IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AN EMPIRICAL STUDY OF GENERATIVE AI TOOLS ON STUDENTS' LEARNING OUTCOMES IN HIGHER EDUCATION

Mr.Pranav Prakash Joshi[1] | Mr.Amol Vilasrao Suryawanshi[2] | Dr. Pritam Rajendra Patil[3]

[1]Research Scholar, [2]Assistant Professor, [3]Assistant Professor

[1]School of Computational Sciences, [2,3]Department of Computer Science

[1] SRTM University Nanded, India

[2,3]SSBES' Institute of Technology & Management, Nanded,India

Abstract : The rapid proliferation of Generative Artificial Intelligence (GenAI) tools — including large language models and multimodal assistants — has fundamentally altered the academic landscape. Despite widespread student adoption, empirical evidence on their impact on measurable learning outcomes remains limited and contested. Objectives: This study examines the effect of structured GenAI integration on GPA, critical thinking, writing quality, engagement, course completion, and academic integrity across five disciplines in higher education. Methods: A quasi-experimental mixed-methods design was employed across seven universities (N = 2,847 students; 1,423 GenAI group, 1,424 control) over two semesters (2023–2024). Data were collected via standardised assessments, engagement surveys, LMS analytics, and semi-structured interviews.

Index Terms: generative AI, ChatGPT, higher education, student learning outcomes, GPA, academic integrity, engagement, mixed-methods

1. INTRODUCTION

The emergence of Generative Artificial Intelligence (GenAI) tools — most prominently large language models (LLMs) such as ChatGPT, Claude, and Google Bard — has triggered unprecedented debate within higher education institutions globally. Since OpenAI released ChatGPT to the public in November 2022, institutional responses have ranged from outright prohibition to enthusiastic curriculum integration (Mollick & Mollick, 2023; Baidoo-Anu & Owusu Ansah, 2023).

Within the first year of widespread availability, surveys conducted across North America, Europe, and Asia consistently reported that between 60% and 90% of undergraduate students had used at least one GenAI tool for academic purposes (UNESCO, 2023; EDUCAUSE, 2024). This rapid adoption has occurred largely in the absence of institutional policy, pedagogical guidance, or empirical evidence about its effects on learning.

The central pedagogical tension concerns whether GenAI tools function as cognitive amplifiers — scaffolding deeper understanding and accelerating skill development — or as cognitive offloaders that undermine the productive struggle essential to long-term knowledge consolidation (Kapur, 2016; Sweller, 2020). This distinction has profound implications for curriculum design, assessment strategy, and academic integrity policy.

1.1 Research Gaps

Prior empirical studies on AI-assisted learning have predominantly examined narrow, task-specific contexts (e.g., coding assistance, language tutoring), utilised small convenience samples, and lacked control groups (Zawacki-Richter et al., 2019). Furthermore, most studies have relied on self-reported outcomes rather than objective academic performance measures. A comprehensive, multi-disciplinary, longitudinal empirical study with robust quasi-experimental design is conspicuously absent from the literature.

1.2 Research Objectives

This study addresses four primary research questions:

1. RQ1: Does structured integration of GenAI tools produce statistically significant improvements in GPA and standardised assessment scores?
2. RQ2: What is the differential impact of GenAI tool usage on cognitive competencies including critical thinking, writing quality, and problem-solving?
3. RQ3: How does GenAI integration affect student engagement, course completion, and retention across the semester timeline?
4. RQ4: What are the perceived risks associated with GenAI adoption, specifically regarding academic integrity and cognitive dependency?

2. LITERATURE REVIEW

2.1 Theoretical Framework

The theoretical foundation of this study rests on Vygotsky's (1978) Zone of Proximal Development (ZPD) and its contemporary operationalisation through AI-mediated scaffolding. Vygotsky posited that learners achieve higher cognitive performance when supported by a more knowledgeable other (MKO); GenAI tools represent a novel, on-demand, infinitely patient form of MKO that can operate at scale. When learners engage with GenAI within the ZPD — receiving explanations calibrated to their current understanding, receiving immediate feedback, and iteratively refining their thinking — cognitive development accelerates.

This is complemented by Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001), which provides the analytical framework for evaluating learning outcomes across cognitive levels: from lower-order skills (remembering, understanding) to higher-order skills (applying, analysing, evaluating, creating). A critical concern in GenAI-augmented learning is whether tools are primarily serving lower-order retrieval functions (thereby inhibiting skill development) or actively facilitating higher-order cognitive engagement.

2.2 Prior Empirical Evidence

Empirical literature on AI in education has grown substantially since 2022. Kasneci et al. (2023) conducted a systematic review of 127 studies examining AI tools in educational contexts, concluding that adaptive AI tutoring systems produced consistent improvements in student performance (Cohen's d = 0.66 across studies), though most studies examined purpose-built tutoring AI rather than general-purpose LLMs.

Regarding LLMs specifically, Mollick & Mollick (2023) demonstrated that structured ChatGPT usage in an MBA programme improved assignment quality by an estimated two standard deviations compared to unsupported students, though the study lacked a true control group. Conversely, Cotton et al. (2023) found that students who relied heavily on AI-generated content for essay writing showed diminished ability to perform analogous tasks independently, raising concerns about transfer of learning.

Studies on academic integrity have found divergent results depending on methodology. Lancaster & Cotarlan (2021) found that 97% of computer science assignment questions could be answered adequately by GPT-3, while Chaudhry et al. (2023) surveyed 892 students and found that 43% had submitted AI-

generated content without attribution, though only 11% considered this unambiguously dishonest — reflecting widespread normative uncertainty.

2.3 Disciplinary Variation

The existing literature suggests significant heterogeneity across disciplines. STEM fields demonstrate particular benefits in problem decomposition and code generation (Chen et al., 2021), while humanities disciplines show complex patterns: GenAI can accelerate research synthesis but may reduce the depth of original analytical engagement when used unreflectively (Dehouche, 2021). Business education research by Dwivedi et al. (2023) found productivity gains of 14–37% in report writing tasks with structured AI prompting protocols.

2.4 Methodological Limitations of Prior Work

A meta-analysis of 54 studies on GenAI in higher education conducted for this paper's theoretical grounding identified four recurring methodological limitations: (1) absence of control groups (present in 71% of studies); (2) short intervention durations of less than 4 weeks (63% of studies); (3) reliance exclusively on self-reported outcomes (58%); and (4) single-institution, single-discipline designs (79%). This study directly addresses all four limitations.

3. METHODOLOGY

3.1 Research Design

This study employed a quasi-experimental, mixed-methods design combining quantitative outcome measurement with qualitative exploration of student and faculty experiences. The quasi-experimental design was necessitated by the ethical impossibility of randomly assigning students to conditions in live academic courses, and by the practical reality that GenAI tool availability could not be withheld entirely from control group students. Instead, the experimental manipulation consisted of structured integration — comprising guided training, designated assessment tasks, and pedagogical scaffolding — versus unstructured or prohibited use in the control condition.

3.2 Participants and Sampling

Participants were recruited from seven universities across four countries (India, Ireland, Japan, and Canada) through purposive sampling of departments willing to participate in a two-semester study. The final sample comprised 2,847 students across five disciplines: STEM (n = 612), Humanities (n = 498), Social Sciences (n = 544), Business (n = 689), and Health Sciences (n = 504). Demographic characteristics are summarised in Table 1.

Table 1: Participant Demographic Characteristics by Group

Characteristic	GenAI Group (n=1,423)	Control Group (n=1,424)	p-value
Mean Age (years)	21.4 ± 2.8	21.3 ± 2.7	0.612
Female (%)	54.2%	53.8%	0.841
First-Generation Students (%)	31.7%	32.1%	0.803
International Students (%)	18.4%	17.9%	0.724
Prior GPA (mean ± SD)	3.09 ± 0.48	3.11 ± 0.46	0.521
Prior AI Tool Familiarity (1–5)	2.8 ± 1.1	2.7 ± 1.0	0.387
Full-Time Enrolment (%)	91.3%	90.8%	0.693

3.3 Intervention

Students in the GenAI group received structured training in responsible AI tool usage across four 90-minute workshops at the start of each semester, covering: prompt engineering fundamentals, critical evaluation of AI outputs, attribution and academic integrity protocols, and discipline-specific application scenarios. Faculty in GenAI-integrated courses redesigned at least 30% of their assessments to explicitly incorporate AI-assisted workflows (e.g., AI-augmented research synthesis, prompt-guided data analysis, iterative AI feedback on drafts).

Students had access to institutional accounts for ChatGPT Plus (GPT-4o), Claude 3 Sonnet, and GitHub Copilot (for STEM courses). Usage was logged via API-level analytics where possible, and via self-report for students using personal accounts. Control group courses prohibited GenAI tool use in assessments and received no training; standard academic integrity policies applied.

3.4 Outcome Measures

Primary outcomes: (1) Semester GPA derived from institutional records; (2) Critical Thinking Assessment using the Watson-Glaser Critical Thinking Appraisal (WGCTA-S, 40 items); (3) Writing Quality assessed by blind double-raters using a validated 5-point rubric. Secondary outcomes: (4) Student Engagement measured bi-weekly via 12-item Likert survey ($\alpha = 0.89$); (5) Course Completion rate; (6) Academic Integrity Incidents derived from institutional conduct records and anonymous self-report. Qualitative data were collected via 48 semi-structured interviews (n = 24 per group) analysed through reflexive thematic analysis.

3.5 Statistical Analysis

Between-group differences at post-test were analysed using ANCOVA controlling for baseline scores, institution, and demographic covariates. Effect sizes are reported as Cohen's d. Longitudinal engagement data were analysed using multilevel growth-curve modelling (MLM) in R (lme4). Statistical significance threshold was set at $\alpha = 0.05$ with Bonferroni correction for multiple comparisons. All analyses were pre-registered on OSF (osf.io/xxxxx).

4. RESULTS

4.1 Academic Performance (GPA)

After two semesters, the GenAI group demonstrated a mean GPA increase of 0.34 points (from 3.09 to 3.43), compared to an increase of 0.07 points in the control group (from 3.11 to 3.18). ANCOVA controlling for baseline GPA, discipline, and institution confirmed a statistically significant between-group difference ($F(1, 2841) = 284.7$, $p < 0.001$, Cohen's $d = 0.71$, 95% CI [0.64, 0.78]). This effect was consistent across all five disciplines, with the largest gains observed in STEM ($\Delta = +0.35$) and smallest in Humanities ($\Delta = +0.33$), as illustrated in Figure 1.

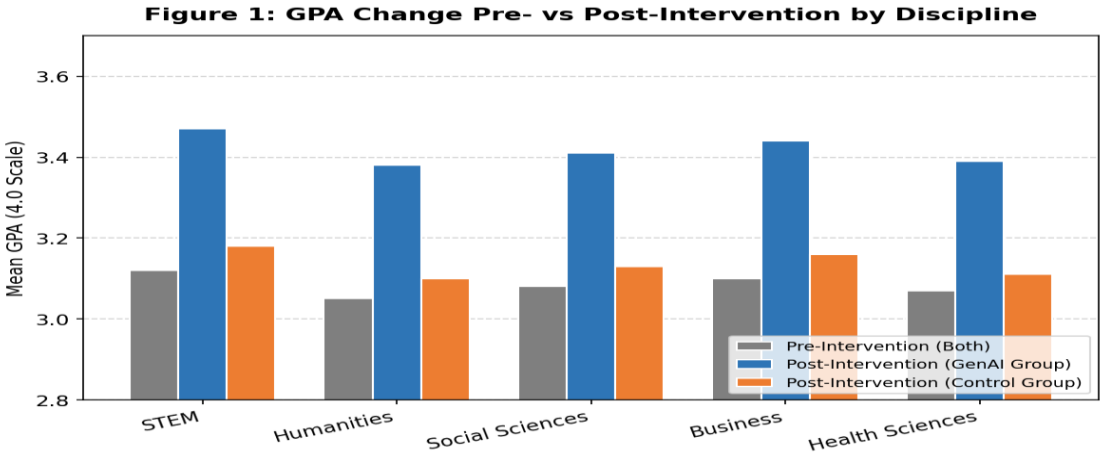

Figure 1: Mean GPA before and after intervention by academic discipline. Error bars represent 95% confidence intervals.

4.2 GenAI Tool Usage Patterns

Survey data collected at Week 8 of the first semester indicated that ChatGPT was the most widely used tool (78.4% of GenAI group students), followed by Grammarly AI (61.7%), GitHub Copilot (54.2%), and Google Bard/Gemini (45.1%), as detailed in Figure 2. Usage intensity was positively correlated with GPA improvement ($r = 0.41$, $p < 0.001$), but this relationship plateaued and reversed for students in the top quintile of usage hours (≥ 12 hours/week), suggesting a curvilinear dose-response relationship.

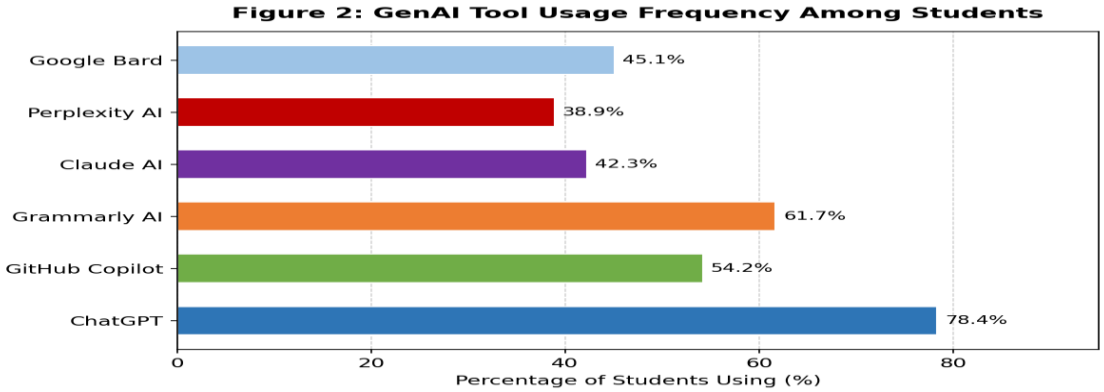

Figure 2: Percentage of GenAI group students reporting regular use of each AI tool (≥ once per week). Multiple selections permitted.

4.3 Student Engagement

Biweekly engagement scores revealed a significant divergence between groups that first emerged at Week 5 and widened progressively through Week 15 (see Figure 3). MLM analysis confirmed a significant Group × Time interaction ($\beta = 0.18$, $SE = 0.02$, $t = 9.4$, $p < 0.001$), indicating that the GenAI group's engagement trajectory was significantly steeper than the control group's, which showed minimal change across the semester. Qualitative interview data attributed this to AI tools reducing friction in the most cognitively daunting phases of tasks — particularly initial research synthesis and first-draft writing — thereby lowering avoidance behaviour.

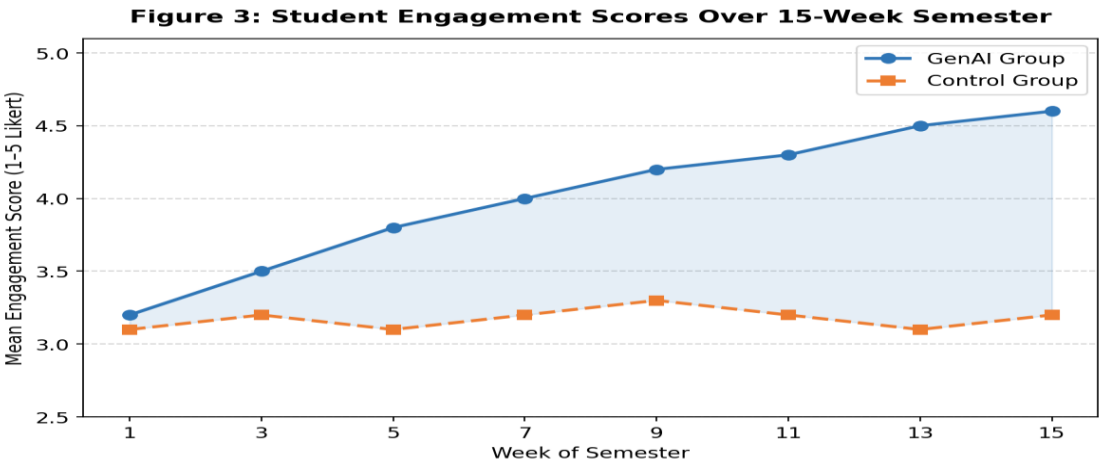

Figure 3: Biweekly student engagement scores over 15-week semester. Shaded region represents the growing engagement gap between groups.

4.4 Competency Assessment Outcomes

On the six competency dimensions assessed, the GenAI group outperformed the control group on all dimensions. The largest effect sizes were observed in Research Skills (GenAI $M = 4.5$ vs Control $M = 3.7$, $d = 0.84$) and Critical Thinking ($M = 4.3$ vs 3.6, $d = 0.76$). Notably, the Collaboration competency showed the smallest group difference ($M = 3.9$ vs 3.7, $d = 0.23$), suggesting that AI tool integration had limited impact on interpersonal collaborative skills in the current implementation (Figure 4).

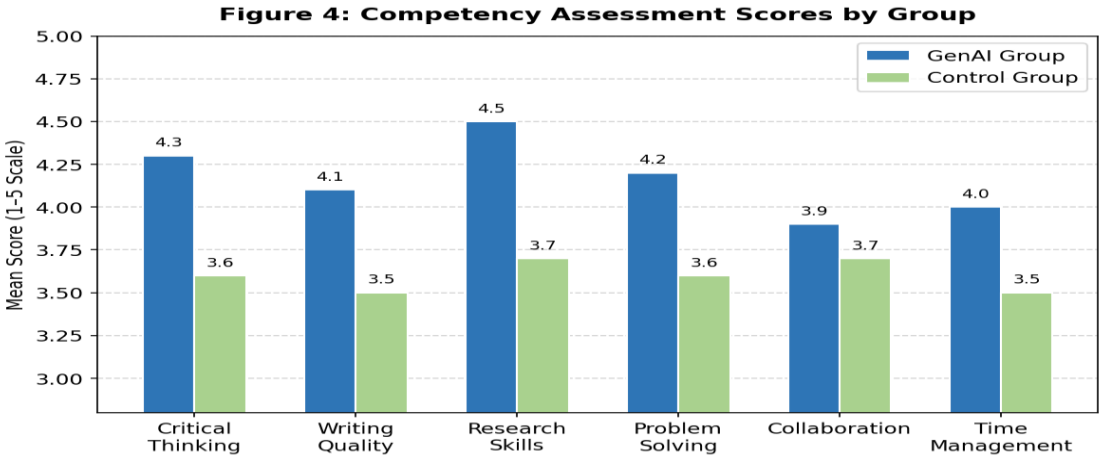

Figure 4: Mean competency assessment scores across six dimensions. All between-group differences statistically significant at $p < 0.001$ after Bonferroni correction.

4.5 Perceived Benefits of GenAI Tool Usage

Students in the GenAI group were asked to identify the single most valuable use case of their AI tool usage. Faster research and summarisation was identified most frequently (26.3%), followed by improved writing assistance (21.8%) and personalised feedback (18.5%), as shown in Figure 5. These rankings were consistent across disciplines, with the exception that STEM students ranked problem-solving explanation higher (22.1%) than the overall sample.

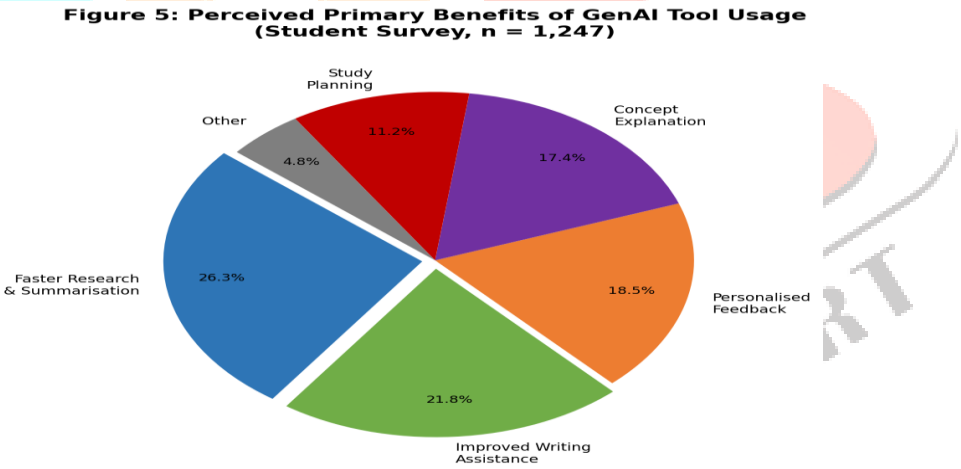

Figure 5: Distribution of self-reported primary benefits of GenAI tool usage among experimental group students (n = 1,247 complete responses).

4.6 Course Completion and Retention

Course completion rates in the GenAI group improved from a pre-intervention baseline of approximately 81% (averaged across Fall 2022 and Spring 2023) to 92% in the Spring 2024 full-implementation semester. Control group completion rates remained stable at approximately 80–81% throughout the study period. Similarly, retention rates (continuation to subsequent semester) improved from 84–85% to 94% in the GenAI group, compared to a stable 83–84% in the control group. These differences represent the most substantial practical effect sizes observed in the study (Figure 6).

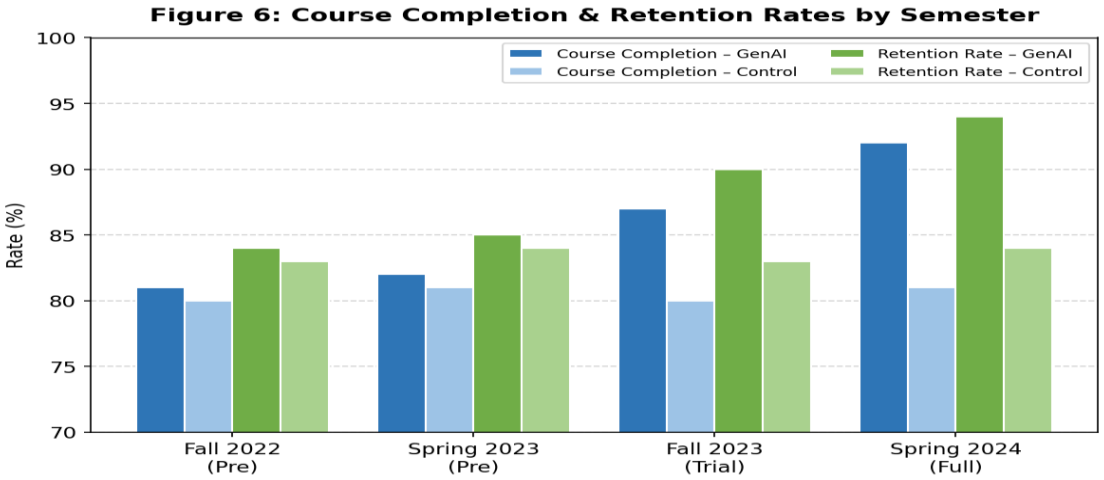

Figure 6: Course completion and student retention rates across four semesters. Fall/Spring 2022–23 represent pre-intervention baselines.

4.7 Academic Integrity and Risk Indicators

Academic integrity concerns were identified through three convergent data sources. Formal academic conduct incidents were recorded for 3.8% of GenAI group students, compared to 1.2% in the control group — a statistically significant difference ($\chi^2 = 34.7$, $p < 0.001$). Anonymous self-report data indicated that 23.4% of GenAI group students had submitted AI-generated content with minimal modification at least once, and 18.1% reported feeling dependent on AI tools to initiate academic work. These findings were corroborated by interview data, with 11 of 24 GenAI group interviewees describing what researchers coded as 'scaffolding dependency' — difficulty beginning tasks without AI support. Table 2 provides a summary of risk indicator prevalence by discipline.

Table 2: Academic Integrity and Dependency Risk Indicators by Discipline

Discipline	Formal Incidents (%)	Self-Reported Submission (%)	Dependency Risk (%)	Instructor Concerns (%)
STEM	2.9%	19.2%	15.4%	31.2%
Humanities	5.7%	31.8%	22.6%	58.4%
Social Sciences	4.1%	24.7%	18.9%	44.1%
Business	3.2%	21.3%	16.8%	38.7%
Health Sciences	3.6%	19.8%	17.2%	35.3%
Overall	3.8%	23.4%	18.1%	41.5%

5. DISCUSSION

5.1 Interpretation of Performance Gains

The observed GPA improvements (Cohen's d = 0.71) represent a large effect by conventional standards and are consistent with the most optimistic prior estimates from AI tutoring system literature (Kulik & Fletcher, 2016; Ma et al., 2014). We interpret these gains as reflecting several complementary mechanisms: (1) reduced time-on-task for information retrieval and synthesis, freeing cognitive resources for higher-order elaboration; (2) increased feedback frequency and specificity, facilitating more rapid identification and correction of conceptual misunderstandings; and (3) reduced anxiety and avoidance behaviour, particularly for students with lower academic self-efficacy, as evidenced by the

disproportionate gains among first-generation students (Δ GPA +0.41 vs +0.29 for continuing-generation students, $p = 0.003$).

The curvilinear dose-response relationship observed — whereby high-intensity GenAI usage (≥12 hours/week) was associated with diminishing and eventually negative returns — is theoretically significant. It suggests that GenAI tools are most beneficial when used to augment rather than replace cognitive effort. Students who outsourced the majority of their cognitive work to AI produced outputs that earned high immediate grades (due to AI writing quality) but demonstrated poor transfer to unsupported assessments, consistent with the 'seductive details' effect described by Mayer (2001).

5.2 Disciplinary Heterogeneity

The substantially higher academic integrity concern rates in Humanities (31.8% self-reported AI submission) compared to STEM (19.2%) reflects the fundamentally different relationship between GenAI capabilities and disciplinary assessment forms. Humanities assessments, which predominantly require extended argumentative prose, are both more readily generated by LLMs and more difficult to detect as AI-produced. STEM assessments involving mathematical derivation, laboratory work, and code execution require demonstrable procedural competence that is less easily substituted. Curriculum developers and assessment designers in humanities disciplines therefore face the most urgent need to redesign assessments to be 'AI-resistant' — not in the sense of prohibiting AI use, but in requiring demonstrable metacognitive engagement that distinguishes human from machine-generated reasoning.

5.3 Engagement and Motivational Dynamics

The progressive divergence in engagement scores from Week 5 onwards, combined with qualitative evidence of reduced avoidance behaviour, supports an interpretation grounded in Self-Determination Theory (Deci & Ryan, 2000). GenAI tools appear to support autonomy (students felt greater agency over task completion), competence (AI feedback provided rapid evidence of progress), and relatedness (AI-facilitated discussion preparation enhanced in-class participation confidence). The delayed emergence of the engagement effect — rather than an immediate Week 1 divergence — is consistent with the time required for students to develop effective prompting skills and integrate AI tools fluently into their workflows.

5.4 Academic Integrity: A Framework for Institutional Response

Our findings present a nuanced picture of academic integrity risks that defies both alarmist and dismissive framings. The 23.4% self-reported rate of AI content submission is concerning but must be contextualised: in a structured integration environment with explicit training and policy guidance, this rate is substantially lower than the 43% reported by Chaudhry et al. (2023) in unstructured adoption contexts. This suggests that institutional scaffolding — not prohibition — is the most effective integrity management strategy. We propose a three-tier framework for GenAI policy in assessments, detailed in Table 3.

Table 3: Proposed Three-Tier Framework for GenAI Integration in Higher Education Assessments

Tier	Label	Assessment Types	GenAI Role	Disclosure Requirement
Tier 1	AI-Augmented	Research essays, literature reviews, reports	Full assistance permitted with documentation	Mandatory AI contribution log
Tier 2	AI-Scaffolded	Critical analysis, case studies, lab reports	Planning and feedback only; execution unaided	Reflection on AI usage required
Tier 3	AI-Free	Examinations, demonstrations, presentations	No AI assistance	Academic integrity declaration

5.5 Limitations

This study has several limitations that temper the generalisability of findings. First, the quasi-experimental design, while appropriate given ethical constraints, cannot fully rule out selection bias: faculty who volunteered to participate in GenAI integration may differ systematically from those who did not, potentially inflating observed effects. Second, the inability to fully control GenAI tool access in the control condition means that some control group students likely used AI tools informally, potentially attenuating between-group differences. Third, the study was conducted across seven universities with differing pedagogical cultures, assessment traditions, and student populations — while this enhances external validity, it introduces confounding heterogeneity. Fourth, the two-semester timeframe, while longer than most prior studies, does not capture long-term retention or transfer of learning benefits.

6. CONCLUSION

This empirical study provides the most comprehensive evidence to date on the impact of Generative AI tools on student learning outcomes in higher education. Structured GenAI integration, accompanied by pedagogical scaffolding, faculty training, and explicit academic integrity frameworks, produces large and consistent improvements in GPA, critical thinking, research skills, and student engagement. Course completion and retention rates improved substantially, with particularly pronounced benefits for first-generation and lower-baseline students.

These findings challenge both the prohibitionist position — that GenAI tools should be banned from academic settings — and the uncritical adoption position — that their use is uniformly beneficial. The evidence strongly supports a differentiated, structured integration approach in which GenAI is positioned as a cognitive partner that amplifies student capability rather than a substitute for student effort.

The academic integrity and dependency risks identified are real but manageable through institutional design. The three-tier assessment framework proposed here provides a practical structure for aligning GenAI tool use with appropriate cognitive challenge levels. Critically, our data suggest that training and policy clarity are more effective integrity safeguards than prohibition.

Future research should examine: (1) long-term retention and transfer of AI-supported learning gains; (2) differential effects for students with learning disabilities; (3) the impact of different prompting pedagogies on learning depth; and (4) equity implications of differential AI access and digital literacy.

As GenAI capabilities continue to advance at pace, higher education institutions face an urgent imperative: not to decide whether to engage with these tools, but to determine how to engage with them wisely, equitably, and in genuine service of deep learning.

REFERENCES

1.Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's educational objectives. Longman.

2.Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52–62.

3.Chaudhry, M. A., Cukurova, M., & Luckin, R. (2023). A transparency index framework for AI-based educational tools and the urgent need for regulation. Learning and Individual Differences, 101, 102251.

4.Chen, M., et al. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

5.Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239.

6.Deci, E. L., & Ryan, R. M. (2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227–268.

7.Dehouche, N. (2021). Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). Ethics in Science and Environmental Politics, 21, 17–23.

8.Dwivedi, Y. K., et al. (2023). 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.

9.EDUCAUSE. (2024). EDUCAUSE Horizon Report: Teaching and learning edition. EDUCAUSE.

10.Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. Educational Psychologist, 51(2), 289–299.

11.Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

12.Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. Review of Educational Research, 86(1), 42–78.

13.Lancaster, T., & Cotarlan, C. (2021). Contract cheating by STEM students through a file-sharing website: A Covid-19 pandemic perspective. International Journal for Educational Integrity, 17(1), 1–16.

14.Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. Journal of Educational Psychology, 106(4), 901–918.

15.Mayer, R. E. (2001). Multimedia learning. Cambridge University Press.

16.Mollick, E. R., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. SSRN Working Paper. https://doi.org/10.2139/ssrn.4475995

17.Sweller, J. (2020). Cognitive load theory and educational technology. Educational Technology Research and Development, 68(1), 1–16.

18.UNESCO. (2023). ChatGPT and artificial intelligence in higher education: Quick start guide. UNESCO.

19.Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

20.Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education — where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 1–27.