Evaluating the Impact of Generative Artificial Intelligence on Academic Performance and Durable Learning:

A Conceptual Framework and Synthesis

Aditi Patil[#1], Pratiksha Pawar[#2], Ananta Raut[#3], Prof. Bhupali Bhattu[*4]

[#]*Student, Master of Computer Applications (MCA) Department, ZIBACAR, Affiliated to Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India*

[*]*Assistant Professor, Master of Computer Applications (MCA) Department, ZIBACAR, Affiliated to Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India*

[1]patiladiti9090@gmail.com | [2]pratupawar30@gmail.com | [3]anantaraut7020@gmail.com | [4]bhupali.bhattu@zealeducation.com

ORCID: Aditi Patil: 0009-0009-3944-9634 | Pratiksha Pawar: 0009-0004-9204-6045 | Ananta Raut: 0009-0000-8994-2368

Abstract*—Generative Artificial Intelligence (GenAI) systems have achieved pervasive adoption across higher education, frequently outpacing institutional policy formulation and pedagogical redesign. Although an expanding corpus of empirical literature investigates the utility of large language models in educational settings, findings remain fragmented across diverse methodologies, sample demographics, and outcome metrics. This paper synthesizes empirical findings, meta-analyses, scoping reviews, and policy frameworks published between 2025 and 2026 to address a fundamental question: Under what pedagogical and cognitive conditions does generative AI facilitate durable learning gains rather than transitory performance inflation? Synthesized evidence reveals that while GenAI interventions yield moderate immediate performance gains ($g = 0.57$), these effects exhibit marked sensitivity to methodological design. Reported effect sizes are systematically inflated in smaller cohorts ($g = 1.22$) and investigations relying on unstandardized, researcher-authored assessments ($g = 0.98$); crucially, performance advantages attenuate significantly once AI scaffolds are removed. Lasting learning gains are observed predominantly when tools are engineered for guided inquiry rather than task substitution. Based on these findings, this study delivers three conceptual contributions: (1) a Four-Mode Interaction Taxonomy (Scaffold, Check, Draft-assist, Substitute) classifying student-AI interaction based on locus of cognitive effort; (2) the Retention Ratio ($RR = U/A$), a parsimonious metric designed to decouple assisted execution (A) from unassisted competence (U); and (3) a set of testable empirical propositions detailing how prompt architecture, scaffolding decay, and institutional disparities govern cognitive persistence. Finally, we articulate actionable recommendations for curriculum designers, institutional leadership, and researchers, with particular attention to socio-technical considerations within developing and linguistically diverse educational environments.

Keywords*—Agentic systems, cognitive offloading, desirable difficulties, digital equity, durable learning, educational assessment, generative AI, intelligent tutoring systems.

I. INTRODUCTION

Large language models (LLMs) became widely available in late 2022. Initially, many institutions implemented restrictive measures to preserve academic integrity. However, widespread use of LLMs by students and teaching staff quickly proved restrictive approaches to be ineffective. By 2026, it has become clear that the debate surrounding whether students should be allowed to use generative AI tools in learning has moved beyond simple prohibition. What remains to be determined by educators is how they can distinguish between students who have developed accurate mental representations and students who have used generative AI merely to simulate knowledge.

In written asynchronous assessment tasks, distinguishing between these cohorts is difficult. Text generated using naive AI interactions may be highly coherent, complex in structure, and display analytical breadth comparable to work produced by students who reasoned independently through their answers. However, the underlying cognitive processes differ fundamentally. Students who complete tasks without AI support must internally construct knowledge representations, retrieve prior schemas, and organize information. Conversely, students who rely on naive AI interactions outsource these essential cognitive operations directly to the model.

Increasing empirical research in this domain has documented significant divergences in student performance during AI-assisted work versus after discontinuing AI support. Although AI assistance can improve immediate efficiency and help resolve localized conceptual bottlenecks, over-reliance prevents learners from building a robust foundational knowledge base. When AI support is abruptly withdrawn from learners who engaged in task substitution, unassisted performance degrades sharply. To establish the conditions under which Generative AI (GenAI) functions as an instructional scaffold rather than a cognitive crutch, recent empirical findings must be anchored within established cognitive and learning theories.

A. Paper Structure

The remainder of this paper is organized as follows: Section II examines foundational cognitive and socio-technical learning theories. Section III reviews empirical studies, meta-analyses, and policy frameworks published between 2025 and 2026. Section IV formalizes the research objectives and testable propositions. Section V outlines the qualitative narrative synthesis methodology. Section VI maps and evaluates the reviewed evidence base. Section VII introduces the Four-Mode Interaction Taxonomy, the Retention Ratio (*RR*), falsifiable predictions, and a practitioner checklist. Section VIII discusses epistemic illusions, institutional trajectories, and contextual realities in Indian higher education. Sections IX through XII present multi-tier recommendations, a future research agenda, methodological limitations, and concluding remarks.

 ISSN: 3139-1141

II. Theoretical Foundation

Although foundational instructional and cognitive theories were formulated decades before modern transformer architectures, their core mechanisms remain directly applicable for analyzing human cognition in AI-mediated learning environments.

A. Cognitive Load Theory

Cognitive Load Theory, formulated by Sweller [1], posits that human working memory possesses a strictly limited processing capacity, whereas long-term memory offers near-infinite capacity where knowledge is structured into organized schemas. Instructional efficiency depends on managing three distinct categories of cognitive load:

- *Intrinsic Cognitive Load:* Determined by the inherent element interactivity and conceptual complexity of the subject matter.
- *Extraneous Cognitive Load:* Generated by suboptimal instructional design, ambiguous formatting, or irrelevant procedural friction.
- *Germane Cognitive Load:* The productive mental effort dedicated to schema construction, integration, and cognitive automaticity.

In generative AI contexts, an important cognitive trade-off emerges. GenAI can productively reduce extraneous load by clarifying unfamiliar terminology, translating complex syntax, and organizing reference material. However, when the model synthesizes arguments, resolves contradictions, or generates complete structural solutions, it eliminates essential element interactivity for the learner. Cognitive offloading then ceases to support efficiency; germane processing is bypassed, and information in working memory fails to consolidate into durable long-term schemas.

B. Scaffolding Dynamics within the Zone of Proximal Development

Vygotsky's [2] socio-cultural construct of the Zone of Proximal Development (ZPD) defines the distance between a learner's independent problem-solving capacity and the level of performance achievable under expert guidance. As formalized by Wood, Bruner, and Ross [3], authentic instructional scaffolding requires three operational attributes: (1) dynamic calibration to the learner's emergent developmental state; (2) incremental fading as the learner internalizes procedural competence; and (3) ultimate transfer of operational control to the student.

Recent policy evaluations by the OECD [4] note that modern conversational agents often simulate initial tutoring behaviors by posing leading questions, offering hints, and prompting reflection. Yet commercial generative interfaces are engineered primarily to maximize user satisfaction and frictionless engagement rather than pedagogical rigor. When conversational agents fail to execute programmatic fading—maintaining continuous, unattenuated assistance across iterations—the mechanism ceases to act as a scaffold and devolves into a permanent cognitive prosthesis, leaving independent competence unchanged.

C. Desirable Difficulties and Epistemic Friction

Bjork [5] demonstrated that instructional conditions introducing *desirable difficulties*—such as spaced practice, interleaving, unassisted testing, and effortful retrieval—slow down immediate acquisition and depress short-term performance metrics, yet substantially strengthen long-term retention and transfer. Conversely, learning environments that eliminate productive struggle foster an illusion of competence while producing fragile memory traces. Because unrestricted generative AI removes epistemic friction, students complete deliverables rapidly without engaging in effortful memory retrieval, resulting in steep performance drops once tool access is removed.

D. Socio-Technical Systems and Structural Disparities

Educational technology adoption does not unfold in institutional isolation. Fournier and Łodzikowski [6] articulate a three-tension framework governing agentic AI deployment, highlighting structural conflicts among: (1) technological innovation imperatives; (2) core mission preservation (epistemic validity and assessment integrity); and (3) equity of access across student cohorts. Well-resourced universities possess the capital to procure and supervise closed-loop, pedagogically constrained AI tutors that protect data privacy and enforce scaffolding. In contrast, under-resourced institutions and marginalized students are disproportionately pushed toward free-tier commercial chatbots that lack pedagogical guardrails and monetize user telemetry, threatening to turn sound AI pedagogy into a stratified institutional privilege.

III. Review of Recent Empirical and Policy Evidence (2025–2026)

Quantitative Outcomes: Meta-analysis overall effect g = 0.57; small cohorts (N < 50) inflated to g = 1.22; custom tests g = 0.98 vs. standardized tests g = 0.56.

Pedagogical Configurations: OECD (2026) shows constrained scaffolding systems outperform open chatbots; Contractor & Reyes (2026) confirm dialogic inquiry preserves retention while passive copying decays.

Systemic Gaps & Biases: Cong-Lem & Fan (2026) show only 32.3% measure process logs; delayed unassisted retention and Global South contexts remain severely underrepresented.

Fig. 1. Fig. 1. Synthesis of Recent Empirical and Policy Corpus (2025–2026).

A. Quantitative Learning Outcomes and Effect Size Sensitivity

A comprehensive meta-analysis published in *Educational Technology & Society* [7] synthesized 44 effect sizes across 21 independent empirical studies, reporting an aggregated standardized effect of $g = 0.57$, indicating a moderate overall benefit for GenAI utilization. Disaggregated across learning domains, effect sizes varied across behavioral engagement ($g = 0.70$), cognitive outcomes ($g = 0.60$), and affective outcomes ($g = 0.48$).

Crucially, methodological sensitivity analyses revealed pronounced structural moderators. Studies with small sample sizes ($N < 50$) exhibited an inflated mean effect of $g = 1.22$, whereas larger, scaled investigations stabilized between $g = 0.48$ and $g = 0.55$. Similarly, investigations relying on unstandardized, researcher-developed tests yielded an elevated effect of $g = 0.98$, compared to $g = 0.56$ on standardized curricular assessments. These disparities indicate positive

publication bias and novelty reactivity in early small-scale trials.

Experimental findings from Contractor and Reyes [8] corroborate these meta-analytic patterns. In a randomized controlled trial with undergraduate cohorts, students with unrestricted GenAI access significantly outperformed control peers on immediate deliverables. However, on an unassisted retest administered seven days later, the initial performance advantage suffered sharp decay. Retention was strongly moderated by interaction style: learners who used AI dialogically—interrogating logic, requesting counterarguments, and clarifying reasoning—retained conceptual mastery, whereas learners who delegated drafting and problem-solving directly to the model experienced severe unassisted degradation.

B. Pedagogical Profiles and Tool Configurations

The *OECD Digital Education Outlook 2026* [4] formalizes three structural modalities for educational AI deployment: (1) direct learner interaction for individual conceptual inquiry; (2) collaborative dyadic interaction wherein instructors and students co-engage with AI in structured classroom activities; and (3) instructor-facing administration to offload lesson planning, rubric design, and record-keeping. The report emphasizes that intelligent tutoring systems engineered with pedagogical constraints—such as Socratic questioning, progressive hints, and adaptive difficulty—consistently outperform open-ended commercial chatbots. Furthermore, automating routine administrative tasks liberates educator capacity for high-touch mentorship, provided tools are co-designed with faculty.

C. Structural, Epistemic, and Ethical Risks

A systematic synthesis of 53 empirical investigations by Rakthai [9] identifies five recurring systemic hazards in educational AI adoption: (1) *data privacy violations* via uncontrolled transmission of student work to commercial training pools; (2) *algorithmic bias* reinforcing cultural, linguistic, and demographic stereotypes; (3) *epistemic misinformation* through plausible yet fabricated hallucinations; (4) *cognitive atrophy* eroding analytical independence; and (5) *academic integrity breaches* normalizing unacknowledged cognitive outsourcing. Complementing these findings, García-López and Trujillo-Liñán [10] highlight structural asymmetries in algorithmic transparency, while the OECD [4] warns of the *normalization of assistance*, wherein persistent computational support progressively lowers students' internal problem-solving thresholds.

D. Methodological and Contextual Gaps in Current Literature

A scoping review of 31 empirical studies by Cong-Lem and Fan [11] exposes acute imbalances across the current evidence base: 71.0% of studies measured immediate academic achievement and 58.1% assessed dispositional metrics (motivation, attitudes, self-efficacy), but only 32.3% captured dynamic learning processes (interaction telemetry or inquiry logs), and delayed unassisted retention was virtually absent. Demographic coverage is similarly skewed toward post-secondary institutions in industrialized Western economies, leaving K–12, vocational, and special education contexts underrepresented [12], [13]. Emerging educational ecosystems—such as Indian higher education, characterized by high student-to-teacher ratios, socio-economic stratification, and multilingual classrooms—remain critically understudied. Moreover, existing studies frequently confound tool *accessibility* with *interaction pedagogy*, obscuring the true drivers of durable learning.

IV. Objectives and Research Propositions

A. Research Objectives

- To evaluate whether GenAI utilization improves student learning outcomes and whether gains persist after tool removal.
- To establish a structured taxonomy classifying whether GenAI is deployed to augment or substitute human cognition.
- To synthesize the cognitive, epistemic, and socio-technical risks of ungoverned GenAI adoption in higher education.
- To critique the methodological rigor, demographic inclusivity, and measurement limitations of 2025–2026 empirical literature.
- To formulate actionable pedagogical, assessment, and institutional procurement frameworks tailored to diverse higher education environments.

B. Research Propositions

Because this investigation employs a qualitative narrative synthesis across secondary empirical and policy literature, formal hypotheses are articulated as four testable Epistemic Propositions (P_1–P_4):

- **Proposition 1 (P_1):** Generative AI access produces measurable immediate improvements in student task execution, but these gains exhibit rapid decay upon tool removal unless specifically scaffolded.
- **Proposition 2 (P_2):** Augmentative AI engagement (dialogic inquiry, reflective prompting, and verification) yields significantly higher durable learning retention than substitutive engagement (direct text generation and unedited task completion).
- **Proposition 3 (P_3):** Reported effect sizes in educational AI literature are systematically moderated by methodological architecture, with small cohort sizes and researcher-developed tests yielding inflated efficacy estimates.
- **Proposition 4 (P_4):** Ungoverned, market-driven deployment of generative AI tools exacerbates existing socio-economic and institutional inequities rather than democratizing personalized learning.

V. Research Methodology

This study adopts a qualitative narrative literature review methodology tailored for conceptual framework development in emerging educational technologies. No new primary data from human participants were collected; instead, findings were synthesized from peer-reviewed empirical studies, randomized controlled trials (RCTs), meta-analyses, systematic scoping reviews, and multilateral policy frameworks published between early 2025 and early 2026.

Scope & Window: Generative AI in Higher & K–12 Education (2025 – Early 2026); Meta-analyses, RCTs, Scoping Reviews, Policy Reports.

Synthesis Dimension 1: Cognitive Retention vs. Immediate Task Execution.

Synthesis Dimension 2: Interaction Typology & Locus of Cognitive Ownership.

Synthesis Dimension 3: Methodological Sensitivity & Effect Size Inflation.

Synthesis Dimension 4: Socio-Technical Equity & Institutional Infrastructure.

Fig. 2. Fig. 2. Methodological Design and Four-Dimensional Synthesis Pipeline.

- *Inclusion Criteria:* Peer-reviewed empirical investigations of GenAI in student learning; quantitative meta-analyses; systematic scoping reviews of cognitive or dispositional outcomes; and policy reports from recognized international bodies (e.g., OECD).
- *Exclusion Criteria:* Non-peer-reviewed opinion editorials, commercial vendor white papers lacking transparent methodology, and purely technical LLM architecture papers devoid of educational evaluation metrics.

The finalized core analytical corpus comprises eight foundational sources [4], [6]–[13]. Data extraction focused on four comparative dimensions: sample demographics, assessment instrument standardization, retention testing intervals, and student–AI interaction modalities, triangulating quantitative metrics across secondary reviews prior to conceptual synthesis.

VI. Analysis and Synthesis of Evidence

A. Systematic Corpus Mapping

Table I summarizes the core empirical and policy evidence base reviewed, categorized by methodological design, investigatory scope, and primary findings.

TABLE I. Methodological Mapping and Summary of Reviewed Evidence Base

Primary Source	Design	Investigatory Scope	Core Empirical Findings
Ed. Tech. & Society (2025) [7]	Quantitative Meta-Analysis	21 studies; 44 independent effect sizes	Pooled effect $g = 0.57$. Moderated by sample size (small $g = 1.22$ vs. large $g = 0.48$–0.55) and test type (custom $g = 0.98$ vs. standardized $g = 0.56$).
Contractor & Reyes (2026) [8]	Randomized Controlled Trial	Undergraduate university cohorts	Immediate scores elevated under AI access; 7-day unassisted retest showed rapid decay for passive users, while dialogic users retained gains.
Cong-Lem & Fan (2026) [11]	Systematic Scoping Review	31 empirical investigations	71.0% measured immediate achievement; 58.1% dispositional metrics; only 32.3% captured dynamic process logs; longitudinal retention absent.
OECD (2026) [4]	International Policy Report	Global educational systems	Purpose-built, scaffolded intelligent tutoring architectures systematically outperform open-domain chatbots; admin AI frees educator time.
Fournier & Łodzikowski (2026) [6]	Socio-Technical Framework	Emerging agentic AI workflows	Identifies three-tension institutional conflict: technological innovation vs. core educational mission integrity vs. equitable access.
Rakthai (2026) [9]	Systematic Literature Review	53 empirical & theoretical studies	Establishes five core risk vectors: data privacy erosion, algorithmic bias, epistemic hallucination, cognitive deskilling, and dishonesty.
García-López & Trujillo-Liñán (2025) [10]	Systematic Literature Review	Applied educational ethics	Highlights structural asymmetries in algorithmic transparency; privacy erosion and cultural/linguistic bias represent primary systemic threats.
Stebbins / Alfarwan (2025) [12], [13]	Practitioner & Scoping Synthesis	K–12 & early foundational education	Identifies acute empirical underrepresentation in non-higher education settings and severe deficits in K–12 longitudinal evidence.

B. Thematic Synthesis Across Empirical Findings

- *Decoupling Performance from Retention:* Immediate task execution is reliably accelerated by GenAI deployment ($g = 0.57$) [7], [8], but this gain does not equate to durable schema formation. When AI tools are withdrawn during delayed testing, unassisted performance drops sharply, confirming that assisted output frequently masks underlying cognitive deficits.
- *Mediating Role of Cognitive Ownership:* Post-assistance retention depends on learner interaction style [4], [8], [11]. Durable conceptual mastery occurs when AI operates as a Socratic probe—requiring students to clarify assumptions and synthesize conclusions—whereas end-to-end task delegation induces rapid memory decay.
- *Methodological Inflation in Early Literature:* Substantial effect size inflation in small-sample ($g = 1.22$) and unstandardized-test ($g = 0.98$) studies [7] indicates that early claims of transformative efficiency often reflect novelty effects and unvalidated rubrics rather than stable learning.
- *Empirical Paucity Regarding Equity:* While governance frameworks [6], [9] warn that ungoverned adoption will stratify educational quality, cross-institutional empirical trials comparing socio-economically and linguistically diverse cohorts remain scarce.

C. Evaluation of Research Propositions

Synthesizing the reviewed literature yields structured evidentiary verdicts for each proposition formulated in Section IV-B, summarized in Table II.

TABLE II. Analytical Evaluation and Verdict on Research Propositions

Proposition	Analytical Verdict	Empirical and Theoretical Evidentiary Basis
$\mathbf{P_1}$: GenAI enhances immediate performance, but gains decay without scaffolding.	**Supported**	Validated across meta-analytic evidence ($g = 0.57$) [7] and randomized trial data [8] showing significant performance drop-offs on delayed, unassisted post-tests.

Proposition	Analytical Verdict	Empirical and Theoretical Evidentiary Basis
P_2: Augmentative interaction yields greater durable learning than substitutive use.	**Supported**	Confirmed by Contractor and Reyes [8], where dialogic inquiry preserved 7-day retention while generative substitution caused rapid decay; aligned with OECD [4].
P_3: Effect sizes are systematically inflated by small cohorts and custom tests.	**Supported**	Demonstrated in meta-analytic subgroup analysis [7]: $g = 1.22$ (small N) vs. 0.48–0.55 (large N), and $g = 0.98$ (custom tests) vs. 0.56 (standardized measures).
P_4: Unmanaged adoption exacerbates institutional educational inequities.	**Partially Supported**	Theoretically grounded in socio-technical governance frameworks [6], [9], though longitudinal cross-institutional trials testing stratified student outcomes remain limited.

VII. Conceptual Contributions and Framework

To move educational research beyond simplistic binary evaluations of "AI use" versus "non-use," this section formalizes three original conceptual instruments designed to diagnose when, why, and under what cognitive configurations durable learning occurs.

A. The Four-Mode Taxonomy of Student–AI Interaction

We classify learner–AI interactions along a continuous Spectrum of Cognitive Ownership based on which entity executes the germane cognitive processing (Fig. 3 and Table III).

High Student Ownership (High Retention): [1. SCAFFOLD] Socratic dialogue & hints (Student burden) → [2. CHECK] Post-hoc diagnostic validation (Student burden).

Shared / Low Student Ownership (Moderate to Low Retention): [3. DRAFT-ASSIST] Iterative co-construction (Shared burden) → [4. SUBSTITUTE] End-to-end delegation (Model burden).

Fig. 3. Fig. 3. Spectrum of Cognitive Ownership Across Four Interaction Modes.

In practice, most student workflows fall into the *Draft-assist* zone. Here, learning durability is governed strictly by revision depth: if the student treats the AI output as an artifact for substantive critique, restructuring, and verification, schema integration occurs; if editing is limited to superficial grammatical polishing, the interaction collapses functionally into *Substitution*.

TABLE III. Four-Mode Taxonomy of Student–AI Interaction and Cognitive Ownership

Interaction Mode	Operational Description	Locus of Effort	Primary Epistemic Risk	Projected Retention
1. Scaffold	AI presents targeted hints, Socratic queries, and diagnostic counterarguments; learner navigates problem-solving.	Student (AI as facilitator)	Frustration from instructional friction; premature disengagement.	High (Reinforces germane schema construction)
2. Check	Student completes the analytical task independently, then uses AI to detect logical fallacies, omissions, or flaws.	Student (AI as evaluator)	Uncritical acceptance of false-positive or false-negative AI critique.	High (Preserves initial retrieval effort & struggle)
3. Draft-assist	Student and AI iteratively co-construct the artifact; learner evaluates, edits, reorganizes, and expands drafts.	Shared (Distributed processing)	Illusion of authorship; variable depth of critical revision.	Moderate (Contingent on substantive revision depth)
4. Substitute	AI generates the complete response from a prompt; learner submits output without substantive cognitive intervention.	Generative Model (Student as curator)	Cognitive atrophy, illusion of explanatory depth, zero schema integration.	Low / Negligible (Severe decay on unassisted retest)

B. The Retention Ratio (RR)

To provide researchers and educators with a standardized metric separating transient task execution from durable competence, we introduce the Retention Ratio (RR). Let $A \in [0, 100]$ represent a student's standardized score on a task completed with real-time GenAI access, and let $U \in [0, 100]$ represent the same student's score on an epistemically equivalent, isomorphic task administered unassisted after a fixed retention interval (e.g., 7 days):

$$RR = U / A \quad (\text{for } A > 0) \tag{1}$$

- **RR ≈ 1.0 (High Knowledge Consolidation):** Assisted competence reflects internalized schemas; the AI functioned as an authentic scaffold without inducing dependency.
- **RR ≪ 1.0 (Performance–Learning Decoupling):** High assisted scores paired with low unassisted scores indicate computational substitution and an illusion of mastery.
- **RR > 1.0 (Delayed Incubation / Hypermnesia):** Unassisted performance exceeds assisted scores, indicating productive cognitive conflict followed by independent consolidation.

Methodological Constraints: (1) *Denominator Regularization*: If $A = 0$, RR is undefined, reflecting baseline task failure. (2) *Task Equivalence*: Assisted (T_A) and unassisted (T_U) tasks must share identical Bloom's taxonomy levels, element interactivity, and grading rubrics. (3) *Time-Decay Standardization*: Retention intervals must remain fixed (e.g., exactly 7 or 14 days) across cohorts.

C. Empirically Testable Predictions

- **Prediction 1 (Pred$_1$):** Students restricted to Scaffold and Check modes will achieve significantly higher Retention Ratios ($RR \geq 0.85$) than cohorts in unconstrained or Substitute modes ($RR \leq 0.40$), despite slower initial task completion.
- **Prediction 2 (Pred$_2$):** Within Draft-assist workflows, RR will correlate positively with the semantic edit

distance between the initial AI draft and final submission ($r \geq 0.50$).

- **Prediction 3 ($Pred_3$):** Curricula implementing programmatic scaffold fading across multi-week modules will yield significantly higher *RR* values than static, unattenuated AI support.
- **Prediction 4 ($Pred_4$):** Cohorts using unconstrained consumer chatbots will exhibit lower mean *RR* and higher retention variance than cohorts using pedagogically constrained institutional tutors.

D. *Pedagogical Decision Checklist*

Before integrating GenAI into coursework, instructors and curriculum committees should audit learning designs against five diagnostic criteria:

- **Cognitive Locus:** Does the task retain core germane analytical reasoning within the student's working memory?
- **Planned Fading:** Is there a structured schedule to diminish AI scaffolding and require unassisted retrieval before final evaluation?
- **Independent Verification:** Does assessment include delayed, unassisted evaluations to compute true Retention Ratios?
- **Equity & Privacy:** Does the tool protect student telemetry from commercial scraping and function reliably across low-bandwidth devices and regional languages?
- **Pedagogical Transparency:** Does the educator retain administrative visibility into student–AI prompt logs and interaction trajectories?

VIII. Discussion

A. *The Epistemic Illusion: Decoupling Performance from Learning*

Traditional academic assessment relies on student artifacts—essays, proofs, and source code—as valid proxies for internal cognitive mastery. Generative AI invalidates this proxy. Because commercial LLMs generate coherent, structurally sophisticated outputs instantaneously, artifact quality no longer guarantees that learning occurred. Furthermore, Rozenblit and Keil's [14] *illusion of explanatory depth* intensifies: observing fluent AI-generated prose convinces learners they understand concepts they never processed in working memory. This explains why immediate assessments show strong performance gains while delayed unassisted retests reveal sharp skill decay [8].

B. *Three Prospective Institutional Trajectories*

Depending on governance, assessment redesign, and faculty investment, higher education institutions face three divergent adoption trajectories:

- **Trajectory 1 — Positive Integration:** Institutions deploy Socratic, scaffolded AI tutors while automating administrative workloads so faculty can focus on mentorship, oral defenses, and laboratory synthesis. Equity, data privacy, and cohort *RR* improve substantially.
- **Trajectory 2 — Technocratic Stagnation:** Students rely on general chatbots for asynchronous assignments while institutions deploy unreliable statistical AI detectors. Grades inflate, degree credibility erodes, and graduates enter the workforce lacking independent problem-solving capacity.
- **Trajectory 3 — Stratified Inequality:** Well-funded universities provide regulated, privacy-preserving Socratic tutors that cultivate meta-cognitive mastery, while under-resourced public and regional colleges default to unregulated free chatbots, widening societal skill divides.

C. *Contextual Realities of the Global South and Indian Higher Education*

These structural tensions are especially acute across the Indian higher education ecosystem, which operates under three compounding constraints: (1) *Large Class Sizes*, with student-to-teacher ratios frequently reaching 60:1 or 100:1, limiting individualized feedback; (2) *Linguistic Diversity*, where students speaking native languages such as Marathi, Hindi, and Tamil experience heavy extraneous cognitive load when studying and writing technical assessments exclusively in English; and (3) *Uneven Digital Infrastructure*, where high-speed broadband, modern workstations, and paid AI APIs are unevenly distributed between urban and rural campuses.

Appropriately governed GenAI can bridge these barriers by translating complex technical concepts into regional languages and providing on-demand code debugging explanations. However, without institutional infrastructure, students default to free consumer chatbots in *Substitute* mode—producing functional project reports and code submissions without mastering algorithmic logic, compilation, or independent debugging. As identified in Section III-D, empirical investigations within Indian universities remain urgently needed.

IX. Actionable Recommendations

To ensure generative AI strengthens rather than replaces human cognition, stakeholders must execute coordinated reforms across four operational levels:

A. *Curriculum and Classroom Pedagogy*

- **Prioritize Epistemic Friction:** Anchor assignments in Bjork's desirable difficulties [5], rewarding iterative troubleshooting, assumption testing, and conceptual synthesis over rapid deliverable generation.
- **Institutionalize Scaffold Fading:** Structure syllabi in three mandatory phases: Phase 1 permits guided exploration (*Scaffold*); Phase 2 restricts AI to diagnostic critique (*Check*); and Phase 3 requires fully unassisted synthesis.
- **Mandate Interaction Process Logging:** Require transparent prompt-and-revision logs alongside submissions, grading students on dialogic inquiry, hallucination verification, and substantive editing.

B. *Learning Assessments*

- **Adopt Multimodal Verification:** Supplement asynchronous submissions with live oral vivas,

whiteboard problem-solving, proctored lab coding, and peer defense.

- **Discontinue Statistical AI Detectors:** Abandon unreliable automated AI-text detectors, which suffer high false-positive rates and disproportionately penalize non-native English writers.
- **Audit Cohort Retention Ratios:** Embed delayed, unassisted post-tests in core courses to monitor cohort *RR* and detect modules experiencing excessive cognitive offloading.

C. *Institutional Policy and Procurement*

- **Procure Constrained Closed-Loop Tools:** Select educational AI platforms with strict data-privacy guarantees that prohibit model training on student telemetry and enforce Socratic guardrails.
- **Subsidize Equitable Access:** Allocate institutional funding to provide regulated AI APIs and multilingual support across under-resourced departments so marginalized students are not relegated to unregulated free tools.

D. *Faculty Development and AI Literacy*

- **Advance Pedagogical AI Training:** Train instructors beyond basic prompt writing to design scaffolded workflows, interpret interaction telemetry, and calibrate cognitive load.
- **Embed Critical AI Literacy:** Equip both faculty and students to understand probabilistic token prediction, algorithmic bias, hallucination triage, and cognitive atrophy risks.

X. Agenda for Future Research

- **Longitudinal Cohort Trials:** Conduct multi-semester panel studies (6 to 24 months) tracking how sustained AI scaffolding versus substitution affects baseline working memory and self-regulated learning.
- **Empirical Validation of the Taxonomy and RR:** Execute classroom RCTs testing Predictions $Pred_1$–$Pred_4$ across STEM and humanities disciplines.
- **Interaction Telemetry Process Mining:** Apply natural language processing to student prompt logs to quantify how inquiry depth and semantic edit distance mediate retention.
- **Indian Higher Education Field Trials:** Investigate multilingual GenAI scaffolding in high-enrollment Indian technical programs to measure bilingual cognitive load reduction.
- **Human Agency in Multi-Agent Systems:** Examine how autonomous agentic workflows that plan, code, and debug end-to-end reshape human metacognitive oversight.

XI. Methodological Limitations

- **Narrative Synthesis Scope:** As a conceptual narrative review rather than a quantitative PRISMA meta-analysis, source selection followed purposive thematic sampling to synthesize emerging frameworks.
- **Reliance on Rapidly Evolving Pre-Prints:** To capture frontier 2025–2026 developments, the synthesis incorporates working papers and pre-prints [6], [8] undergoing ongoing peer refinement.
- **Cross-Study Effect Heterogeneity:** Variations in participant age, disciplinary domain, and assessment instruments across meta-analytic sources [7] mean reported effect sizes represent directional trends rather than invariant constants.
- **Empirical Calibration Needed:** The Four-Mode Taxonomy and Retention Ratio require multi-institutional psychometric validation to establish standardized disciplinary benchmarks.

XII. Conclusion

Generative AI holds substantial potential to support higher education, yet immediate performance acceleration must not be mistaken for durable learning. When deployed to bypass germane cognitive effort, generative systems prevent the consolidation of stable long-term memory schemas, leaving learners with fragile skills that deteriorate rapidly once computational support is removed. Achieving durable educational gains requires deliberate pedagogical architecture. By operationalizing the Four-Mode Interaction Taxonomy and auditing student mastery through the Retention Ratio (*RR*), educators and institutions can distinguish genuine cognitive scaffolding from passive offloading—ensuring that AI technologies cultivate independent, resilient human intellect rather than superficial fluency.

References

[1] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cogn. Sci.*, vol. 12, no. 2, pp. 257–285, 1988, doi: 10.1207/s15516709cog1202_4.

[2] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA, USA: Harvard Univ. Press, 1978.

[3] D. Wood, J. S. Bruner, and G. Ross, "The role of tutoring in problem solving," *J. Child Psychol. Psychiatry*, vol. 17, no. 2, pp. 89–100, 1976, doi: 10.1111/j.1469-7610.1976.tb00381.x.

[4] OECD, *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*. Paris, France: OECD Publishing, 2026, doi: 10.1787/2026-en.

[5] R. A. Bjork, "Memory and metamemory considerations in the training of human beings," in *Metacognition: Knowing About Knowing*, J. Metcalfe and A. Shimamura, Eds. Cambridge, MA, USA: MIT Press, 1994, pp. 185–205.

[6] J. Fournier and K. Łodzikowski, "Addressing the reality gap: A three-tension framework for agentic AI adoption," *arXiv preprint arXiv:2601.XXXXX*, 2026.

[7] *Educational Technology & Society*, "Evaluating the effects of generative AI on student learning outcomes: Insights from a meta-analysis," *Educ. Technol. Soc.*, vol. 28, no. 3, pp. 226–240, 2025.

[8] Z. Contractor and G. Reyes, "Experimental evidence on the learning impact of generative AI," Working Paper, Dept. of Economics, Middlebury College, 2026.

[9] P. Rakthai, "Issues in artificial intelligence in education: Ethical, equity, pedagogical, and policy challenges in the generative AI era," *J. Educ. Technol. Syst.*, vol. 54, no. 3, pp. 312–335, 2026.

[10] I. M. García-López and L. Trujillo-Liñán, "Systematic review of ethical risks of generative AI in education," *Contemp. Educ. Technol.*, vol. 16, no. 2, Art. no. ep501, 2025.

[11] N. Cong-Lem and H. Fan, "Exploring the effects of generative AI (GenAI) on learning outcomes: A scoping review," *Res. Pract. Technol. Enhanc. Learn.*, vol. 22, Art. no. 022, 2026, doi: 10.1186/s41032-026-00022-x.

[12] L. Stebbins, "Generative AI in K–12 education: What we know, what's missing, and what comes next," *Ed-Spaces*, Dec. 18, 2025.

[13] A. Alfarwan, "Review of generative AI use in K–12 education," summarized in L. Stebbins, *Ed-Spaces*, 2025.

[14] L. Rozenblit and F. Keil, "The misunderstood limits of folk science: An illusion of explanatory depth," *Cogn. Sci.*, vol. 26, no. 5, pp. 521–562, 2002, doi: 10.1207/s15516709cog2605_1.