Inspira-Journal of Commerce, Economics & Computer Science (JCECS)
(A National bi-lingual Quarterly Double Blind Open Access Peer Reviewed Refereed Journal)
ISSN : 2395-7069(Print), Impact Factor: (Cosmos 7.650 & I2OR 5.120), Volume 12, No. 04, October-December, 2026, pp. 156-165

Evaluating the Impact of Generative AI Tools on Analytical Skill Acquisition and Problem-Solving Among Business School Students

Krishna Murari Chaturvedi[1] & Dr. Bharti Chaturvedi[2*]

[1]Assistant Professor, Department of Commerce and Management, Agrawal P.G. College, Jaipur, Rajasthan.
[2]Assistant Professor, Department of Commerce, Pandit Naval Kishore Government P.G. College, Dausa, Rajasthan.

[*]Corresponding Author: bhartichaturvedi01@gmail.com

Citation: *Chaturvedi, K. & Chaturvedi, B. (2026). Evaluating the Impact of Generative AI Tools on Analytical Skill Acquisition and Problem-Solving Among Business School Students. Journal of Commerce, Economics & Computer Science, 12(04), 156–165.*

Abstract

The fast incorporation of GenAI applications (including large language models like GPT-4, Claude, and specific financial/analytical copilots) has irrevocably transformed the pedagogic framework within which graduate business education is being conducted. Whereas previous computational tools were used to perform routine calculations and code execution, GenAI conducts more advanced cognitive tasks involving semantic synthesis, problem structuring, and conceptual abstraction. Based on Cognitive Load Theory (Sweller, 1988) and the Extended Mind Thesis (Clark & Chalmers, 1998), the present study empirically explores the two-sided influence that constant availability of GenAI exerts on the development of analytical abilities and diagnostic reasoning as well as metacognitive calibration in business school students. Employing a randomized controlled design for a semester-long period ($N = 340$ full-time MBA students) divided into the AI-Enabled Cohort ($n = 170$) and the AI-Restricted Cohort ($n = 170$). The performance was assessed on the basis of the weekly quantitative case deliverables as well as an unassisted proctored comprehensive final examination that assessed Procedural Execution Index (PEI), Diagnostic Reasoning Score (DRS), and Metacognitive Confidence Delta (ΔMC). The empirical results support the existence of a phenomenon that we call Synthesized Fluency Without Structural Anchoring. Though the AI-Enabled group performed much better on the coursework, scoring 88.4 versus 76.2, $p<0.001$, completing the assignments 44.8% faster in 3.2 hours versus 5.8 hours, $p<0.001$, yet their unassisted Diagnostic Reasoning Score was very low at 59.4 versus 78.6, $p<0.001$, and their metacognitive overconfidence was very high, $\Delta MC = +18.7\text{ points}$ versus $+3.1\text{ points}$, $p<0.001$.

Keywords: *Generative Artificial Intelligence, Business Education, Cognitive Load Theory, Cognitive Offloading, Diagnostic Reasoning, Metacognitive Calibration, Analytical Skill Acquisition, Pedagogical Design.*

Introduction

Background and Context of Business Education

Education in management and business has historically involved inquiry-based learning processes, most prominently through case studies, mathematical modeling, and diagnostic methods to build executives' decision-making abilities in ambiguous environments. Whereas scientific disciplines operate based on strict determinacy laws, the practice of business administration involves dealing with

Copyright © 2026 by Author's and Licensed by Inspira. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work properly cited.

unstructured problem areas, working with incomplete data sets, discovering hidden limitations, and generating integrated strategic solutions. Traditionally, such skills were attained through considerable mental labor from students who had to manually generate DCF schedules, define constraints for linear programming, conduct econometric diagnostics, and repeatedly check their assumptions.

But the appearance and widespread use of Generative Artificial Intelligence (GenAI), which spans from 2023 to 2026, has shaken up these pedagogical standards. LLMs like OpenAI's GPT-4/o series, Anthropic's Claude, Google's Gemini, and specialized financial copilot models have moved beyond being mere experimental novelties to becoming standard academic aids. The current crop of business students frequently uses GenAI for writing code in Python and R languages for regression analysis, creating sophisticated financial formulas in Excel, synthesizing several pages of Harvard and Ivey business cases, doing competitive comparisons, and writing detailed memos.

The Qualitative Departure from Prior Cognitive Technologies

Advocates for open-ended use of LLMs rely on historical comparisons to prior technological disruptions in quantitative education – including pocket electronic calculators in the 1970s, spreadsheet applications (Lotus 1-2-3, Microsoft Excel) in the 1980s, and statistical programs (SPSS, Stata, R) in the 1990s and 2000s. The core idea is that offloading basic, mechanistic calculations enables students to think more strategically.

This comparison, however, misses a crucial point. Earlier technologies offloaded mechanical processes – a calculator performs the calculation of an internal rate of return only after the user figures out the appropriate cash flow sequence to use; statistical software performs the variance inflation factor (VIF) test only after the analyst understands the risk of multicollinearity. GenAI offloads semantic synthesis and abstraction. If the user poses an unstructured business challenge to the language model, the latter will formulate the problem, identify the methodological approach, develop the code or the financial model, interpret the statistical results, and write the recommendation for management. Thus, the technology not only helps the execution of the student's reasoning process but replaces it as well.

Statement of the Problem and Research Gaps

In the context of India's higher education system, and particularly postgraduate commerce and management courses under UGC guidelines and NEP 2020 regulations, outcome-based education (OBE) values critical thinking, analytical thinking, and problem-solving skills. Although there has been increasing emphasis on academic integrity and anti-plagiarism checks with respect to AI-generated writing within institutional policies, there is an alarming lack of research studies exploring the cognitive and development aspects of using AI regularly in the learning process.

Indeed, a major pedagogical conundrum has developed whereby students who employ GenAI software submit assignments demonstrating extremely high levels of polish, syntactical fluency, and analytical insight; consequently, teachers end up overestimating students' level of competence. However, reports by university professors and corporate recruiters demonstrate that there is an increasing disparity between the polished nature of student work done with the help of GenAI and their ability to conduct an analysis without any assistance in face-to-face or closed book exams. Current literature has mostly considered productivity among workers in corporations (Dell'Acqua et al., 2023) or more conceptual essays about AI in higher education (Mollick & Mollick, 2023).

Research Questions and Objectives

To help fill this void, the current research explores the long-term impact of integrated GenAI use on MBA students during an entire semester. The research is motivated by the following two research questions:

RQ1 (Skill Development and Efficiency): Is the use of GenAI-based tools more likely to speed up or hinder the skill development of problem solving and financial modeling among business students?

RQ2 (Transfer of Learning and Anomaly Detection): In what way does the process of generation and synthesis influence cognitive transfer, particularly the identification of hidden financial, operations, and structural anomalies without technology?

Consequently, the aims of this study are:

- To assess the performance disparity between AI-enabled and AI-disabled students in their respective assisted and unassisted assessments.
- To determine the level of metacognitive miscalibration (gap between perception of skill and actual unassisted performance) that may be generated due to fluent interaction with artificial intelligence.
- To design appropriate changes in the curriculum and assessments in line with UGC guidelines and retain the efficiency of GenAI.

Literature Review and Theoretical Framework

- **Cognitive Load Theory and Schema Acquisition**

The theoretical framework underpinning this study is derived mainly from Cognitive Load Theory (CLT) developed by John Sweller (Sweller, 1988; Sweller, van Merriënboer, & Paas, 1998). The theory of cognitive load holds that human cognitive structure is restricted by an extremely low working memory capacity in the case of new material, and an effectively infinite long-term memory system comprising cognitive schemas. Cognitive schemas refer to a conceptual framework that organizes pieces of information according to how they are going to be applied to solve a problem or task.

According to Sweller et al. (1998), there are three components of cognitive load in the process of learning that can add up:

- Intrinsic Cognitive Load: Defined by the intrinsic complexity and the element interactivity of the topic of instruction (e.g., comprehending the interaction between changes in working capital, depreciation, and terminal value in DCF valuation).
- Extraneous Cognitive Load: Inflated due to sub-optimal instructional design, formatting friction, or irrelevant tasks that burden working memory but do not contribute to learning.
- Germane Cognitive Load: The cognitive work of building schemas in long-term memory.

The use of appropriate scaffolding helps to minimize extraneous load in order for the individual to be able to devote all working memory capacity towards germane load. But there is one important condition to watch out for, where the technology becomes too good at doing everything and thus takes away the intrinsic and germane processes on its own. We refer to this phenomenon as epistemic offloading – where the LLM does everything such as structuring the problem, selecting the right variables, resolving any edge case issues and in doing so provides for very little cognitive friction for the individual.

- **The Extended Mind Thesis and Cognitive Offloading**

Extended Mind Thesis (EMT) developed by Clark and Chalmers (1998) states that cognitive processes are not confined by the physiological limits of the human brain. External tools such as notebooks, computers, or software may serve as an integral part of an extended cognitive system in case of reliable access and usage of these tools. Although EMT offers an explanation as to how professionals improve their performance due to human-artefact coupling, the issue of developmental risk is discussed by Risko and Gilbert (2016) in their concept of Cognitive Offloading defined as “the process of using physical or digital action to change the information processing demands of a task to lower the cognitive demands of the task.”

In educational setting, cognitive offloading implies a compromise between immediate task performance and internal knowledge accumulation. While offloading spatial navigation to GPS allows immediate improvement of wayfinding efficiency, it makes the process of forming internal cognitive maps impossible. Similarly, offloading problem framing, assumptions' checking, and financial modeling design to the language model will allow better immediate performance but no development of internal diagnostic heuristics among business students.

- **Productive Struggle, Desirable Difficulties, and Illusions of Competence**
  - The research in learning science clearly indicates that temporary performances achieved through training sessions are poor indicators of retention and transfer of learning. The idea of Desirable Difficulties has been proposed by Bjork and Bjork (2011), illustrating how the learning process becomes slower due to conditions that cause intentional cognitive effort

such as spaced practice, interleaving, and generation of answers. Analogously, Kapur's (2016) research on productive failure shows that struggling students in finding solutions for complex and ill-defined problems before getting formal instruction end up having a much better conceptual grasp compared to students who receive optimal solution paths instantly.

- Generative Artificial Intelligence essentially removes desirable difficulties since it gives instantaneous, plausible, and structurally sound outputs once there is any sort of ambiguity. This is a dangerous situation when we look at it in combination with two well-known cognitive biases:
  - Illusion of Explanatory Depth (IoED) (Rozenblit & Keil, 2002): Human beings systematically underestimate the complexity of complex causal systems until they try to explain or rebuild them from scratch. Fluency triggered through reading of a coherent, generatively AI-produced strategic analysis causes people to confuse familiarity of the text with mastery of the underlying reasoning.
- **The Dunning-Kruger Effect** (Kruger & Dunning, 1999): Students who do not possess basic proficiency within the field do not have a metacognitive schema that would enable them to determine whether or not their (or AI's) output is accurate. If the LLM generates a plausible financial story with mathematical errors, the ill-calibrated student accepts it without thinking and rates themselves proficiently.

- **Empirical Evidence on AI in Knowledge Work and Education**

Empirical research on human-AI cooperation provides contradictory yet valuable findings. According to Dell'Acqua et al. (2023), in a pioneering field experiment conducted with Boston Consulting Group consultants, there was the "jagged technological frontier". In tasks within the AI capability frontier, consultants with the help of AI performed tasks 25.1% faster and their output quality was evaluated 40% higher. However, when consultants faced the tasks slightly beyond the frontier and required nuanced contextual synthesis as well as recognition of contradictions between data, consultants with the help of AI performed 19 percentage points lower and generated incorrect solutions, since they were inclined to "fall asleep at the wheel" and uncritically use fluent AI fabrications.

In terms of higher education, according to Mollick and Mollick (2023), even though AI may serve as a personalized tutor or simulator, unmonitored access tends to foster cognitive shortcuts on the part of students. There are not many experimental researches in which the impact of GenAI during an entire semester is examined.

- **Formulation of Hypotheses**

Based on the theories of Cognitive Load, Cognitive Offloading, and Illusion of Explanatory Depth, we propose three testable hypotheses:

- Hypothesis 1 (Performance & Quality): Students who have constant access to GenAI technologies during their learning process will deliver initial quantitative case solutions with significantly better grades and faster times than students who use only traditional analytics software.
- Hypothesis 2 (Conceptual Sensitivity & Diagnosis Shortcoming): In proctored and unassisted examination conditions, students that have been exposed to GenAI will show marginal gains in terms of procedural setup but significantly poorer diagnostic skills for uncovering financial, math, and operational errors compared to their AI-restricted counterparts.
- Hypothesis 3 (Illusion of Expertise): Students that are heavily reliant on AI will demonstrate marked overconfidence in their competence level, as measured by significantly higher values of the Metacognitive Confidence Delta (ΔMC).

Research Methodology

- **Research Design and Institutional Context**

A randomized controlled trial (RCT) longitudinal cohort design over a period of 16 weeks of a postgraduate semester was used to effectively establish causality between GenAI usage and analytical skills development. This study took place in several sections of a foundational quantitative management

course (Financial Modeling and Quantitative Decision Analysis) at an accredited business school curriculum, which follows postgraduate management norms.

Approval was granted from the Institutional Review Board (IRB), and the informed consent of all participants was obtained. In order to promote fairness in the learning process and avoid disadvantages in future career development, an intensive 2-week boot camp on enterprise GenAI processes was offered to the AI-Restricted cohort after the final exam.

- **Sample Selection and Cohort Randomization**

The total number of respondents was $N=340$ Full Time MBA students. At the start of the semester, demographic information, academic records of the undergraduates and results of a standardized Quantitative Aptitude Diagnostic Test were obtained.

In order to control confounding caused by previous mathematical ability, Stratified Random Sampling technique was employed. Respondents were first stratified according to their undergraduate discipline background - STEM/Quantitative Commerce (Engineering, Mathematics, Statistics, B. Com Honours with Accounting/ Finance) or Humanities /Social Sciences/General Arts - and quantitative diagnostic quartiles at baseline. Within each stratum, students were randomly allocated at a ratio of 1:1 to either of the two experimental tracks:

- **Instructional Protocol and Case Curriculum**

Both classes had the exact same lecture content presented by the same professors across the four mandatory modules:

  - Module I (Weeks 1-4): Deterministic Financial Modeling, Working Capital Management, and Discounted Cash Flow (DCF) Valuation.
  - Module II (Weeks 5-8): Linear and Integer Programming for Optimal Operations and Logistics.
  - Module III (Weeks 9-12): Multivariate Regression Diagnostics, Heteroscedasticity, and Demand Forecasting.
  - Module IV (Weeks 13-16): Monte Carlo Simulation and Decision Making Under Uncertainty.

In each week of class, both classes had to solve the same challenging business case that entailed data preparation, building models, sensitivity analysis, and writing an executive summary. In the AI-Augmented Cohort, the students were advised to make use of LLM environments in order to generate formulas, write programs, fix any errors, and create managerial insight. In the AI-Banned Cohort, the students used spreadsheets and statistical tools exclusively.

- **Operationalization of Variables and Evaluation Metrics**
  - To ensure measurement of both the superficial and structural aspects of performance, three core performance metrics were defined and measured on a $0\text{--}100$ scale:
  - Project Quality & Time Taken to Complete: The weekly case reports were independently marked by two evaluators, who used a standardized rubric worth 100 points evaluating mathematical correctness (40%), analytical rigor (35%), and executive communication skills (25%). The inter-rater agreement was high (Cohen’s $\kappa = 0.87$). The time taken to complete the tasks in hours was reported and verified using timestamp records each week.
  - Procedural Execution Index (PEI): Assessed in the unaided final examination, the PEI assesses a student's ability to remember and use standard formulas, set up baseline mathematical programming parameters, and construct a standard financial statement projection.
  - **Diagnostic Reasoning Score (DRS):** Also assessed in the unassisted final examination is the ability of DRS to perform higher-order cognitive transfer and anomaly detection. The final exam consisted of embedding six intentional structural anomalies into realistic corporate databases (for example, a sudden shift mid-stream from FIFO to LIFO accounting methodology affecting gross margin calculations; non-linear corporate tax brackets making flat tax assumptions invalid; highly correlated variables hiding behind a high R^2; and circularity between debt and interest payments in a leveraged buyout schedule). DRS

represents the proportion of embedded anomalies that were successfully identified, quantified, and fixed.

- Metacognitive Confidence Delta (ΔMC): Following the completion of the unassisted final exam—prior to any grading—the students were required to fill out a reflective questionnaire estimating their score (E_{score}) exactly. ΔMC is calculated as follows:

$$\Delta MC = E_{\text{score}} - A_{\text{score}}$$

where A_{score} is the student's actual score on the unassisted exam. Positive values represent overconfidence while values near zero represent accurate metacognitive calibration.

- **Statistical Analysis Plan**

Analysis of data was done using SPSS software (version 29) and R software (version 4.3). Two-tailed independent samples t-tests were run to compare means for coursework performance, time taken to complete test, PEI, DRS, and ΔMC. Effect sizes were computed using Cohen's d method in addition to significance levels (p-values). In addition, a 2×2 analysis of variance (ANOVA) was also run to investigate interaction between Cohort Track (AI-Enabled vs. AI-Restricted) and Prior Academic Background (STEM/Quant vs. Non-Quant) on Diagnostic Reasoning Scores.

Empirical Results and Data Analysis

- **Performance Comparison Descriptive and Inferential**

All 340 participants completed the 16 weeks study protocol and took the proctored final exam, thus having 100% retention rate. Table 1 presents comparative empirical performance of both cohorts on all primary evaluation measures.

- **Assisted Coursework Productivity vs. Unassisted Examination Outcomes (H1 & H2)**

These results give strong support to the Hypothesis 1. In all 12 weekly case assignments, AI-Enabled students significantly outperformed their AI-restricted classmates with an average gap of 12.20 percentage points (88.40 against 76.20, $t = 16.98$, $p < 0.001$, $d = 1.84$). Moreover, students in the AI-Enabled category managed to score higher while spending 44.8% less time weekly on task completion (3.20 hours vs. 5.80 for control group students, $p < 0.001$). Visual inspection of submission records showed that GenAI tools helped to virtually eliminate syntax errors in R/Python code, speed up data cleaning procedures, and create highly coherent executive summaries.

Nonetheless, once the technological scaffolding was stripped away for the proctored final exam, the performance gap reversed dramatically, providing strong evidence for Hypothesis 2:

- Surface Procedural Execution Intact: Using the Procedural Execution Index (PEI), there was no statistical difference in performance between the AI-Enabled ($M = 72.10$, $SD = 16.40$) and AI-Restricted ($M = 74.80$, $SD = 17.55$) cohorts ($t = -1.47$, $p = 0.142$). Thus, exposure to AI-generated problems did not harm students' abilities to memorize standard formulas and perform routine algebraic setup.
- Diagnostic Reasoning Collapse: Using the Diagnostic Reasoning Score (DRS), the AI-Enabled group performed 19.20 percentage points worse than the AI-Restricted group (59.40 vs. 78.60, $t = -13.15$, $p < 0.001$, Cohen's $d = -1.43$). At the item level, whereas 82% of AI-Restricted students spotted the mid-period accounting switch from FIFO to LIFO inventory accounting, which required adjustments to free cash flow, only 34% of AI-Enabled students did so. Also, 76% of the control group students identified the tax rate threshold nonlinear trap, while just 29% of AI-reliant students did so, despite having solved very similar edge cases successfully during weekly classes aided by LLM prompts

- **Subgroup Analysis: Impact of Prior Quantitative Background**

A two-way ANOVA test to determine if previous mathematical instruction acted as a shield to cognitive offloading was performed by testing the impact of Cohort Track and Undergraduate Background (STEM/Quantitative versus Non-Quantitative) on Unassisted Diagnostic Reasoning Scores (DRS).

The ANOVA showed significant main effects for both Cohort Track ($F(1, 336) = 184.2$, $p < 0.001$) and Previous Background ($F(1, 336) = 78.6$, $p < 0.001$) as well as an interaction effect ($F(1,

336) = 9.42, p = 0.002$). It is important to note that while both groups of students exhibited large diagnostic reasoning gaps as a result of GenAI use, those with a background in non-quantitative courses had a significantly larger gap ($\Delta = -22.90$ versus -15.60 in STEM students). Without the previous schema of quantitative analysis, non-quant students substituted their cognitive reasoning for the AI-generated output.

- **Metacognitive Miscalibration Analysis (H3)**

Post-exam self-prediction analysis supported Hypothesis 3. Although the AI-Enabled group performed significantly worse in the composite unassisted exam (65.75 vs. 76.70), they showed significantly higher self-predictions of score ($M = 84.45$) as compared to the AI-Restricted group ($M = 79.80, p < 0.001$).

As a result, the Metacognitive Confidence Delta (ΔMC) was significantly higher in the AI-Enabled group ($+18.70$ percentage points, $SD = 11.20$) than in the control group ($+3.10$ percentage points, $SD = 8.40, t = 14.53, p < 0.001$). About 68% of students in the AI-Enabled group overestimated their final exam score by more than 15 points, while 79% of students in the AI-Restricted group estimated their score within ± 6 points margin of accuracy.

Discussion

- **Synthesized Fluency Without Structural Anchoring**

The major empirical finding of this research is the discovery of a novel cognitive pattern specific to business education mediated by artificial intelligence: Synthesized Fluency Without Structural Anchoring.

In classical education contexts, the skill that enables a learner to construct an elegant financial model or a well-articulated strategic memo is deeply connected with his/her knowledge; he/she cannot explain why the increase in working capital lowers free cash flow without first grappling with the accounting principles linking the balance sheet with the cash flow statement. The connection is broken by GenAI. As LLMs respond to short natural language queries with perfectly written text, elegantly formatted tables, and flawless code, learners often create expert artifacts with novice mental models.

- **Mechanisms of Cognitive Degradation: Heuristic Atrophy and Contextual Inelasticity**
  - Careful analysis of student error patterns and LLM prompt logs show two cognitive risks created by unchecked reliance on GenAI:
  - Heuristic Atrophy (Loss of Sanity-Checking Reflexes): Financial analysts and operations managers have developed an instinctive set of “rules of thumb” and sanity-checking heuristics for themselves, such as automatically questioning the validity of a terminal growth rate that is greater than the growth in nominal GDP, or investigating the cause of an increase in inventory turnover ratio accompanied by a footnote to the accounting policy. In the AI-Restricted group, the students spent 5.8 hours each week performing manual reconciliations; in doing so, they internalized those diagnostic reflexes. However, in the case of the AI-Enabled group, students left the task of anomaly detection to the LLM during their studies.
  - **Contextual Inelasticity:** While typical LLMs have no issues with finding patterns in the familiar structures of textbook concepts (e.g., Porter's Five Forces, WACC computations, simple OLS regressions), when dealing with more disorganized, less structured or unconventional problem formulations on the exam without assistance, AI-dependent students demonstrated high inelasticity with respect to context. Having become accustomed to instructing an AI to “structure a financial model according to the attached data,” they found it hard to start the process of problem structuring from scratch. While their procedural knowledge was good enough to do the math after being told what to do, their executive skills had deteriorated.

- **The Metacognitive Trap in Management Education**

Most worrisome for business education and corporate recruiting is the $+18.7$-point Metacognitive Confidence Delta (ΔMC). In managerial decision-making processes, it is much more dangerous to be incompetent without knowing that one is incompetent than ignorant. Manager

aware of their quantitative limitations will check, ask specialists, or stress test assumptions. On the contrary, an MBA graduate, plagued by an AI-generated Illusion of Explanatory Depth (Rozenblit & Keil, 2002), would exhibit confidence in faulty reasoning.

Every contact with the LLM during weekly work on the subject led the learner to a mistaken belief of mastery: type a prompt, receive a clear explanation, understand the reasoning in the text, and think "I can do that." From the standpoint of cognitive psychology, this is a typical confusion of recognition (understanding of an explanation) and generation (creation of a solution). The illusion breaks in an unassisted context, but learners remain unaware of their analytical weaknesses straight after the test.

- **Implications for Indian Higher Education and UGC Quality Mandates**

Given that the Indian education system is governed by the Learning Outcomes-based Curriculum Framework (LOCF) by the University Grants Commission (UGC) and the National Education Policy (NEP) 2020, these findings have significant policy implications in this domain. According to both NEP 2020 and UGC guidelines, emphasis is placed on moving from rote learning to higher-order cognitive skills: critical thinking, analytical problem solving, and innovation.

Allowing free GenAI usage in business schools and commerce departments in the form of continuous internal evaluation (CIE), which includes assignments, case studies, and project reports, will lead to grade inflation and the loss of analytical skills. Students will graduate with excellent CGPAs but fail at their analytical work in corporate companies and be unable to perform their managerial duties because of their inability to think critically and analytically. While banning GenAI is also impossible since working with humans and AI is now an essential workplace skill, the transition should be made from allowing unrestricted GenAI integration to building cognitive architecture.

Pedagogical Recommendations for Business Curricula

In order to harness the huge productivity and analytical power of Generative AI while ensuring the acquisition of foundational schemas, business schools and faculties of commerce need to develop an instructional and evaluation framework comprising of three levels:

- **Level 1:** Bifurcated Assessment Design (Disentangling Acquisition from Synthesis)

The curricula need to bifurcate foundation schema learning from AI-assisted strategic synthesis through bifurcated evaluation weightages:

  - Closed-Loop Foundational Assessments (Minimum 50-60% Weightage): The core quantitative mechanics including financial statement reconciliation, adjustments, linear programming formulation, and statistical assumptions must be assessed through proctored, closed-network, unassisted assessments (for example, mid-term exams, in-class viva voce, and live whiteboard case defense). This ensures the existence of "desirable difficulties" (Bjork & Bjork, 2011) and forces students to develop internal cognitive schemas prior to using generative technology.
  - Open-Loop Augmented Projects (40-50% Weightage): After assessing the foundational competence of the students, capstone projects and multi-variable simulations must use GenAI explicitly, thereby elevating the standards of grading to assess large scale data synthesis and strategic creativity that cannot be done manually.

- **Tier 2: Adversarial AI Critiques ("Red-Teaming" GenAI Outputs)**

Instead of assigning students to write answers using GenAI, educators need to reverse engineer their pedagogical strategy to devise Adversarial AI Critique assignments that:

  - Teachers produce highly elaborated case analyses, financial calculations, or strategic memos through LLMs that include mathematical, accounting, and logical hallucinations purposefully inserted into them (for example, the use of currencies for discount rates not matching, survivorship bias in the markets' data, or violations of regression assumptions).
  - The students need to play the role of senior auditors and scrutinize this AI-produced document to find all the mistakes in it, understand the underlying theoretical breach, and calculate the corrected values themselves.
  - This method directly attacks Heuristic Atrophy and Metacognitive Miscalibration by forcing the students to develop professional skepticism towards LLM outputs.

- **Tier 3: Mandatory Prompt-Audit Logs and Epistemic Traceability**
  - In all cases involving homework and research papers where GenAI is allowed, institutions must require a Prompt Audit Log as part of the final submission:
  - Iterative Prompt Transcript: All documentation related to the prompts that the student used, showing how they broke down the problem into manageable bits instead of using a one-shot "solve this case" instruction.
  - Verification & Benchmark Trace: A compulsory reflection table with at least three times when the student verified an AI-generated calculation or assertion by comparing it with actual financials, textbooks, or empirical data.
  - Modification Transparency: Clear delineation between the parts drafted by the AI and those written by the student themselves. The grading scale for such projects should give around $25\text{--}30\%$ credit to the verification trace alone.

Limitations and Future Research Directions

Although this investigation provides strong experimental data over an entire academic semester period, some of its limitations need to be noted for future research:

- Limitation of Single-Discipline Study Group: Our sample involved only MBA postgraduates specializing in quantitative finance and operations modules. It needs to be investigated whether the same diagnostic reasoning deficiencies persist in other qualitative management fields, like organizational behavior, human resource strategy, and marketing ethics.
- Longitudinal Career Monitoring: This experiment ended after 16 weeks of the semester course. Longitudinal monitoring of graduates for the first 12 to 24 months of their career will determine whether the problem of cognitive offloading at early stages hinders future executive performance, or on the contrary, workplace responsibility quickly adjusts their diagnostic skills.
- Efficacy of Interventions Test: Although we suggest a three-level solution model (Bifurcated Assessment, Adversarial Critiques, and Prompt-Audit Logging), future randomized experiments need to investigate the precise recovery rate of Diagnostic Reasoning Score (DRS) and Metacognitive Calibration (ΔMC) after structured interventions training of students.

Conclusion

Generative AI is the most impactful cognitive technology to have entered the business curriculum since personal computing. As shown in our 340-student randomized controlled trial, constant exposure to LLMs significantly speeds up task completion time (44.8% increase in speed) and improves the surface quality of case assignments ($+12.2$ percentage points). But in the absence of well-defined pedagogic controls, GenAI will cause Synthesized Fluency Without Structural Anchoring. Offloading the relevant cognitive load involved in schema creation causes heuristic impairment, impairs autonomous diagnostic thinking (-19.2 percentage points) and induces metacognitive overconfidence ($\Delta MC = +18.7$ points).

For modern business schools, commerce departments, and regulatory agencies like the UGC, the path forward is obvious. Academic rigour cannot be measured only on the basis of how polished the student's AI-assisted work is. If Generative AI is to be sustainably adopted into educational systems, then careful curriculum design involving unaided foundational tests, AI audits and mandatory epistemic logbooks is essential.

References

1. Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.
2. Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.
3. Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on worker productivity and quality* (Harvard Business School Technology & Operations Management Unit Working Paper No. 24-013).

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

5. Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134.

6. Mollick, E. R., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *SSRN Electronic Journal*.

7. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688.

8. Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.

9. Sweller, J., van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296.

10. University Grants Commission (UGC). (2020). *Learning Outcomes-based Curriculum Framework (LOCF) for Postgraduate Programmes in Commerce and Management*. Ministry of Education, Government of India.

❑❑❑