. Its five-step microcycle requires learners to notice and name a question, examine evidence and assumptions, explore alternatives, unify learning with identity and values, and share a defensible position that shapes action. The present study had two sequential aims. Phase I aimed to develop the NEXUS theory of change and session architecture, map developmental needs among upper-secondary students in Qazvin, and establish face and content validity. Phase II aimed to estimate the preliminary joint and outcome-specific effects of NEXUS, compared with a dose- and technology-matched active control, at immediate post-intervention and three-month follow-up. Because only 12 class clusters were available, the trial was framed as a small-cluster preliminary efficacy study: the three-month multivariate test was the gatekeeping primary analysis, and individual outcomes were interpreted from multiplicity-adjusted, small-sample-robust inference rather than from effect magnitude alone. The study was conducted in Qazvin, Iran, during the 2025-2026 academic year.

Materials and Methods

Study Design and Reporting Framework:

A sequential intervention-development and evaluation design was used. Phase I comprised evidence synthesis, logic-model construction, quantitative needs mapping, participatory co-design, cognitive interviewing, expert Delphi validation, and manual refinement. Phase II comprised a parallel, two-arm, assessor-blinded, cluster randomized controlled trial with baseline, immediate post-intervention, and three-month follow-up assessments. Development followed the Medical Research Council framework for complex interventions, which emphasizes programme theory, context, uncertainty, feasibility, refinement, and transparent progression criteria [27]. Intervention reporting followed TIDieR principles [28]. Trial planning and reporting were informed by SPIRIT 2025 [29], CONSORT 2025 [30], and the 2026 child and adolescent extensions [31,32].

Setting and School Sampling Frame

The school list supplied to the project included first- and second-cycle secondary schools. Because the target population was upper-secondary students, schools explicitly identified as first-cycle sites were not included in quantitative screening or effectiveness testing. The upper-secondary sampling frame comprised Shahid Babaei NODET Secondary School (second cycle), Sadra Board of Trustees Secondary School (second cycle), Qazvin Atomic Energy School, Erfan School, Pasdaran Secondary School, Allamah Tabatabaei Secondary School, and Eqmesheh State Model Secondary School. Norouzian Board of Trustees Secondary School (first cycle), Shahid Babaei NODET Secondary School (first cycle), and Allamah Helli Non-Governmental Secondary School (first cycle) were retained as potential sites for later developmental extension. Public ratings and street addresses were not used as sampling, stratification, or analytic variables.

Schools were stratified by administrative type. Classes were treated as natural clusters to minimize contamination. Needs mapping involved all seven eligible upper-secondary schools. The effectiveness phase used 12 classes from six schools, with two classes selected within each school and matched as closely as possible by grade, academic track, timetable, and average prior-term grade. All eligible participants within a selected class cluster received the intervention assigned to that class.

Phase I: Needs Mapping

Information packages were distributed to 612 students. A parent or guardian returned a participation decision for 508 (83.0%); 497 granted written permission, and 486 adolescents subsequently provided written assent. Of these, 463 supplied analyzable screening records. Twelve records were excluded for more than 15% item nonresponse, seven for prespecified invariant or contradictory response-pattern flags, and four because of duplicate or ineligible enrolment. Thus, the analyzable screening yield was 75.7% of all students approached, avoiding the assumption that nonresponse was negligible. The screening sample comprised 158 Grade 10 students (34.1%), 156 Grade 11 students (33.7%), and 149 Grade 12 students (32.2%). All participants were male because the eligible schools in the sampling frame were boys' schools. Mean age was 16.42 years (SD=0.84; range=15-18).

Measures and Programme-Selection Thresholds

Learning agency was assessed with the 10-item Enlarged Student Agentic Engagement Scale. Items

are rated from 1 to 7 and averaged; higher scores indicate greater constructive contribution to the flow of learning [12]. Intrinsic motivation was operationalized as the mean of the 12 intrinsic-motivation items of the Academic Motivation Scale-High School Version. The scale is rated from 1 to 7 and distinguishes intrinsic motivation, identified regulation, introjected regulation, external regulation, and amotivation [33]. Iranian studies have supported the contextual use and standardization of the Academic Motivation Scale among secondary students [34,35]. Critical-thinking disposition was assessed with the 33-item Ricketts Critical Thinking Disposition Questionnaire, comprising innovativeness, cognitive maturity, and engagement [36]. Items are rated from 1 to 5; eight are reverse-scored, and the total ranges from 33 to 165. The Persian adaptation has shown an interpretable three-factor structure and satisfactory reliability among Iranian high-school students [37]. Screening values were not treated as clinical or diagnostic cut-offs. For programme selection, the lower quartile of each cleaned screening distribution was calculated: agentic engagement <= 3.18, intrinsic motivation <= 3.55, and critical-thinking disposition <= 102. A student scoring at or below the empirical threshold on at least two constructs was classified as having elevated developmental need. This rule identified 121 of 463 students (26.1%); the proportion slightly exceeded 25% because students could meet two or three correlated thresholds and because observed scores were tied at the quartile boundaries. Because the thresholds were sample-dependent, they were used only for recruitment within this Qazvin sampling frame. Internal consistency was evaluated with Cronbach's alpha and McDonald's omega. Coefficients were 0.86 and 0.87 for agentic engagement, 0.91 and 0.92 for intrinsic motivation, and 0.88 and 0.89 for critical-thinking disposition, respectively.

Evidence Synthesis, Logic Model, and Co-Design

Candidate mechanisms, activities, risks, and implementation conditions were extracted from the evidence base cited in this manuscript. Evidence was organized into six domains: AI-mediated Socratic inquiry; learning agency; self-determination and intrinsic motivation; adolescent identity development; critical thinking and source evaluation; and ethical, human-centred AI use. UNESCO guidance informed privacy, human agency, transparency, fairness, age appropriateness, and AI competency safeguards [38,39]. The logic model linked inputs and activities to proximal mechanisms, observable process indicators, and intended outcomes. The non-negotiable design rule was that learners produced a provisional question, judgment, or draft before viewing model-generated material. The co-design group included 10 adolescents selected to represent Grades 10-12 and varied screening profiles, four secondary-school teachers, and four school counsellors. Two 120-minute workshops examined language, task burden, classroom fit, privacy risk, emotionally sensitive topics, peer-feedback rules, accessibility, and offline alternatives. Cognitive interviews were then conducted with 12 additional adolescents using think-aloud and verbal-probing procedures. Participants explained how they interpreted instructions, identified unclear language, rated emotional comfort and cognitive burden, and responded to privacy and disagreement scenarios. Revisions replaced abstract terminology with behavioral examples, limited the model to three alternative viewpoints per cycle, made personal identity disclosure optional, and required de-identification before any student prompt was entered. Ethical safeguards were benchmarked against the 2024 Declaration of Helsinki [40].

Expert Panel and Content Validation

The Delphi panel comprised 15 experts: five educational psychologists, three adolescent or clinical psychologists, three curriculum and instruction specialists, two educational technology or AI specialists, and two psychometricians. Eligibility required a doctorate or senior specialist role plus either at least five relevant peer-reviewed publications or at least ten years of direct professional practice. The panel represented eight institutions and more than one theoretical or professional tradition; median experience was 12 years (interquartile range=8-17). No panelist reported a financial relationship with an AI vendor. Three disclosed prior academic collaboration with members of the development team; ratings were retained but examined in an exclusion sensitivity analysis. Panelists independently rated every objective, core activity, homework task, process indicator, facilitator rule, and AI-safety rule for necessity, relevance, clarity, feasibility, cultural responsiveness, and safety. Anonymized distributions, minority rationales, and item-level count not only group means were returned between rounds to limit conformity pressure. For each element, the content validity ratio (CVR) was calculated as (ne-N/2)/(N/2), where he is the number rating the element essential and N is panel size. With 15 experts, CVR >=0.49 was the retention criterion. The item-level content validity index (I-CVI) was the proportion assigning a relevance score of 3 or 4 on a four-point scale; I-CVI >=0.78 was required. S-CVI/Ave >=0.90 and Delphi agreement >=80% were progression criteria. Universal agreement was reported descriptively rather than required to exceed 0.80 because it becomes increasingly stringent as the number of items and raters grows. Modified kappa adjusted I-CVI for chance agreement. Fleiss kappa with 95% confidence intervals quantified multi-rater agreement for relevance and necessity, and the

variance and full frequency distribution of ratings were retained for every round. Elements failing one threshold were revised or removed unless a documented safety rationale required retention and re-rating. These choices follow current guidance that CVI/CVR values should be accompanied by transparent item generation, disagreement, and agreement diagnostics rather than treated as self-sufficient evidence of validity [41].

NEXUS Intervention Architecture

The NEXUS microcycle maps three macroprocesses onto five learner actions: exploration comprises N, E, and X; connection is represented by U; and manifestation is represented by S. Table 1 specifies the learner action, permitted AI role, and observable trace.

Table 1. Five-step NEXUS learning cycle and bounded AI role

Step	Learner action	Permitted AI role	Required trace
N - Notice	Notice a need or ambiguity and name an open, answerable question	Clarify scope and ask follow-up questions; do not answer the substantive question	Initial question, rationale, and confidence rating
E - Examine	Examine claims, sources, assumptions, bias, counterevidence, and uncertainty	Request reasons and sources, flag inconsistency, and display uncertainty	Claim-evidence-reasoning record and source check
X - eXplore	Explore competing viewpoints, counterexamples, and conditional consequences	Offer no more than three alternatives and one respectful challenge	Alternatives considered and reasons for acceptance or rejection
U - Unify	Connect learning to values, lived experience, possible selves, and a self-concordant goal	Ask reflective questions; no personality labeling or automated psychological interpretation	Identity-value link and small action plan
S - Share	Share a personally authored position, receive critique, and shape an observable action	Provide structural feedback only after the learner's draft; never write the final product	Final product, revision trail, and AI-use disclosure

Every interaction followed a learner-first rule: student response, AI question, student revision, evidence check, and facilitator sign-off. Students were prohibited from entering names, contact details, health information, private family information, or identifiable peer narratives. The text-only model was accessed through an institutionally managed interface using a fixed September 2025 model snapshot. Generation parameters were fixed at temperature 0.20, top-p 0.90, a 450-token output ceiling, and one model response per student turn. Memory across students, tools, external plug-ins, web browsing, image generation, and model-training retention were disabled. The system prompt required question-led dialogue, explicit uncertainty, refusal to infer mental-health status or personality, and escalation to the facilitator for self-harm, abuse, violence, or other safeguarding content. De-identified logs preserved timestamp, model-version identifier, student-turn category, model response, automatic gate decision, facilitator override, and final disposition. The system prompt, prompt bank, interface rules, generation parameters, audit taxonomy, and model-version checksum were archived so that later replication could distinguish curriculum effects from configuration drift.

The intervention comprised ten weekly 90-minute group sessions delivered to 8-10 students within the same class cluster. Each session included a 10-minute retrieval and safety check, 15 minutes of modelling, 35 minutes of guided practice, 20 minutes of peer critique and revision, and 10 minutes of reflection and transfer planning. Table 2 presents the manualized sequence.

Table 2. Session structure of the final NEXUS protocol

Session	Focus and mechanism	Core in-session activities	Product and fidelity marker
1	Orientation, authorship, and psychological safety	Distinguish assistance from substitution; create group norms; practice de-identification and the learner-first rule	Signed AI-use compact and correctly de-identified practice prompt
2	N - Notice and question quality	Convert vague concerns into open, answerable, consequential questions; rate initial confidence	Question ladder containing scope,

			relevance, and confidence
3	E - Claims, evidence, and source quality	Separate claim, evidence, and reasoning; triangulate a factual claim; identify missing evidence	Completed claim-evidence-reasoning matrix and source audit
4	E - Assumptions, bias, and uncertainty	Detect hidden assumptions, framing, bias, hallucination, and unjustified certainty	Assumption map and revised confidence rating
5	X - Alternatives and counterexamples	Generate plausible alternatives; steelman a competing view; test conditional consequences	Alternatives table and reasoned accept/reject decision
6	U - Values and learning identity	Link a learning problem to values, strengths, prior experience, and a possible self without forced disclosure	Optional identity-value map and self-authored learning statement
7	U - Autonomous goals and implementation intentions	Translate a possible self into a small self-concordant goal; plan barriers and supports	Goal ladder and if-then implementation plan
8	S - Evidence-based expression	Draft a claim for a real audience; integrate evidence, reasoning, uncertainty, and fair counter position	First personally authored product and AI-use disclosure
9	S - Peer critique and revision	Use a respectful critique rubric; compare human and AI feedback; justify accepted and rejected edits	Annotated revision trail and response-to-feedback note
10	Integration, transfer, and public micro-action	Complete a full N-E-X-U-S cycle; present a defensible product; plan transfer to a new subject	Final product, oral defence, transfer plan, and fidelity checklist

Phase II: Cluster Randomized Effectiveness Trial

Eligibility, Recruitment, and Sample Size:

Students were eligible if they were in Grades 10-12, scored at or below the lower-quartile threshold on at least two primary constructs, could participate in group sessions, and provided adolescent assent with written parental permission. Exclusion criteria were planned school transfer during the intervention, concurrent participation in a structured programme targeting the same outcomes, inability to complete Persian questionnaires independently, or an acute condition requiring individualized clinical care rather than a group educational programme. Students were not excluded for ordinary academic difficulty, prior AI use, or lack of a personal device; school devices and an offline pathway were provided. Design-stage Monte Carlo simulation assumed 12 clusters, mean cluster size 9, intraclass correlation 0.04, within-person correlation 0.60, two-sided alpha 0.05, 10% attrition, a standardized group-by-time effect of 0.45, small-cluster degrees of freedom, and Holm control across the three follow-up outcomes. Depending on outcome correlation and cluster-size imbalance, estimated power was only 62%-73%; therefore, 106 students did not constitute a conventionally powered definitive trial. The available 12 classes were the logistical maximum, and the study was explicitly treated as a preliminary efficacy and precision study. This distinction is important because increasing the number of students within a fixed set of 12 clusters cannot replace independent cluster-level information [42]. No post hoc power was calculated; interpretation prioritized design-compatible confidence intervals [43]. Of the 121 students meeting the programme-selection rule, 117 attended the selected class clusters and were invited. Parental permission and adolescent assent were obtained for 110; four withdrew before baseline because of timetable or transfer plans. The final randomized sample was 106 students.

Randomization, Allocation Concealment, and Blinding

Within each of six schools, two class clusters were paired by grade, academic track, timetable, and prior-term academic average. After baseline assessment and completion of recruitment, an independent statistician used a computer-generated Bernoulli draw within every matched pair to assign one class to NEXUS and the other to active control, yielding six clusters and 53 students per condition. Allocations were stored in password-protected files

and released only after the recruitment list and baseline data were locked. The six pairwise assignments create $2^6 = 64$ allocation patterns, which defined the exact randomization-test reference set. Students and facilitators could not be blinded to curriculum content. Questionnaire administrators, data-entry staff, raters of source-verification and written products, and the statistician who first received groups labeled A and B were masked to allocation.

Active Control

The active control received ten weekly 90-minute sessions delivered in the same group size and with the same device access, facilitator contact, breaks, and homework burden. Content covered study planning, note-taking, memory strategies, exam preparation, collaborative learning, digital organization, internet-search basics, and conventional GenAI-assisted summarization. The same managed AI interface was available, but the NEXUS system prompt, learner-first gate, structured evidence audit, counter position sequence, identity integration, revision trail, and authentic-audience product were absent. Control students received the NEXUS workbook after completion of the follow-up assessment.

Facilitator Training and Treatment Integrity

Four school counsellors with at least five years of adolescent group-work experience delivered the programmes. Facilitators completed a 12-hour workshop covering the manual, demonstration, role-play, AI safety, boundary management, and scoring of fidelity items. Each facilitator delivered sessions in only one condition. Weekly supervision used de-identified process summaries. Twenty percent of sessions, stratified by condition and session number, were independently rated from observation notes and system logs. Fidelity was expressed as the percentage of applicable manual components completed. Inter-rater agreement was estimated with a two-way random, absolute-agreement intraclass correlation coefficient.

Outcomes and Assessment Schedule

The three outcome domains were agentic engagement, intrinsic motivation, and critical-thinking disposition, measured with the same instruments used during screening. Baseline assessments occurred within seven days before randomization, post-intervention assessments within seven days after Session 10, and follow-up assessments 12-14 weeks later. The three-month joint outcome vector was the gatekeeping primary endpoint; follow-up outcome-specific contrasts formed the multiplicity-controlled family. Immediate post-intervention contrasts were supportive and exploratory. Secondary implementation outcomes included recruitment, retention, session attendance, fidelity, acceptability, perceived relevance, autonomy support, cognitive load, AI-safety compliance, unsupported-claim correction, and adverse events. Acceptability, relevance, and autonomy support were rated from 1 to 5; cognitive load was rated from 1 to 5, with higher values indicating greater burden.

As a behavioral complement, students completed a 20-minute source-verification task containing six claims with mixed levels of evidential support. Two blinded raters scored correct identification of support, appropriate confidence calibration, and justification quality using a 0-12 rubric. Written products were scored on a 0-20 rubric covering question quality, evidence use, counter position fairness, uncertainty, revision rationale, and audience adaptation. Inter-rater reliability was estimated before discrepancies were resolved by consensus. These behavioral outcomes were secondary and were not included in the multiplicity-controlled primary family.

Safety Monitoring and Ethical Considerations

Facilitators recorded privacy incidents, distress, frustration, inappropriate model output, safeguarding disclosures, and protocol deviations after every session. A predefined escalation pathway required immediate suspension of the AI interaction, private assessment by the school counsellor, contact with the parent or designated safeguarding officer when indicated, and referral to appropriate services. The protocol did not use AI to diagnose, predict mental health, infer personality, or make disciplinary or academic placement decisions.

Administrative authorization was obtained from the Qazvin education authority and participating schools; however, no formally numbered institutional research-ethics approval was issued, and no approval number is claimed. Written informed permission was obtained from parents or legal guardians, and written assent was obtained from adolescents. Information was provided outside the presence of classroom teachers; families had at least 48 hours to decide; refusal or withdrawal had no effect on grades, services, or teacher relationships; no performance-linked payment was offered; and students could use a private opt-out route through an independent school counsellor. Students could skip a personal reflection or use the offline pathway without penalty. Data were coded, access-restricted, and reported in aggregate. No direct identifiers, health information, or private identity narratives were entered into the AI interface. Procedures were designed to accord with the privacy, voluntariness, risk proportionality, assent, and safeguarding principles of the 2024 Declaration of Helsinki [40]. Consent and reference to Helsinki do not substitute for prospective ethics-committee review. The trial was not prospectively or retrospectively registered, no registration number is

claimed, and the analysis plan was not publicly time-stamped before analysis.

Statistical Analysis

Analyses followed an intention-to-treat estimand and included all 106 randomized students. Descriptive statistics summarized recruitment, participant characteristics, distribution shape, missingness, attendance, fidelity, acceptability, and harms. Baseline balance was evaluated with standardized mean differences rather than significance tests. Scale reliability was estimated with alpha and omega. Content-validity indices, rating dispersion, modified kappa, and Fleiss kappa were calculated as defined above. For each outcome, a restricted-maximum-likelihood mixed model included condition, categorical time, condition-by-time interaction, matched school pair, grade, and the baseline outcome; random intercepts represented class and repeated observations within student. The model supplied covariate-adjusted marginal mean differences, but conventional model-based standard errors were not used as the sole inferential basis. Primary follow-up inference used class-cluster-robust CR2 covariance estimation with Satterthwaite degrees of freedom; CR3 was a more conservative jackknife-style sensitivity analysis [44]. Restricted wild-cluster bootstrap-t intervals used Rademacher weights and enumerated all $2^{12}=4096$ cluster sign patterns [45]. Kenward-Roger inference was retained only as a comparative sensitivity analysis. A global multivariate test combined the three standardized follow-up score statistics and was calibrated by exact permutation across all 64 allocations allowed by the six matched pairs. Conditional on this gatekeeping test, raw follow-up p values for the three outcomes were adjusted by Holm's procedure. The correlated outcomes were not analyzed with false-discovery-rate control because the family was confirmatory; immediate post-intervention contrasts were labeled supportive.

Standardized effects were computed from the same model as $gMB = J(\nu)\Delta / \sqrt{(\tau^2 class + \sigma^2 student)}$, where Δ is the adjusted marginal mean difference, $\tau^2 class$ and $\sigma^2 student$ are unconditional time-specific class and student variances, and $J(\nu)$ is the Hedges correction based on the effective CR2 degrees of freedom. Thus, class-level heterogeneity entered the denominator directly. Baseline-pooled-SD Hedges g was not the principal effect size. Unconditional variance components and $ICC = \tau^2 class / (\tau^2 class + \sigma^2 student)$ were reported at every time point.

Maximum likelihood addressed incomplete repeated outcomes under missing at random. Multilevel multiple imputation used 50 datasets in wide format with a class random intercept (two-level pan specification), fixed matched-pair indicators, treatment arm, all repeated outcomes, baseline covariates, attendance, prior grade, AI-use history, and observed predictors of missingness. Imputation was conducted within randomized arm to preserve treatment-by-time structure. Because only 12 clusters can destabilize multilevel imputation, adjusted-group-mean fully conditional specification was added as a distinct sensitivity analysis [46]. A multidirectional tipping-point grid shifted missing NEXUS outcomes from -0.10 to -1.00 total SD and missing control outcomes from -0.50 to +0.50 SD in 0.10-SD increments. Further analyses included complete cases, attendance of at least eight sessions, leave-one-pair-out models, CR3, exact matched-pair permutation, and a cluster-level rank test.

Diagnostics included student-level and class-level quantile-quantile plots, scale-location plots, leverage and influence summaries, and inspection of empirical cluster means. No low-powered formal normality test was used to certify assumptions. Secondary blinded behavioral outcomes used analogous small-cluster inference but remained secondary because no prospective registry or publicly time-stamped protocol justified post hoc promotion to primary status. All tests were two-sided; adjusted $p < 0.05$ was the decision threshold, and confidence intervals were interpreted as the principal description of precision.

Results and Discussion

Results

Needs Mapping and Face Validity:

The screening distributions showed sufficient dispersion to support empirical programme-selection thresholds. Mean agentic engagement was 4.08 (SD=0.92; observed range=1.80-6.70), mean intrinsic motivation was 4.31 (SD=1.04; range=1.42-6.83), and mean critical-thinking disposition was 116.8 (SD=17.6; range=68-158). The three primary constructs were positively but not redundantly associated: agentic engagement correlated with intrinsic motivation at r=0.48, with critical-thinking disposition at r=0.37, and intrinsic motivation correlated with critical-thinking disposition at r=0.41. These relationships supported a multicomponent intervention rather than treating the constructs as interchangeable.

Application of the prespecified lower-quartile rule identified 121 of 463 students (26.1%). Fifty-three students (11.4%) met the threshold on exactly two outcomes and 68 (14.7%) met it on all three. Students classified as having elevated developmental need reported the greatest difficulty initiating questions, sustaining interest when a task lacked immediate relevance, tolerating ambiguity, seeking counterevidence, and revising confidence. Co-design participants recommended short repeated cycles, visible decision trails, student-selected topics, offline alternatives, and a clear separation between voluntary reflection and compelled personal disclosure.

Cognitive interviews showed that several students initially interpreted "critical thinking" as criticism of another person and "AI partner" as permission to delegate drafting. The manual was revised to define critical thinking through observable actions and to describe the model as a "bounded dialectical coach" in facilitator instructions. Identity tasks were reframed as optional reflection on learning patterns, values, and possible selves. Formal face-validity impact scores ranged from 3.20 to 4.73, exceeding the prespecified criterion of 1.50 for every instruction and activity.

Delphi Content Validation

Version 0.7 contained 58 separately rated content and implementation elements. In round 1, CVR ranged from 0.20 to 1.00, I-CVI from 0.67 to 1.00, and S-CVI/Ave was 0.84. Item-level rating variance had a median of 0.58 (interquartile range 0.44-0.73). Six elements were removed because they were duplicative, weakly linked to the theory of change, or judged unnecessarily burdensome; seven were substantively revised, and three explicit safety elements were added: mandatory de-identification, an equivalent offline pathway, and facilitator escalation for harmful or developmentally inappropriate output. Version 0.9 therefore contained 55 elements. All 55 elements met the minimum retention criteria in round 2, but agreement was not portrayed as unanimity. CVR ranged from 0.60 to 1.00 (median 0.73), I-CVI ranged from 0.80 to 1.00, and S-CVI/Ave was 0.91. Universal agreement was 17/55=0.31. Modified kappa ranged from 0.76 to 0.99. Fleiss kappa was 0.66 (95% CI 0.61-0.70) for relevance and 0.61 (95% CI 0.56-0.66) for necessity, compared with 0.58 and 0.52 in round 1. Median item-level variance declined to 0.41 (interquartile range 0.30-0.55), indicating convergence without erasing minority positions. Nine retained items received a relevance rating below 3 from two or three experts, and six received a non-essential judgment from two or three experts. Excluding the three panelists with prior academic collaboration yielded S-CVI/Ave=0.90. Minor edits standardized terminology, added examples of acceptable uncertainty statements, and clarified that facilitators not the model retained responsibility for safeguarding decisions.

Table 3. Final content-validity results by protocol domain

Domain	Elements, n	Median CVR (range)	Mean I-CVI (range)	Median non-essential ratings per item (range)
AI-mediated Socratic inquiry	12	0.87 (0.60-1.00)	0.92 (0.87-1.00)	1 (0-3)
Evidence appraisal and critical thinking	10	0.73 (0.60-1.00)	0.90 (0.80-1.00)	1 (0-3)
Agency and autonomy support	8	0.87 (0.73-1.00)	0.93 (0.87-1.00)	1 (0-2)
Identity meaning-making	8	0.73 (0.60-0.87)	0.88 (0.80-0.93)	2 (1-3)
Impactful expression and peer critique	7	0.73 (0.60-1.00)	0.90 (0.80-1.00)	1 (0-3)
AI safety, privacy, and fidelity	10	0.87 (0.73-1.00)	0.94 (0.87-1.00)	1 (0-2)
Overall	55	0.73 (0.60-1.00)	0.91 (0.80-1.00)	1 (0-3)

CVR, content validity ratio; I-CVI, item-level content validity index. Overall S-CVI/Ave=0.91; S-CVI/UA=0.31; modified kappa range=0.76-0.99. Fleiss kappa: relevance 0.66 (95% CI 0.61-0.70), necessity 0.61 (95% CI 0.56-0.66). Item-level distributions and minority comments are retained in Supporting Information.

Trial Flow and Baseline Characteristics

Of 117 trial-eligible students attending the selected matched class clusters, 110 (94.0%) provided parental permission and adolescent assent. Four withdrew before baseline because of timetable conflict (n=3) or a planned transfer (n=1). Twelve class clusters containing 106 students were randomized: six clusters (n=53) to NEXUS and six clusters (n=53) to active control. At post-intervention, data were available for 51 NEXUS students and 49 control students. At three-month follow-up, data were available for 49 and 46 students, respectively, corresponding to 89.6% overall retention. NEXUS follow-up losses reflected school transfer (n=1), examination conflict (n=2), and assent withdrawal (n=1); control losses reflected transfer (n=2), examination conflict (n=2), assent withdrawal (n=1), and inability to contact the family (n=2). No cluster withdrew. The modestly greater individual loss in control was described rather than tested because only 11 follow-up observations were missing. Baseline characteristics were acceptably balanced. Absolute standardized mean differences ranged from 0.00 to 0.13 and did not exceed the

prespecified 0.20 descriptive flag. Participants had a mean age of 16.4 years, approximately one-third were enrolled in each grade, and slightly more than half reported weekly GenAI use before the study. Figure 2 provides the complete participant flow and Table 4 presents baseline data.

Table 4. Baseline characteristics of randomized participants

Characteristic	NEXUS (n = 53)	Active control (n = 53)	Absolute SMD
Age, years, mean (SD)	16.40 (0.82)	16.44 (0.86)	0.05
Grade 10, n (%)	18 (34.0)	17 (32.1)	0.04
Grade 11, n (%)	17 (32.1)	18 (34.0)	0.04
Grade 12, n (%)	18 (34.0)	18 (34.0)	0.00
Prior weekly GenAI use, n (%)	29 (54.7)	31 (58.5)	0.08
Personal smartphone access, n (%)	49 (92.5)	48 (90.6)	0.07
Prior-term grade average, 0-20, mean (SD)	16.31 (1.42)	16.24 (1.47)	0.05
Agentic engagement, 1-7, mean (SD)	3.18 (0.79)	3.24 (0.81)	0.07
Intrinsic motivation, 1-7, mean (SD)	3.45 (0.84)	3.50 (0.86)	0.06
Critical-thinking disposition, 33-165, mean (SD)	101.9 (13.0)	102.6 (13.2)	0.05

SMD, standardized mean difference; GenAI, generative artificial intelligence. Percentages may not total 100 because of rounding.

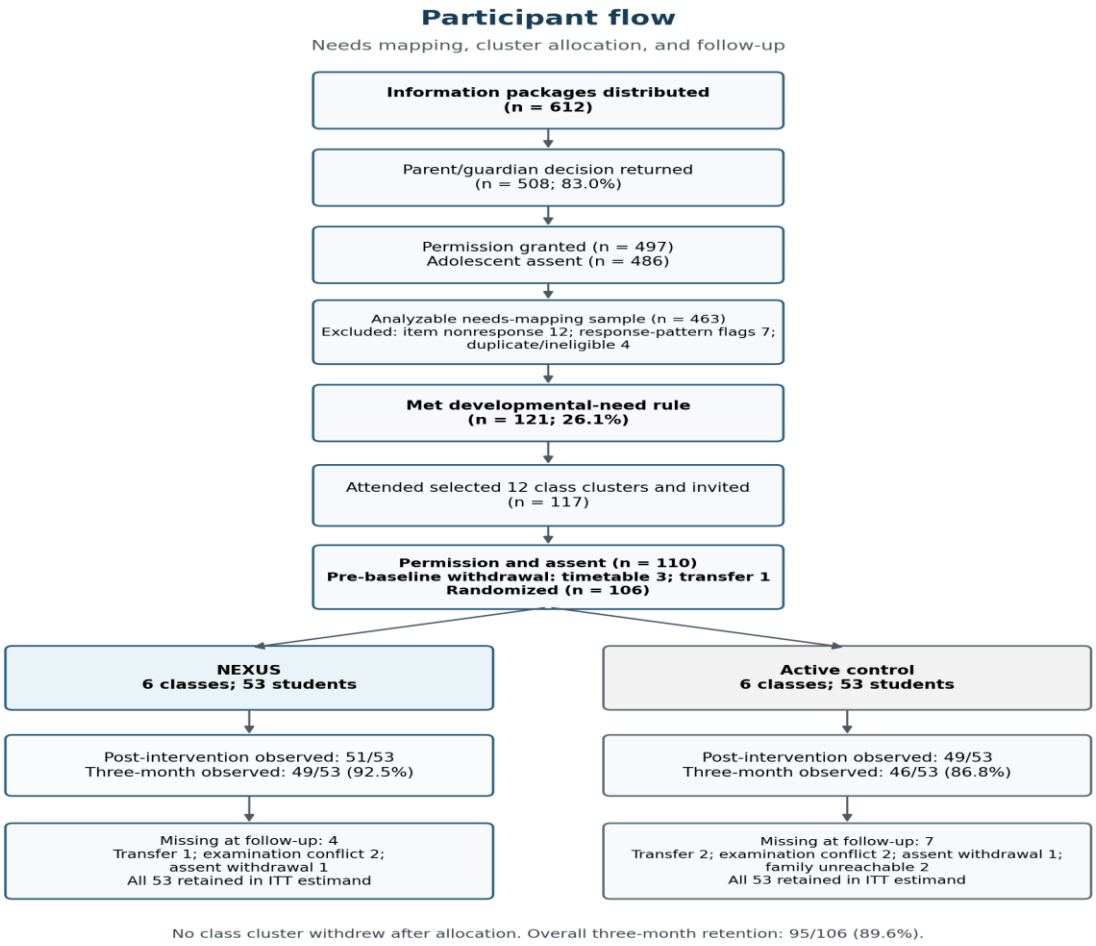

Fig. 2. Participant flow through needs mapping, eligibility, cluster allocation, post-intervention assessment, and three-month follow-up

Implementation, Fidelity, Acceptability, and Safety

NEXUS students attended a mean of 8.7 sessions (SD=1.5); 45 of 53 (84.9%) attended at least eight. Control students attended a mean of 8.5 sessions (SD=1.6); 43 of 53 (81.1%) attended at least eight. NEXUS fidelity averaged 86.8% (session range 78.4%-94.3%), and control fidelity averaged 88.1% (81.0%-94.8%). Independent raters audited 24 of 120 sessions, stratified by condition and session number; absolute-agreement ICC was 0.82 (95% CI 0.67-0.91). The broader range and imperfect agreement were retained because fidelity judgments involved human interpretation rather than mechanical box checking.

NEXUS acceptability was 4.08 (SD=0.56), perceived relevance was 4.16 (SD=0.55), and perceived autonomy support was 4.12 (SD=0.61). Corresponding active-control values were 3.89 (SD=0.59), 3.94 (SD=0.58), and 3.82 (SD=0.64). Mean cognitive load was 2.91 (SD=0.66) in NEXUS and 2.55 (SD=0.63) in control. The NEXUS tasks were therefore experienced as more demanding, although mean burden remained below the progression threshold of 3.50.

Across 1,187 logged NEXUS interactions, 93 direct-answer requests were detected. The automatic Socratic gate redirected 78 (83.9%); the remaining 15 were intercepted by facilitators before students incorporated a direct answer. Audit identified 21 inappropriate or overconfident outputs (1.8%): nine unverifiable citations, six instances of unjustified certainty, four instances of drift toward drafting the final product, and two incomplete privacy warnings. Twenty were corrected before the student product was finalized, and one was retained only with an explicit uncertainty label. Students independently flagged 104 potentially unsupported claims; 87 (83.7%) were verified, corrected, qualified, or removed, and 17 remained explicitly unresolved. One privacy near miss was blocked before transmission. These observations show that the safeguards reduced but did not eliminate model and user error. No serious intervention-related adverse event occurred. Six NEXUS students reported transient frustration, two reported discomforts during optional identity reflection, and the privacy near miss generated a facilitator debrief; three control students reported technical frustration. All were minor and resolved through clarification, topic change, or the offline pathway. No safeguarding disclosure was generated by the intervention and no external referral was required. Three control students accessed Socratic-prompt examples outside class; a contamination sensitivity analysis remained directionally similar but imprecise.

Table 5. Implementation and safety outcomes

Indicator	Progression criterion	NEXUS	Active control
Recruitment among invited students	>= 80%	94.0% overall	94.0% overall
Three-month completion	>= 85%	92.5% (49/53)	86.8% (46/53)
Attendance at >= 8 sessions	>= 75%	84.9% (45/53)	81.1% (43/53)
Mean implementation fidelity	>= 80%	86.8%	88.1%
Mean acceptability, 1-5	>= 3.5	4.08 (0.56)	3.89 (0.59)
Mean cognitive load, 1-5	< 3.5	2.91 (0.66)	2.55 (0.63)
Serious related adverse events	0	0	0

Primary Effectiveness Outcomes

The exact global multivariate randomization test for the three follow-up outcomes was p=0.031, supporting a joint between-condition shift while leaving outcome-specific precision to be evaluated separately. Agentic engagement showed an adjusted follow-up difference of 0.44 points (restricted wild-cluster-bootstrap 95% CI 0.10 to 0.78; raw p=0.014; Holm-adjusted p=0.042; model-based g=0.43). Critical-thinking disposition showed a 6.1-point difference (95% CI 0.6 to 11.6; raw p=0.039; adjusted p=0.078; g=0.38). Intrinsic motivation showed a 0.31-point difference (95% CI -0.04 to 0.66; raw and adjusted p=0.086; g=0.30). Thus, only agentic engagement crossed the multiplicity-adjusted outcome-specific threshold; the intervals for motivation and critical-thinking disposition remained compatible with smaller or null effects.

Supportive post-intervention differences were 0.43 for agentic engagement (95% CI 0.08 to 0.78; wild-bootstrap p=0.024; g=0.44), 0.30 for intrinsic motivation (95% CI -0.05 to 0.65; p=0.082; g=0.30), and 6.7 for critical-thinking disposition (95% CI 1.2 to 12.2; p=0.021; g=0.42). These p values were unadjusted and exploratory. Importantly, the active control improved from baseline rather than remaining static: at follow-up, observed changes were +0.22, +0.15, and +3.5 points, respectively. Study-planning practice, facilitator attention, technology familiarity, repeated measurement, and ordinary academic development are plausible contributors to those gains; attributing them to a single mechanism is not warranted. Student-level quantile-quantile plots showed mild tail departure, while six-arm cluster-level plots were necessarily sparse. Scale-location patterns did not indicate gross heteroscedasticity. One class exerted moderate

leverage, but leave-one-pair-out estimates remained positive and varied materially in width, as expected with only six pairs. These diagnostics motivated robust and randomization-based inference rather than a claim that normality had been established.

Table 6. Primary outcomes at baseline, post-intervention, and three-month follow-up

Outcome	Group	Baseline mean (SD)	Post mean (SD)	Follow-up mean (SD)	Adjusted NEXUS-minus-control contrast; model-based g
Agentic engagement, 1-7	NEXUS	3.18 (0.79)	3.84 (0.88)	3.78 (0.90)	Post: 0.43 (0.08 to 0.78); g = 0.44. Follow-up: 0.44 (0.10 to 0.78); g = 0.43
-	Active control	3.24 (0.81)	3.49 (0.84)	3.46 (0.87)	Reference
Intrinsic motivation, 1-7	NEXUS	3.45 (0.84)	3.94 (0.90)	3.88 (0.92)	Post: 0.30 (-0.05 to 0.65); g = 0.30. Follow-up: 0.31 (-0.04 to 0.66); g = 0.30
-	Active control	3.50 (0.86)	3.70 (0.88)	3.65 (0.91)	Reference
Critical-thinking disposition, 33-165	NEXUS	101.9 (13.0)	112.2 (14.1)	111.3 (14.5)	Post: 6.7 (1.2 to 12.2); g = 0.42. Follow-up: 6.1 (0.6 to 11.6); g = 0.38
-	Active control	102.6 (13.2)	106.4 (13.8)	106.1 (14.0)	Reference

Confidence intervals are restricted wild-cluster-bootstrap-t intervals. Follow-up raw p values were 0.014, 0.086, and 0.039; Holm-adjusted values were 0.042, 0.086, and 0.078. Post-intervention p values are supportive and unadjusted. Model-based g includes unconditional class and student variance and Hedges correction.

The unconditional ICCs were non-zero and increased modestly over time (Table 7). This pattern rules out an implicit zero-cluster-variance explanation for the confidence intervals and shows why using only a baseline or within-student SD would overstate standardized effects.

Table 7. Unconditional time-specific variance components and intraclass correlations

Outcome	Time	Class variance	Student variance	ICC	Total SD
Agentic engagement	Baseline	0.040	0.616	0.061	0.810
-	Post	0.059	0.717	0.076	0.881
-	Follow-up	0.069	0.762	0.083	0.912
Intrinsic motivation	Baseline	0.035	0.710	0.047	0.863
-	Post	0.052	0.761	0.064	0.902
-	Follow-up	0.060	0.785	0.071	0.919
Critical-thinking disposition	Baseline	12.0	162.0	0.069	13.19
-	Post	16.9	189.2	0.082	14.36
-	Follow-up	18.7	189.1	0.090	14.42

ICC = class variance/ (class variance + student variance); total SD is the square root of the summed variance components. Components came from outcome-specific unconditional models at each time point.

Secondary Behavioral Outcomes and Sensitivity Analyses

Blinded scoring of source-verification tasks showed inter-rater ICC=0.84 (95% CI 0.75-0.90), and written-product scoring showed ICC=0.86 (95% CI 0.78-0.92). At post-intervention, adjusted NEXUS-minus-control differences were 1.2 points for source verification (0-12 scale; 95% CI 0.2 to 2.2; g=0.42) and 1.8 for written-product quality (0-20 scale; 95% CI 0.3 to 3.3; g=0.46). Follow-up differences were 0.9 (95% CI -0.1 to 1.9; g=0.31) and 1.4 (95% CI -0.2 to 3.0; g=0.34), respectively. These blinded outcomes reduce sole reliance on self-report, but their follow-up intervals include no effect and their

secondary status was not changed after seeing the data. Sensitivity analyses did not support an all-or-none claim of robustness (Table 8). CR3, both imputation strategies, complete-case analysis, and per-protocol analysis preserved direction but widened uncertainty. Exact pairwise randomization p values were constrained by 64 possible allocations: 0.031 for agentic engagement, 0.125 for motivation, and 0.063 for critical-thinking disposition. Cluster-rank p values were 0.031, 0.156, and 0.063. Leave-one-pair-out point estimates remained positive, but the widest intervals crossed zero for every outcome in at least one omission, illustrating dependence on a small number of clusters. The multidirectional tipping-point analysis showed that the agentic-engagement interval lost its positive lower bound when missing NEXUS follow-up scores were approximately 0.55 total SD worse than predicted while missing control scores were 0.20 SD better. The corresponding adverse combination was approximately 0.65 and 0.20 SD for critical-thinking disposition. Intrinsic motivation already included zero under missing at random. These thresholds are more informative than a single one-directional -0.30-SD shift because they display simultaneous departures in both arms.

Table 8. Sensitivity analyses for three-month outcome-specific contrasts

Analysis	Agentic engagement, adjusted difference (95% CI)	Intrinsic motivation, adjusted difference (95% CI)	Critical-thinking disposition, adjusted difference (95% CI)
Mixed model point estimate with CR2/WCB interval	0.44 (0.10 to 0.78)	0.31 (-0.04 to 0.66)	6.1 (0.6 to 11.6)
CR3 cluster jackknife	0.43 (0.08 to 0.79)	0.30 (-0.07 to 0.67)	5.9 (-0.1 to 11.9)
Multilevel MI, 50 datasets	0.42 (0.07 to 0.77)	0.29 (-0.08 to 0.66)	5.8 (-0.2 to 11.8)
Adjusted-group-mean FCS MI	0.43 (0.08 to 0.78)	0.30 (-0.06 to 0.66)	5.9 (0.0 to 11.8)
Complete cases	0.46 (0.09 to 0.83)	0.33 (-0.06 to 0.72)	6.4 (0.1 to 12.7)
Per protocol, >= 8 sessions	0.47 (0.10 to 0.84)	0.34 (-0.05 to 0.73)	6.6 (0.2 to 13.0)
Exact matched-pair randomization, p value	0.031	0.125	0.063
Cluster-level rank test, p value	0.031	0.156	0.063

CR2/CR3, small-sample cluster-robust covariance corrections; WCB, restricted wild-cluster bootstrap-t; MI, multiple imputation; FCS, fully conditional specification. Model-based rows report adjusted differences and 95% confidence intervals. Randomization and rank rows report two-sided p values.

Discussion

This study developed and content-validated a ten-session NEXUS protocol and then estimated its preliminary effects in a comparison with a dose- and technology-matched active control. The main design contribution was a redistribution of epistemic authority: students had to formulate the question, expose provisional reasoning, evaluate evidence, consider a counter position, connect the task to a meaningful goal, and author the final communication. The model could scaffold these acts but was not intended to replace them. The results support feasibility and a promising joint follow-up signal, not definitive effectiveness across every outcome. Content-validation findings were credible precisely because they were not uniformly perfect. The final S-CVI/Ave of 0.91 met the progression criterion, whereas universal agreement of 0.31 and Fleiss kappa values of 0.66 and 0.61 showed meaningful residual disagreement. Reporting item-level variance, minority ratings, conflicts, and the exclusion sensitivity analysis makes clear how consensus was reached. Round-to-round improvement followed deletion, revision, and added safeguards rather than re-rating an unchanged instrument. These decisions align with current content-validation guidance [41], UNESCO's human-centred emphasis on agency, privacy, transparency, fairness, and age-appropriate AI competency [38,39], and evidence identifying hallucination, privacy exposure, inequity, and cognitive offloading as central K-12 risks [5].

At three months, the global multivariate randomization test supported a joint shift, but outcome-specific evidence differed. Agentic engagement retained a positive wild-bootstrap interval and survived Holm adjustment; its model-based g of 0.43 was far smaller than the baseline-SD estimate that would have resulted from ignoring class variance. This pattern is theoretically coherent because every session required learner-initiated questions, evidence requests, task-direction

decisions, and defended revisions behaviours closely aligned with agentic engagement [11,12]. Even so, the interval remains wide enough to encompass modest rather than transformative benefit. For intrinsic motivation, the follow-up interval included zero, and for critical-thinking disposition the raw interval excluded zero but the Holm-adjusted p value did not cross 0.05. These outcomes should therefore not be described as conclusively improved. Their positive point estimates are compatible with self-determination theory and structured higher-order-thinking accounts [3,4,13,14,22-24], but theoretical plausibility does not repair imprecision. The blinded behavioral measures were directionally concordant at post-intervention, yet their follow-up intervals also included zero. They strengthen triangulation while remaining secondary; promoting them after observing the results would compound selective-reporting risk. The active control improved on all three outcomes. This is plausible for ten sessions of study planning, memory strategy, digital organization, collaborative learning, and guided GenAI summarization. Adult attention, peer activity, repeated testing, practice, and normal academic development may also have contributed. The between-arm contrasts therefore estimate the incremental value of NEXUS over a credible educational comparator, not change against inactivity. This pattern reduces the concern created by an implausibly stationary control and appropriately yields smaller effect estimates.

Identity meaning-making remains a plausible mechanism connecting autonomous motivation and agency, but mediation was not tested. NEXUS prohibited automated personality interpretation, made personal disclosure optional, and allowed non-personal topics. These safeguards are important because identity interventions are intended to support exploration and commitment rather than impose fixed labels [18,19]. Future mechanism work should measure identity-goal integration directly instead of inferring it from outcome change.

Implementation was adequate rather than flawless. Fidelity averaged 86.8%, and independent agreement was 0.82. The Socratic gate intercepted most direct-answer requests but required human rescue in 15 cases. Auditors identified 21 inappropriate or overconfident outputs, and students did not resolve every unsupported claim. These errors are educationally consequential: a bounded system reduces risk only when automatic restrictions, facilitator audit, student verification, and an offline alternative operate together. The absence of a serious related event should not be interpreted as absence of minor burden or system failure. Several design features strengthen interpretation: intervention development preceded testing; adolescent, school-staff, and expert input were integrated; recruitment and baseline measurement preceded allocation release; the active control matched contact, device exposure, and general AI access; and behavioral products were blindly rated. Analytically, CR2/CR3, exhaustive restricted wild-cluster bootstrap, exact within-pair randomization, cluster-rank tests, time-specific ICCs, model-based standardization, multilevel imputation, and multidirectional tipping points made the small-cluster uncertainty visible instead of allowing a conventional mixed model to conceal it [42-46].

The limitations are nonetheless substantial.

First, 12 classes six per condition provide little independent information for variance-component estimation. CR2, CR3, bootstrap, and permutation methods improve calibration but cannot manufacture missing clusters. Effective degrees of freedom remain small, exact p values are coarse, and leave-one-pair-out intervals reveal fragility.

Second, participants were male students from urban upper-secondary schools in Qazvin; transfer to girls, rural schools, first-cycle students, or less-resourced settings is unknown.

Third, students knew their allocation. For self-report outcomes this creates performance, expectancy, and detection threats even though questionnaire administrators and analysts were masked. Blinded behavioral outcomes are valuable corroboration but cannot retrospectively replace the designated family.

Fourth, differential individual attrition was small but somewhat greater in control. Multilevel imputation and tipping-point analyses address observable and hypothetical departures, yet no missing-data method proves the unobserved values.

Fifth, cluster-level residual diagnostics are intrinsically weak with 12 classes; the observed mild tail departure and leverage reinforce reliance on robust and rank-based analyses.

Sixth, the three-month endpoint does not establish maintenance across an academic year.

Seventh, estimates remain conditional on facilitator skill and a specific archived model configuration; changes in model snapshot, temperature, system prompt, interface, or safety classifier may alter both learning and risk.

The most serious transparency limitations are the absence of a formally numbered prospective ethics approval, prospective trial registration, and a publicly time-stamped protocol and analysis plan. Parental permission, adolescent assent, administrative authorization, and reference to the Declaration of Helsinki do not substitute for independent ethics-committee review. Without prospective registration, selective outcome, contrast, or subgroup reporting cannot be excluded, and the results should not be presented as confirmatory. Before dissemination of any real-world replication, investigators should obtain prospective approval and registration. If a completed study is registered retrospectively, the registration must clearly identify

its retrospective date and should be accompanied by the original protocol, analysis code, full outcome matrix, and a transparent account of deviations; retrospective registration does not restore prospective protection.

A definitive evaluation requires substantially more independent schools, sex-inclusive and socioeconomically diverse recruitment, longer follow-up, and a prospectively registered hierarchy of estimands. Prespecified mechanism measures should include autonomy support, epistemic vigilance, confidence calibration, cognitive offloading, identity-goal integration, and revision quality. A process evaluation could examine whether source checking and learner-authored revision mediate later agency, while a dismantling design could separate the contributions of Socratic gating, identity work, and impactful expression. Cost, teacher workload, accessibility, privacy, offline equivalence, and model-version drift should be treated as core implementation outcomes.

Conclusion

NEXUS is a content-valid and implementable framework for converting GenAI from an answer generator into a bounded Socratic coach while retaining question formation, evidence appraisal, revision, and final authorship with the adolescent. In this small-cluster comparison, the joint follow-up test and the multiplicity-adjusted agentic-engagement result were promising; intrinsic motivation, critical-thinking disposition, and blinded behavioral outcomes remained uncertain. The model-based effects were modest, the active control also improved, and uncertainty was materially wider after small-cluster correction. NEXUS should therefore be treated as ready for prospectively approved, registered, multisite replication not as established evidence of broad effectiveness.

Disclosure Statement

No potential conflict of interest reported by the authors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions

All authors contributed to data analysis, drafting, and revising of the paper and agreed to be responsible for all the aspects of this work.

References

[1] L. Yan, S. Greiff, Z. Teuber, D. Gasevic, Promises and challenges of generative artificial intelligence for human learning, Nature Human Behaviour, 2024, 8, 1839-1850.

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

[3] M.Y.I. Helal, I.A. Elgendy, M.A. Alb Ashrawi, Y.K. Dwivedi, M.S. Al-Ahmadi, I. Jeon, the impact of generative AI on critical thinking skills: A systematic review, conceptual framework and future research directions, Information Discovery and Delivery, 2025, advance online publication. https://doi.org/10.1108/IDD-05-2025-0125

[4] N. Cong-Lem, N.T. Thuy-Dung, towards a framework for understanding and assessing critical thinking in generative AI-enhanced learning environments: A scoping review, European Journal of Psychology of Education, 2026, 41(2), Article 41. https://doi.org/10.1007/s10212-026-01089-y

[5] S. Tao, M. Lan, M. Wang, H. Li, Potential risks of generative artificial intelligence integration into K-12 education: A scoping review, Computers and Education: Artificial Intelligence, 2026, 10, 100561. https://doi.org/10.1016/j.caeai.2026.100561

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

[7] G. Brod, Agency does not equal choice: Conceptualizing agency for learning in the age of AI, Learning and Individual Differences, 2026, 125, 102841. https://doi.org/10.1016/j.lindif.2025.102841

[8] G. Brod, E. Holstein, L. Weindorf, J. Colantonio, E. Bonawitz, M. Theobald, do it yourself: Discerning the effects of self-directed activity on conceptual learning, npj Science of Learning, 2025, 10, 70. https://doi.org/10.1038/s41539-025-00364-9

[9] P.-B. Degen, I. Asanov, beyond automation: Socratic AI, epistemic agency, and the implications of the emergence of orchestrated multi-agent learning architectures, arXiv, 2025. https://doi.org/10.48550/arXiv.2508.05116

[10] J. Lee, J.-T. Hung, M.Y. Soylu, D. Popescu, C.Z. Cui, G. Grigoryan, D.A. Joyner, S.W. Harmon, Socratic Mind: Impact of a novel GenAI-powered assessment tool on student learning and

higher-order thinking, arXiv, 2025. https://doi.org/10.48550/arXiv.2509.16262
[11] J. Reeve, C.-M. Tseng, Agency as a fourth aspect of students' engagement during learning activities, Contemporary Educational Psychology, 2011, 36(4), 257-267. https://doi.org/10.1016/j.cedpsych.2011.05.002
[12] C. Mameli, S. Passini, Development and validation of an enlarged version of the Student Agentic Engagement Scale, Journal of Psychoeducational Assessment, 2019, 37(4), 450-463. https://doi.org/10.1177/0734282918757849
[13] R.M. Ryan, E.L. Deci, Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions, Contemporary Educational Psychology, 2020, 61, 101860. https://doi.org/10.1016/j.cedpsych.2020.101860
[14] X. Fang, J. Feng, Research on the application behavior of generative artificial intelligence learning of college students based on self-determination theory, Frontiers in Psychology, 2026, 17, 1805498. https://doi.org/10.3389/fpsyg.2026.1805498
[15] H. Galindo-Dominguez, N. Delgado, M.V. Urruzola, J.M. Etxabe, L. Campo, using artificial intelligence to promote adolescents' learning motivation: A longitudinal intervention from self-determination theory, Journal of Computer Assisted Learning, 2025, 41(2), e70020. https://doi.org/10.1111/jcal.70020
[16] L. Zhang, W. Zhang, integrating large language models into project-based learning based on self-determination theory, Interactive Learning Environments, 2025, 33(5), 3580-3592. https://doi.org/10.1080/10494820.2024.2446535
[17] S. Choi, H. Kim, The impact of a large language model-based programming learning environment on students' motivation and programming ability, Education and Information Technologies, 2025, 30(6), 8109-8138. https://doi.org/10.1007/s10639-024-13107-x
[18] E. Crocetti, M. Pagano, F. De Lise, F. Maratia, B. Bobba, W. Meeus, V. Bacaro, Promoting adolescents' personal and social identities: A meta-analysis of psychosocial interventions, Identity, 2025, 25(2), 167-196. https://doi.org/10.1080/15283488.2024.2394888
[19] E. Crocetti, F. De Lise, M. Pagano, B. Bobba, F. Maratia, W. Meeus, V. Bacaro, What's your vocation? A meta-analysis of interventions tackling youth educational and professional identity, Identity, 2025, 25(2), 197-215. https://doi.org/10.1080/15283488.2024.2394861
[20] Rezai, R. Ahmadi, P. Ashkani, G.H. Hosseini, implementing active learning approach to promote motivation, reduce anxiety, and shape positive attitudes: A case study of EFL learners, Acta Psychologica, 2025, 253, 104704. https://doi.org/10.1016/j.actpsy.2025.104704
[21] J.A. Alqurni, Exploring the role of agentic AI in fostering self-efficacy, autonomy support, and self-learning motivation in higher education, Frontiers in Artificial Intelligence, 2026, 9, 1738774. https://doi.org/10.3389/frai.2026.1738774
[22] M. Badali, M. Shah Alizadeh, M. Alimirzaie, O. Noroozi, S.K. Bani Hashem, Bridging AI chatbot use and critical thinking: The mediating role of AI literacy, Interactive Learning Environments, 2026, advance online publication. https://doi.org/10.1080/10494820.2025.2605487
[23] Y. Zhao, Y. Yue, Z. Sun, Q. Jiang, G. Li, does generative artificial intelligence improve students' higher-order thinking? A meta-analysis based on 29 experiments and quasi-experiments, Journal of Intelligence, 2025, 13(12), 160. https://doi.org/10.3390/jintelligence13120160
[24] F. Shirdel, N. Mobasheri, M.H. Kaveh, J. Hassanzadeh, L. Ghahremani, A randomized controlled trial to evaluate the effectiveness of a theory of planned behavior-based educational intervention in reducing internet addiction among adolescent girls in southern Iran, Adolescents, 2025, 5(3), 33. https://doi.org/10.3390/adolescents5030033
[25] S. Bashirinia, N. Motamed, M. Bahreini, M. Ravani pour, The effects of an empowering self-management model on sense of coherence and self-efficacy among adolescent girls with internet addiction in Bushehr, Iran: A randomized controlled trial, BMC Public Health, 2026, 26, 424. https://doi.org/10.1186/s12889-025-26117-2
[26] K. Skivington, L. Matthews, S.A. Simpson, P. Craig, J. Baird, J.M. Blazeby, K.A. Boyd, N. Craig, D.P. French, E. McIntosh, M. Petticrew, J. Rycroft-Malone, M. White, L. Moore, A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance, BMJ, 2021, 374, n2061. https://doi.org/10.1136/bmj.n2061
[27] T.C. Hoffmann, P.P. Glasziou, I. Boutron, R. Milne, R. Perera, D. Moher, D.G. Altman, V. Barbour, H. Macdonald, M. Johnston, S.E. Lamb, M. Dixon-Woods, P. McCulloch, J.C. Wyatt, A.-W. Chan, S. Michie, better reporting of interventions: Template for Intervention Description and Replication (Tidier) checklist and guide, BMJ, 2014, 348, g1687. https://doi.org/10.1136/bmj.g1687
[28] A.-W. Chan, I. Boutron, S. Hopewell, D. Moher, K.F. Schulz, G.S. Collins, et al., SPIRIT 2025 statement: Updated guideline for protocols of randomized trials, BMJ, 2025, 389, e081477. https://doi.org/10.1136/bmj-2024-081477
[29] S. Hopewell, A.-W. Chan, G.S. Collins, A. Hrobjartsson, D. Moher, K.F. Schulz, et al., CONSORT 2025 statement: Updated guideline for

reporting randomized trials, BMJ, 2025, 389, e081123.
https://doi.org/10.1136/bmj-2024-081123
[30] Baba, M. Smith, B.K. Potter, A.-W. Chan, D. Moher, A. Toulany, et al., CONSORT-Children and Adolescents (CONSORT-C) 2026 extension statement: Enhancing the reporting and impact of pediatric randomized trials, JAMA Pediatrics, 2026, 180(4), 413-425.
https://doi.org/10.1001/jamapediatrics.2026.0104
[31] R.J. Vallerand, L.G. Pelletier, M.R. Blais, N.M. Briere, C. Senecal, E.F. Vallieres, The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education, Educational and Psychological Measurement, 1992, 52(4), 1003-1017.
https://doi.org/10.1177/0013164492052004025
[32] N. Bagheri, M. Shahraray, V. Farzad, Standardization of Academic Motivation Scale (AMS) among the high school students in Tehran, Clinical Psychology and Personality, 2003, 1(1), 11-24.
https://cpap.shahed.ac.ir/article_2771.html?lang=en
[33] J. Cavoukian, P. Kadivar, M. Shahraray, A.A. Sheikhi Fini, V. Farzad, Standardization of the Vallerand Academic Motivation Scale, Journal of Educational Sciences, 2010, 5(4), 103-130. https://www.noormags.ir/view/fa/articlepage/922938/
[34] J.C. Ricketts, The Efficacy of Leadership Development, Critical Thinking Dispositions, and Student Academic Performance, doctoral dissertation, University of Florida, Gainesville, 2003.
https://ufdc.ufl.edu/UFE0000777/00001
[35] H. Pakmehr, F. Mirdoraghi, A. Ghanaei Chamanabad, M. Karami, Reliability, validity and factor analysis of Ricketts' Critical Thinking Disposition Scales in high school, Quarterly of Educational Measurement, 2013, 4(11), 33-54. https://jem.atu.ac.ir/article_2678_en.html
[36] F. Miao, W. Holmes, Guidance for Generative AI in Education and Research, UNESCO, Paris, 2023. https://unesdoc.unesco.org/ark:/48223/pf0000386693
[37] F. Miao, K. Shiohira, N. Lao, AI Competency Framework for Students, UNESCO, Paris, 2024. https://unesdoc.unesco.org/ark:/48223/pf0000391105
[38] S. Gupta, S. Grover, V. Menon, P.V. Indu, K. Vidhu Kumar, D. Chacko, Item generation and establishing face and content validity of a rating scale: A primer, Indian Journal of Psychiatry, 2025, 67(8), 816-822.
https://doi.org/10.4103/indianjpsychiatry_750_25
[39] O. Quintin, J. Kasza, N. Davies, S. Markham, R. Hooper, Small numbers of clusters in cluster randomized trials: A scoping review of problems and proposed solutions, Journal of Clinical Epidemiology, 2026, online ahead of print, 112395.
https://doi.org/10.1016/j.jclinepi.2026.112395
[40] A.-L. Schubert, M. Steinhilber, H. Kang, D.S. Quintana, Improving statistical reporting in psychology, Communications Psychology, 2025, 3, 156.
https://doi.org/10.1038/s44271-025-00356-w
[41] F.L. Huang, when cluster-robust inferences fail, Educational and Psychological Measurement, 2026, 86(3), 579-601.
https://doi.org/10.1177/00131644251393203
[42] I.A. Canay, A. Santos, A.M. Shaikh, The wild bootstrap with a "small" number of "large" clusters, Review of Economics and Statistics, 2021, 103(2), 346-363.
https://doi.org/10.1162/rest_a_00887
[43] S. Grund, O. Lüdtke, A. Robitzsch, Multiple imputation of multilevel data with single-level models: A fully conditional specification approach using adjusted group means, Behavior Research Methods, 2026, 58, 76.
https://doi.org/10.3758/s13428-025-02915-9