

Student Learning and Creativity in the Pre-AI and AI-Assisted Era:

A Research Foundation, Conceptual Framework, and Mixed-Methods Protocol

Prof. Vishwas Kenchi^{#1}, Saloni Palkar^{#2}, Prajakta Pandhare^{#3}

#Master of Computer Applications, ZIBCAR, Savitribai Phule Pune University(SPPU), Pune, Maharashtra, India

¹vishwas.kenchi@zealeducation.com, ²salonipalkar90@gmail.com, ³prajaktapandhare8@gmail.com

ORCID ID: Prof. Vishwas Kenchi: [0009-0006-8832-2291](https://orcid.org/0009-0006-8832-2291)

ORCID ID: Saloni Palkar: [0009-0005-7570-7371](https://orcid.org/0009-0005-7570-7371)

ORCID ID: Prajakta Pandhare: [0009-0002-9277-6385](https://orcid.org/0009-0002-9277-6385)

Abstract—Generative artificial intelligence has entered classrooms faster than the evidence needed to judge its educational value. Most published work so far reports gains in task performance, speed, and perceived personalisation, while far less attention has been paid to what happens to the qualities that education exists to build, namely the ability to think without assistance, to produce original ideas, to solve unfamiliar problems, and to understand concepts deeply enough to transfer them. This paper develops the research foundation and the research contribution for a study that compares student learning and creativity in the era before artificial intelligence and the era of artificial intelligence assistance. Drawing on cognitive offloading theory, the desirable difficulties and productive failure literatures, sociocultural theories of scaffolding, self determination theory, and the componential model of creativity, the paper argues that the educational effect of generative artificial intelligence depends less on whether students use it than on the cognitive role it plays, either a substitute that performs the thinking or a scaffold that supports the student in performing it. The paper contributes an integrative Cognitive Role framework that links patterns of use to learning and creativity outcomes through four mechanisms, a set of testable propositions about moderation by educational level, field of study, and frequency of use, a convergent mixed methods design that combines cohort comparison, a controlled experiment, and a paired student and teacher perception study, and a practice oriented model of responsible use for teachers and institutions. No empirical results are reported. The paper is a theory building and protocol contribution intended to guide and justify the forthcoming data collection.

Keywords—cognitive offloading; creativity; generative artificial intelligence; higher education; independent thinking; mixed methods; problem solving; responsible use

I. INTRODUCTION

For most of the twentieth century, the central technological question in education was how to give students access to information. Libraries, textbooks, and later the internet progressively lowered the cost of finding facts. Generative artificial intelligence (GenAI) changes the question in a more fundamental way, because it lowers the cost of producing the work itself. A student who once needed hours to draft an essay outline, debug a program, or sketch a design concept can now obtain a plausible version in seconds. The output of learning, the artefact that teachers have long used as evidence of understanding, can be separated from the process of learning that was supposed to generate it.

This separation is the source of both the excitement and the anxiety surrounding GenAI in education. Supporters point to

round-the-clock feedback, language support for multilingual learners, patient explanation at any level, and help for students who do not have access to private tutoring. Critics point to the risk that students will submit work they do not understand, stop practising the difficult mental operations that build expertise, and converge on similar, statistically typical ideas rather than developing their own. Both positions contain truth, and both are often argued from anecdote. What the field lacks is a systematic, theory-grounded comparison of how students learn and create under traditional conditions and under assisted conditions, and a clear account of when the difference is beneficial and when it is harmful.

Three features of the existing literature motivate the present paper. First, the majority of studies measure outcomes that artificial intelligence (AI) is good at improving, such as the quality of a finished product, the time taken to complete it, or students' satisfaction with the experience. These outcomes can rise even while underlying capability stagnates or declines, because performance with a tool is not the same thing as learning [1]. Second, much of the evidence comes from a single source, student self-report, which is vulnerable to the well-documented tendency of people to misjudge their own competence [2]. Teachers' observations are rarely collected alongside students' reports, so the degree of agreement between the two is unknown. Third, little work examines whether effects differ across educational level, field of study, and intensity of use, even though there are strong theoretical reasons to expect that a novice in school mathematics and a postgraduate researcher in the humanities will experience AI very differently.

The purpose of this paper is to address these limitations by laying the conceptual and methodological groundwork for a comparative study. The paper does not report new data. Instead, it consolidates the research foundation, identifies the specific gaps, proposes an explanatory framework, and sets out a design capable of testing it. In doing so, it aims to move the debate from the question "Should students use AI?" to the more productive question "What is the AI doing in the student's thinking, and what does that do to the student?"

A. Research Questions

The study is organised around five research questions (RQ).

- RQ1. How do independent thinking, originality, creative problem-solving, and conceptual understanding differ

between students who learned before widespread GenAI availability and students who learn with it?

- RQ2. What positive and negative effects of GenAI on independent learning and creative thinking are reported by students and by teachers?
- RQ3. How closely do students' self-reported experiences align with teachers' observations of the same constructs?
- RQ4. Do these effects vary by educational level, field of study, and frequency and type of AI use?
- RQ5. Which responsible-use practices allow AI to support learning without weakening problem-solving ability and originality?

B. Scope and Key Terms

In this paper, generative AI refers to systems, most commonly large language models (LLMs) and image generators, that produce new text, code, images, or other content in response to natural-language prompts. The pre-AI era refers to the period before such systems were freely and widely available to students, which for practical purposes is the period before late 2022. The AI-assisted era refers to the period afterwards. The boundary is a convenience rather than a sharp line, because many students used earlier forms of automation such as grammar checkers, translation tools, and search engines. The distinction that matters analytically is the availability of a general-purpose tool that can complete an entire academic task on request.

Independent thinking is defined as the capacity to generate, evaluate, and revise ideas and solutions without external cognitive support. Originality is the degree to which a response is statistically uncommon and nevertheless appropriate to the task. Creative problem-solving is the ability to produce solutions to open-ended or ill-structured problems that are both novel and effective. Conceptual understanding is the ability to explain why a principle works, to connect it to other principles, and to apply it in a new context. These definitions follow established usage in the cognitive and creativity literatures and are operationalised in Section IV.

C. Organisation of the Paper

The remainder of the paper is organised as follows. Section II reviews the theoretical and empirical foundations and identifies gaps in the literature. Section III develops the conceptual framework and derives testable propositions and hypotheses. Section IV describes the research design, measures, and analytic strategy. Section V sets out the research contribution. Section VI translates the framework into a model of responsible AI use, Section VII discusses implications, challenges, and directions for future work, and Section VIII concludes.

II. RESEARCH FOUNDATION

A. Learning and Creativity in the Pre-AI Environment

To measure change, the baseline must be described accurately. It would be a mistake to treat the pre-AI classroom as a golden age of independent thought. Students have always had ways of avoiding effortful thinking: copying from peers, relying on commercial essay services, memorising model answers, and using solution manuals. Nor were pre-AI assessments always good at measuring creativity; many rewarded reproduction of taught procedures. Nevertheless,

the pre-AI environment had a structural property that matters for learning: producing an acceptable piece of work generally required the student to do a substantial amount of the cognitive labour personally. Even a student who cheated had to find, adapt, and assemble material. The cost of producing work therefore acted as an implicit, if imperfect, guarantee that some thinking had taken place.

Pre-AI learning also had characteristic strengths and weaknesses. Struggle with a blank page, a failed proof, or a program that would not compile was common and frustrating, but it generated the kind of effortful processing that cognitive psychologists associate with durable learning. At the same time, help was unevenly distributed. Students with educated parents, private tutors, or strong peer networks received the individual feedback that others did not. This inequity is a key reason why many educators greet GenAI with interest rather than hostility, and it is why the research question cannot be reduced to a simple verdict of harm or benefit.

B. Theoretical Foundations

1) Cognitive offloading

Cognitive offloading is the use of physical action or external tools to reduce the mental demands of a task [3]. Offloading is not inherently harmful. Writing a shopping list, using a calculator, and storing a phone number in a contact book are rational uses of external resources that free attention for other purposes. The educational problem arises when the offloaded operation is the very one the curriculum is trying to develop. Research on the so-called Google effect showed that when people expect information to remain accessible, they remember the information less well and remember where to find it instead [4]. Generative AI extends this logic from storing facts to performing reasoning, summarising, planning, and composition. If the student routinely outsources these operations, the opportunity to practise them diminishes, and the capacity to perform them without help may weaken. The offloading perspective therefore generates a clear prediction: the more a student delegates the target cognitive operation to AI, the smaller the gain in unassisted performance on that operation.

2) Desirable difficulties and productive failure

Bjork and Bjork [5] argued that conditions that make learning feel harder in the short term, including spacing, interleaving, and retrieval practice, frequently produce better long-term retention and transfer. The key insight is that performance during learning is an unreliable indicator of learning itself. Conditions that boost immediate performance can leave learners with a false sense of mastery. This is directly relevant to GenAI, which is extraordinarily effective at boosting immediate performance. A student who obtains a correct solution from a chatbot may feel competent without having done the retrieval or reasoning that would make the solution available later. Closely related is the work on productive failure, which shows that students who struggle with a problem before receiving instruction often learn more deeply than students who receive instruction first, even though their initial attempts fail [6]. Immediate AI assistance can remove the struggle phase altogether. Taken together, these literatures predict that unrestricted AI access will improve practice performance while impairing later unassisted performance, and that the harm can be reduced

when the AI is designed or used to preserve productive struggle.

3) *Scaffolding and the zone of proximal development*

Vygotsky's [7] concept of the zone of proximal development describes the gap between what a learner can do alone and what the learner can do with guidance. Later work formalised guidance as scaffolding: temporary, adjustable support that is withdrawn as competence grows. The decisive property of scaffolding is its fading. A scaffold that never fades is not a scaffold but a prosthesis. This distinction offers a principled way of classifying AI use. When an AI tool asks guiding questions, offers a hint rather than an answer, and steps back as the student improves, it functions as a scaffold. When it supplies finished solutions on demand, it functions as a substitute. The same technology can therefore sit at either end of the continuum depending on how it is configured and used, which is the central claim developed in Section III.

4) *Self-determination theory*

Self-determination theory proposes that sustained motivation and psychological well-being depend on satisfaction of three basic needs: autonomy, competence, and relatedness [8]. AI can support these needs when it increases a learner's sense of being able to do things and to choose how to approach them. It can undermine them when the learner begins to feel that the real work is done by the machine, which erodes the sense of competence and ownership, or when the tool displaces interaction with teachers and peers. Because creativity is strongly associated with intrinsic motivation, effects on perceived autonomy and competence are likely to be a pathway through which AI use influences creative outcomes.

5) *The componential model of creativity*

Amabile's [9], [10] componential model holds that creative performance emerges from the interaction of three components within the individual: domain-relevant skills, creativity-relevant processes such as flexible thinking and willingness to take risks, and task motivation. A fourth, external component is the social environment. GenAI can influence each element. It can strengthen domain skills through explanation and feedback, but it can also reduce the need to acquire those skills. It can model creative processes by producing unexpected combinations, but it can also encourage students to accept the first plausible output rather than persist with their own divergent thinking. It can raise motivation by lowering the barrier to starting, or lower it by making the task feel trivial or the result feel not truly one's own. The model is useful because it prevents a one-dimensional reading of AI as simply "good" or "bad" for creativity.

6) *Metacognition and self-regulated learning*

Self-regulated learning describes how learners plan, monitor, and evaluate their own learning [11]. Effective learners know what they do not understand and take steps to address it. Generative AI interacts with this process in two opposite ways. It can support regulation by providing prompt feedback and prompting reflection, but it can also encourage what Fan et al. [12] termed metacognitive laziness, in which learners defer planning and monitoring to the tool. A learner who asks the AI what to do next, rather than deciding, is surrendering part of the regulatory role. Because

metacognition is a strong predictor of achievement and transfer, any adequate account of AI and learning must include it.

C. *Empirical Evidence on Generative AI and Learning*

The empirical literature is young, heterogeneous, and rapidly changing, and findings should be read with caution. Several streams can nevertheless be identified.

1) *Performance and productivity gains*

Evidence from professional settings shows that GenAI can substantially improve the speed and quality of routine writing tasks. In a pre-registered online experiment, Noy and Zhang [13] found that access to ChatGPT reduced the time taken to complete mid-level professional writing tasks and increased rated quality. Reviews of the educational literature report similar benefits, including personalised explanation, immediate feedback, and support for lesson preparation and assessment design [14]. These findings support the claim that AI is a powerful tool. They do not, however, show that it improves the capabilities of the person using it once the tool is removed, which is the question that matters for education.

2) *Evidence of learning harm and its mitigation*

The most directly relevant evidence to date comes from a large field experiment in high-school mathematics, in which students were given either no AI, a basic GPT-style assistant, or a version designed as a tutor that offered hints rather than full solutions [15]. Students with unrestricted access performed markedly better during practice, but when access was removed they performed worse than students who had never had the tool. The tutor-style version largely eliminated this negative effect. The study is important for two reasons. It demonstrates that the performance-learning gap predicted by the desirable difficulties literature appears in real classrooms, and it shows that design and instruction can alter the result. At the same time, a single study in one subject and one age group cannot establish general conclusions, which is why replication across levels and fields is needed.

3) *Engagement, critical thinking, and metacognition*

Other work has examined thinking processes rather than test scores. Kosmyna et al. [16], in a preprint using electroencephalography (EEG) measurements, reported that participants who wrote essays with an LLM showed weaker neural connectivity and lower recall of their own text than participants who wrote unaided or with a search engine. As a small, non-peer-reviewed study its conclusions are provisional, but its direction is consistent with the offloading account. In a survey of knowledge workers, Lee et al. [17] found that higher confidence in AI was associated with less reported critical thinking effort, whereas higher self-confidence was associated with more. Gerlich [18] similarly reported a negative association between frequent AI tool use and critical thinking scores, mediated by cognitive offloading, with younger participants more affected. Fan et al. [12] found in an experimental study that learners supported by ChatGPT improved the quality of their essay revisions but did not show corresponding improvements in knowledge gain or transfer, and displayed signs of metacognitive laziness. Because most of these studies are correlational or short-term, causal interpretation is limited.

4) Creativity evidence

The creativity evidence is mixed in an instructive way. Urban et al. [19] reported that university students who used ChatGPT on a creative problem-solving task produced better solutions than students who did not, and that the tool improved task performance and perceived effort. In a study of short story writing, Doshi and Hauser [20] found that access to AI-generated ideas increased the rated creativity of individual stories, particularly for writers who scored lower on a baseline creativity measure, but that stories written with AI assistance were more similar to one another. Individual creativity improved while collective diversity declined. This pattern points to an important distinction between the quality of a single product and the originality and variety of a population of products, and it raises the possibility that AI may raise the floor of creative performance while lowering the ceiling of distinctiveness.

D. Educational Level, Discipline, and Intensity of Use

Theory suggests that the effect of any instructional support depends on the learner. The expertise reversal effect describes how guidance that helps novices can become redundant, or even detrimental, for more knowledgeable learners, because it competes with the knowledge structures they have already built [21]. Applied to generative AI, this implies that the same tool may play very different roles for a secondary-school student meeting a topic for the first time and for a doctoral student checking a draft in a field they know well. The novice lacks the knowledge needed to evaluate AI output and is therefore more likely to accept it, while the expert can treat the same output as a hypothesis to be tested.

Cognitive load theory adds a complementary point. Working memory is limited, and instructional design should reduce extraneous load while preserving the germane effort that produces durable knowledge structures [22]. AI that removes extraneous load, such as formatting, syntax, or clerical detail, may free capacity for learning. AI that removes germane load, such as the reasoning itself, may leave little to be learned. Which of the two is happening in a given episode is an empirical question, and it depends on what the task is intended to develop.

Disciplines differ in what they ask students to produce and in how easily that production can be automated. Where the central assessment product is a text or a standard program, current systems can often generate an acceptable response with little guidance. Where learning depends on physical practice, fieldwork, studio critique, or interpersonal judgement, the artefact is harder to outsource and the AI is more likely to serve as an adjunct. These differences do not make any field immune. They alter the form that substitution takes and the cues by which teachers might detect it. A study that pools disciplines risks obscuring such differences, which is why field of study is treated here as a moderator and not as a nuisance variable.

Frequency of use is commonly treated as a dose, with the implicit assumption that more use produces more of any effect. This is unlikely to hold. A student who consults AI once a week to check understanding may differ qualitatively from one who consults it constantly to generate every deliverable, and a student who consults it constantly but only to clarify concepts may differ from both. Frequency must

therefore be interpreted jointly with purpose and role. Active engagement with material is generally associated with deeper learning than passive reception, a distinction captured by the Interactive, Constructive, Active, Passive (ICAP) framework, in which interactive and constructive engagement outperform active and passive modes [23]. The role distinctions proposed in Section III can be read as an application of this principle: scaffold and collaborator roles invite constructive and interactive engagement, whereas substitute use encourages passive reception of finished output.

E. Teachers' Perspectives and the Perception Gap

Teachers occupy a distinctive vantage point. They observe many students over time, they know the course standards, and they see the products that students submit. They do not, however, observe the process by which those products were created, and in an AI-assisted setting the process is exactly what determines educational value. Teachers' impressions may therefore be driven by cues such as polish, generic phrasing, or sudden changes in quality rather than by direct evidence of how the work was produced. Such cues are fallible. A student who writes fluently without AI may be viewed with suspicion, while a student who edits AI output carefully may pass unnoticed.

Students, by contrast, have privileged knowledge of their own process but limited ability to judge how much they have learned. People systematically misjudge their competence, and learners often equate fluency with understanding, a tendency that is especially relevant when an AI delivers fluent explanations on demand [1], [2]. Social desirability is a further complication, since students may under-report behaviour they believe is disapproved of, or over-report use they believe is expected. For these reasons neither group can be treated as a gold standard. Collecting both sets of perceptions, and comparing each with performance on unassisted tasks, makes it possible to estimate how far each is a reliable guide. The expectation, formalised in P4, is that the two groups will diverge, and that the direction and size of the divergence will itself be informative about how AI is experienced in classrooms.

F. Gaps in the Existing Literature

Synthesis of the foundations above reveals five gaps that the present study is designed to address.

- *Outcome gap.* Studies emphasise product quality and efficiency. Independent thinking, originality, and depth of conceptual understanding measured without AI are comparatively under-studied, especially together in one design.
- *Comparison gap.* Few studies compare AI-assisted learners with pre-AI learners on the same instruments, so claims of change over time rest mostly on assumption.
- *Perspective gap.* Student self-reports and teacher observations are seldom collected in parallel, leaving the accuracy of each unknown. Students may overestimate their independence, while teachers may overestimate the prevalence of misuse.
- *Moderation gap.* Evidence about differences by educational level, discipline, and intensity of use is thin. Theory predicts that novices, who lack the knowledge to

evaluate AI output, are more vulnerable than experts, but this has rarely been tested.

- *Practice gap.* Guidance on responsible use is plentiful but often normative and untested. Little evidence identifies which specific practices preserve problem-solving and originality.

III. CONCEPTUAL FRAMEWORK AND PROPOSITIONS

A. The Cognitive Role of Artificial Intelligence Framework

The foundations reviewed above converge on a single organising idea. The educational consequence of generative AI is determined primarily by the cognitive role it plays in the learner's work. We distinguish four roles along a continuum from full substitution to full scaffolding. In the substitute role, the AI performs the target cognitive operation and the student transcribes, lightly edits, or submits the result. In the shortcut role, the student attempts the task briefly and then turns to the AI at the first difficulty, so that the productive struggle is truncated. In the collaborator role, the student and the AI exchange ideas and the student retains responsibility for evaluating, selecting, and integrating them. In the scaffold role, the AI supplies hints, questions, examples, and feedback calibrated to the learner's level, and the support is progressively withdrawn.

These roles are defined by what the student does, not by which tool is used or how often. Two students may use the same chatbot for the same length of time and occupy different roles. This is a deliberate departure from studies that treat "AI use" as a single variable, and it explains why such studies frequently yield contradictory results: they average over roles with opposite effects.

TABLE I
FOUR COGNITIVE ROLES OF GENERATIVE ARTIFICIAL INTELLIGENCE IN STUDENT WORK (AUTHORS' FRAMEWORK)

AI role	What the student does	What the AI does	Expected effect on unassisted capability
Substitute	Copies or lightly edits output	Performs the target operation	Negative; practice opportunities lost
Shortcut	Attempts briefly, then requests the answer	Resolves difficulty early	Negative to neutral; struggle truncated
Collaborator	Evaluates, selects, and integrates ideas	Generates options and perspectives	Neutral to positive if evaluation is substantive
Scaffold	Works with hints and feedback	Prompts, questions, and fades support	Positive; support is withdrawn over time

B. Four Mechanisms Linking AI Use to Outcomes

The framework specifies four mechanisms through which the cognitive role of AI is expected to influence learning and creativity. They are grounded in the theories reviewed in Section II.

The first is the practice mechanism. Skills are strengthened through effortful performance, retrieval, and error correction. When AI performs the operation, the amount of relevant practice falls. This mechanism follows from cognitive offloading and desirable difficulties and predicts that unassisted performance will suffer when AI acts as substitute or shortcut.

The second is the monitoring mechanism. Learners who evaluate and verify AI output exercise metacognition and critical thinking, while learners who accept output uncritically do not. Verification is a form of retrieval and comparison against one's own knowledge. This mechanism predicts that the effect of collaboration depends on the learner's prior knowledge, because a student who lacks the knowledge needed to detect errors cannot monitor effectively.

The third is the motivation mechanism. Perceived autonomy, competence, and ownership of work shape intrinsic motivation, which in turn supports creative engagement. AI use can raise these perceptions when it enables students to attempt tasks they would otherwise abandon, and lower them when students feel that the machine is the true author.

The fourth is the convergence mechanism. Large language models generate outputs that reflect the statistical regularities of their training data. When many students draw on the same models, the resulting ideas may cluster, reducing the diversity and distinctiveness of student work even when each individual product is competent [20]. This mechanism concerns originality at the group level and implies that originality should be assessed relative to a reference population, not just by a general quality rating.

Fig. 1 summarises the framework as a causal map. The cognitive role of AI is the central explanatory variable. It is shaped by the student's characteristics and context, and it is open to influence through responsible-use practices. It affects outcomes through the four mechanisms, and the outcomes that matter are those demonstrated without AI.

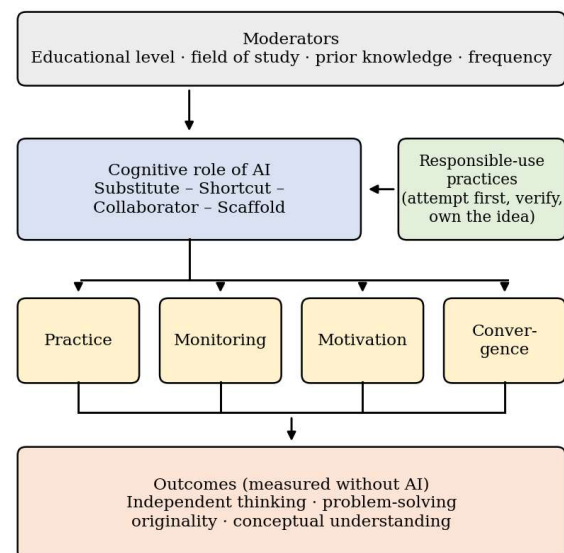


Fig. 1. Cognitive Role of Artificial Intelligence framework: moderators and responsible-use practices act on the role of AI, which affects outcomes through four mechanisms.

C. Propositions

From the framework we derive eight propositions. They are phrased as expectations to be tested and not as findings.

- P1 (Role). The relationship between AI use and unassisted problem-solving performance is determined by the cognitive role of the AI. Substitute and shortcut use is associated with lower unassisted performance,

while scaffold use is associated with equal or higher unassisted performance, compared with non-use.

- P2 (Performance-learning gap). AI-assisted performance during practice overstates learning. The difference between assisted and unassisted performance will be larger for substitute and shortcut use than for scaffold use.
- P3 (Originality). AI-assisted work will show higher average rated quality but lower between-student diversity of ideas than comparable pre-AI work.
- P4 (Perception gap). Students will rate the effect of AI on their own independent thinking more favourably than teachers rate it for their students, and students' self-ratings will correlate only moderately with performance-based measures.
- P5 (Level). Effects of substitute-type use on independent thinking will be stronger for school and early undergraduate learners than for postgraduate learners, because novices have less knowledge with which to monitor AI output.
- P6 (Field). Effects will be larger in fields where the core tasks are readily automated by AI (for example routine writing and standard coding) than in fields where the core task depends on embodied, situated, or interpersonal judgement.
- P7 (Frequency). The association between frequency of use and outcomes will be non-linear and conditional on role: moderate, purposeful use will show little or no decline, while high-frequency substitute use will show the largest decline.
- P8 (Responsible practice). Explicit instruction in AI use, together with assessment designs that require unassisted demonstration of understanding, will weaken the negative association between AI use and independent thinking.

D. Research Hypotheses

The propositions are operationalised as eight directional hypotheses (H), each stated in its alternative form, with the corresponding null being no difference or no association. Each hypothesis is tied to a research question, to the proposition from which it derives, and to the measures described in Section IV. The hypotheses will be registered before data collection (Section IV-H). RQ2 is addressed through the qualitative strand and is exploratory, so no directional hypothesis is stated for it.

- H1 (RQ1; P1). On the delayed unassisted problem-solving and conceptual transfer tasks, participants in the unrestricted-AI condition will score lower than participants in the no-AI control condition, whereas participants in the tutor condition will score no lower than control participants.
- H2 (RQ1; P2). The gap between performance during the learning phase and performance on the delayed unassisted test will be larger in the unrestricted-AI condition than in the tutor condition, and larger for substitute and shortcut use than for scaffold use, as coded from prompt logs.
- H3 (RQ1; P3). Compared with pre-2023 coursework, current-era coursework will show (a) higher mean rated

quality and (b) lower between-student diversity of ideas, indicated by higher mean pairwise similarity and lower statistical rarity of ideas. The same pattern of (b) is expected for the unrestricted-AI condition relative to the control condition in the experiment.

- H4 (RQ3; P4). Students' ratings of the effect of AI on their own independent thinking will be more favourable than teachers' ratings of the same construct for their students, and the correlation between students' self-ratings and their unassisted task scores will be only moderate (approximately .30 to .50).
- H5 (RQ4; P5). The negative association between substitute-type use and unassisted independent thinking scores will be stronger for upper-secondary and early undergraduate students than for postgraduate students, indicated by a significant educational level by role interaction.
- H6 (RQ4; P6). The negative association between substitute-type use and unassisted outcomes will be stronger in fields whose core tasks are readily automated by AI than in fields whose core tasks depend on embodied, situated, or interpersonal judgement, indicated by a significant field by role interaction. Fields will be classified by a task audit completed before data collection.
- H7 (RQ4; P7). The association between frequency of AI use and unassisted outcomes will be non-linear and will depend on role: moderate, purposeful use will not differ reliably from non-use, whereas high-frequency substitute use will be associated with the lowest scores, indicated by a significant quadratic or spline term and a frequency by role interaction.
- H8 (RQ5; P8). The negative association between substitute-type AI use and unassisted independent thinking will be weaker among students who report explicit instruction in AI use and whose courses include unassisted assessment components than among students without them, indicated by a significant moderation effect.

E. Boundary Conditions and Rival Explanations

Any framework should state where it is not expected to apply and which alternative explanations could produce the same observations. Three boundary conditions are noted. First, the framework concerns learning outcomes that depend on the student's own performance of a cognitive operation. It does not address outcomes such as access to information or administrative efficiency, for which substitution may be entirely appropriate. Second, it assumes that the student has some motivation to learn. Where motivation is purely extrinsic, for example when a grade is the only goal, the incentive to use AI as a substitute is stronger, and the effect of role may be amplified by assessment design. Third, it concerns present-generation systems whose outputs are produced in response to prompts. Systems that proactively tutor or that adapt over long periods may behave differently.

Several rival explanations deserve attention. Differences between pre-AI and current cohorts could reflect changes in curriculum, student demographics, pandemic-era disruption, or assessment style rather than AI. Differences between heavy and light users could reflect selection, since students

who are already struggling or already highly motivated may choose to use AI differently. Apparent differences in originality could result from rater bias, particularly if raters can infer the source of a response from its style. Participants in an experiment may behave differently because they know they are observed. The design addresses these threats through random assignment, blinded and format-standardised scoring, covariate adjustment for prior achievement, and triangulation across components, and the discussion of results will note the threats that remain.

IV. METHODOLOGY

A. Research Design

The study uses a convergent parallel mixed-methods design [24]. Quantitative and qualitative strands are collected in the same period, analysed separately, and then integrated in interpretation. Mixing methods is justified because the research questions require both measurement of differences (RQ1, RQ3, RQ4) and understanding of experiences and practices (RQ2, RQ5). A design that relied only on surveys would reproduce the self-report limitation identified in Section II-F, and a design that relied only on tasks would not explain why differences arise.

Because no student can be observed in both eras, the comparison between pre-AI and AI-assisted learning cannot be made within individuals. The study therefore triangulates three complementary comparisons, each with different strengths and weaknesses, as summarised in Table II.

TABLE II
THREE COMPONENTS OF THE COMPARATIVE DESIGN

Component	Comparison logic	Main strength	Main limitation
A. Archival cohort comparison	Work from pre-2023 cohorts versus current cohorts, rated with the same rubric by blinded raters	Direct historical contrast on real coursework	Cohorts differ in other respects; assignments may have changed
B. Controlled experiment	Random assignment to AI-unrestricted, AI-tutor, and no-AI conditions, followed by an unassisted test	Strongest causal evidence on role and learning	Short duration; artificial setting
C. Paired perception study	Parallel surveys and interviews with students and teachers	Reveals agreement and divergence between perspectives	Self-report and recall bias

B. Participants and Sampling

Participants will be students and teachers at schools and higher education institutions. Stratified purposive sampling will ensure representation across three educational levels (upper secondary, undergraduate, postgraduate) and four broad fields of study (science and engineering, humanities and social sciences, business and management, and creative and applied arts). Teachers will be sampled from the same institutions and departments to allow matching at the programme level. A priori power analysis, using conventional parameters [25] of 80 per cent power, a significance level of .05, and a small-to-medium effect size, will determine the final sample size for the experiment and for the moderation analyses, which typically require larger samples than main effects. The survey strand will aim for a sample large enough

to estimate subgroup differences with acceptable precision. For the qualitative strand, purposive sampling will continue until thematic saturation is judged to have been reached, which is typically between 20 and 30 interviews per stakeholder group.

C. Instruments and Measures

Four families of measures will be used. All instruments will be piloted, and reliability and construct validity will be examined before main data collection.

Student survey. The survey will capture frequency of AI use, the tasks for which it is used, and, crucially, the role of AI in each use. Role will be measured with behavioural items, for example whether the student typically attempts the problem before consulting AI, whether the student checks AI output against other sources, and whether the student rewrites AI output substantially. Additional scales will measure creative self-efficacy (adapted from [26]), self-regulated learning and metacognitive awareness, perceived autonomy and competence, and self-rated independent thinking. Items will be adapted rather than reproduced wholesale, with permissions obtained where required.

Teacher survey and interviews. The teacher survey mirrors the student instrument using observer-perspective wording, so that the same construct is rated by two sources. Semi-structured interviews will explore observed changes in originality, depth of understanding, reliance, and classroom practices, and the strategies teachers use to manage AI.

Performance-based tasks. Two tasks will be administered without AI. The first is an open-ended problem-solving or design task scored for fluency, flexibility, originality, and reasoning quality, following the divergent-thinking tradition [27], [28]. The second is a conceptual explanation and transfer task in which students explain a principle in their own words and then apply it in an unfamiliar context. Responses will be rated by at least two trained raters who are blind to condition or cohort, using a rubric developed and calibrated in the pilot. Inter-rater agreement will be reported, and for overall creative quality the Consensual Assessment Technique may be used, in which independent expert judges rate products relative to one another [29]. Originality will additionally be assessed statistically, as the rarity of an idea within the sample.

Archival coursework. For the cohort comparison, anonymised assignments and examination scripts from comparable courses before 2023 will be retrieved with institutional approval. Where original task prompts have changed, only courses with sufficiently similar tasks will be included, and the comparison will be restricted to dimensions that can be rated consistently.

D. Scoring Rubric and Quality Assurance

The validity of the performance-based strand depends on the quality of scoring. Responses will be scored on five dimensions, summarised in Table III, each on a five-point scale with written anchors for every level. Anchors will be drafted from the creativity and problem-solving literature, refined in the pilot using real responses, and frozen before main scoring begins.

TABLE III
SCORING DIMENSIONS FOR THE UNASSISTED TASKS

Dimension	Definition	Scoring focus
Fluency	Number of distinct, relevant ideas or solution steps produced	Count of valid, non-duplicate ideas
Flexibility	Number of different categories or approaches represented	Variety of solution strategies
Originality	Rarity of the idea within the sample, together with its appropriateness	Frequency-based rarity plus rater judgement of fit
Reasoning quality	Soundness and justification of the solution path	Logical coherence, evidence, error handling
Conceptual transfer	Accuracy of explanation and success in applying the concept to a new case	Correctness and depth of explanation; novelty of the application

Raters will be trained on a calibration set, and they will score independently and without knowledge of condition, cohort, or whether the work is a pre-AI or current-era script. Because archival pre-AI scripts and current scripts may differ in format (for example handwritten versus typed), all responses will be transcribed into a uniform format so that surface cues do not reveal origin. Agreement will be estimated using intra-class correlation (ICC) for continuous scores and weighted kappa for ordinal judgements, and a disagreement-resolution procedure will be applied where raters differ by more than one scale point. Raters will also be rotated across batches to limit drift.

A pilot with a small sample from each educational level will precede the main study. Its purposes are to test the clarity of survey items through cognitive interviews, to check the time demands and difficulty of the tasks so that ceiling and floor effects are avoided, to estimate reliability and variance for the final power calculation, and to confirm that the tutor-configured assistant behaves as intended. Items or tasks that perform poorly will be revised or removed, and the pilot data will be excluded from the main analysis.

E. Experimental Procedure

In Component B, participants will be randomly assigned to one of three conditions while they work on a set of learning tasks in a topic appropriate to their level. In the unrestricted condition they have free access to a general-purpose GenAI assistant. In the tutor condition they use an assistant configured to give hints, ask guiding questions, and withhold complete solutions unless the student has attempted the task. In the control condition they work with conventional resources such as notes and textbooks. After the learning phase, all participants complete the unassisted tasks described above, ideally after a delay of several days to separate short-term performance from learning. Process data, including prompts submitted and time on task, will be recorded with consent so that the cognitive role of AI can be coded behaviourally and not only inferred from self-report.

F. Variables

TABLE IV
VARIABLES AND THEIR OPERATIONALISATION

Type	Variables	Operationalisation
Independent	Era or cohort; experimental condition; AI-use frequency; cognitive role of AI	Cohort year; assigned condition; self-reported and logged use; coded behaviour in prompts and survey
Dependent	Independent thinking; originality; creative problem-solving; conceptual understanding	Self-rated scales; rater-scored rubrics; statistical rarity; transfer task score
Moderators	Educational level; field of study; prior achievement; prior knowledge	Institutional records; pre-test
Mediators	Metacognition; perceived autonomy and competence; effort	Validated scales; time-on-task
Perception outcomes	Student-teacher agreement	Paired ratings of the same constructs

G. Data Analysis

Quantitative analysis will begin with descriptive statistics, reliability estimation, and checks of assumptions. Group differences (RQ1) will be tested using analysis of variance or, where covariates such as prior achievement are available, analysis of covariance (ANCOVA), with effect sizes and confidence intervals reported in preference to significance alone. Moderation (RQ4) will be tested through regression models with interaction terms for educational level, field of study, and use frequency. Because students are nested within classes and institutions, multilevel models will be used where ICC justifies them. Non-linear frequency effects (P7) will be examined using polynomial or spline terms. Student-teacher alignment (RQ3) will be assessed using correlation, difference scores, and agreement statistics such as Cohen's kappa for categorical judgements and ICC for continuous ones, with values interpreted against conventional benchmarks [30]. Mediation will be examined with bootstrapped indirect effects. The quantitative plan, including hypotheses and exclusion rules, will be registered before data collection to limit analytic flexibility.

Qualitative analysis will follow reflexive thematic analysis [31]. Transcripts will be coded inductively and then organised around the framework's roles and mechanisms. Two coders will code a subset independently to refine the codebook, and negative cases will be sought actively. Integration of the two strands will use joint displays that place quantitative results and qualitative themes side by side, so that convergence and divergence can be examined directly.

H. Preregistration, Data Management, and Open Science

To promote transparency, the hypotheses, sample-size rationale, exclusion criteria, and primary analyses will be registered before data collection. Deviations from the registered plan will be reported and justified, and exploratory analyses will be labelled as such. Instruments, rubrics, and task materials will be made available subject to copyright and the need to protect the integrity of tasks that may be reused. De-identified quantitative data and analysis code will be shared in a public repository where consent and institutional rules permit. Qualitative transcripts will be shared only in a

form that cannot identify participants, which may mean sharing coded excerpts and the codebook rather than full transcripts. A data-management plan will specify storage, retention, and deletion, and will be reviewed by the ethics committee together with the study protocol.

I. Ethical Considerations

The study will seek approval from the relevant institutional review boards before any data collection. Participation is voluntary, and participants may withdraw without penalty. Informed consent will be obtained, and parental consent will be sought for minors. Student work will be anonymised, stored securely, and used only for the stated research purposes. Because students are asked about behaviour that may breach course rules, the survey will be anonymous, and assurances will be given that responses will not affect grades or be reported to instructors. Experimental conditions will be designed so that no participant is academically disadvantaged; in particular, participants in the control and unrestricted conditions will receive equivalent access to the tutor tool and to feedback after the study. Teachers will be protected from identification in reporting. Finally, the AI systems used in the experiment will be evaluated for privacy and data-handling terms, and no personally identifying information will be entered into external services.

J. Validity and Limitations

Several threats to validity are acknowledged in advance. Cohort comparisons may confound era with other changes such as curriculum, pandemic-related disruption, and shifts in assessment style. The experiment is limited by duration and by the artificiality of laboratory-style tasks, and its results may not generalise to semester-long use. Self-report is subject to social desirability bias, particularly where AI use is stigmatised, and to memory bias in retrospective comparisons. Creativity is a contested construct, and no rubric captures all of its dimensions. GenAI tools also change rapidly, so findings are tied to the tools available at the time of the study. These limitations motivate the triangulated design, in which the weaknesses of one component are partly offset by the strengths of another, but they cannot be eliminated and will be discussed openly in reporting.

V. RESEARCH CONTRIBUTION

The paper makes contributions at four levels: theoretical, methodological, empirical-agenda, and practical.

A. Theoretical Contribution

The principal theoretical contribution is the Cognitive Role of AI framework. Existing accounts either treat AI as a uniform input, whose effect is positive or negative on average, or describe general risks such as overreliance without specifying when they arise. The framework replaces this with a more precise claim: the effect on learning and creativity depends on which cognitive operation the student performs and which the AI performs. By tying that claim to four mechanisms (practice, monitoring, motivation, and convergence), the framework connects previously separate literatures on cognitive offloading, desirable difficulties, scaffolding, self-determination, and creativity. It also reconciles apparently contradictory findings in the literature. Studies reporting benefits typically examine outcomes with

AI present, or scaffold-like use, while studies reporting harm typically examine unassisted outcomes after substitute-like use. Both can be correct, and the framework explains why.

A second theoretical contribution is the distinction between individual and collective originality. Existing creativity research evaluates products one at a time. In an AI-assisted environment in which many learners draw on the same generative models, the diversity of the population of ideas becomes an educational outcome in its own right. The convergence mechanism makes this explicit and generates testable predictions about between-student similarity.

B. Methodological Contribution

Methodologically, the paper contributes a design that confronts the problem that the pre-AI and AI-assisted eras cannot be compared within the same student. The triangulation of archival, experimental, and perceptual evidence provides a template that other researchers can adapt. Three specific features are worth highlighting. First, the cognitive role of AI is measured behaviourally, through prompt logs and behavioural survey items, rather than inferred from frequency of use. Second, assisted and unassisted performance are measured in the same participants, so that the performance-learning gap can be estimated directly. Third, student and teacher perceptions are collected with parallel instruments, allowing the accuracy of each to be examined rather than assumed.

C. Contribution to the Evidence Base

The planned study will extend the evidence beyond the settings in which it is currently concentrated. Most existing studies examine a single educational level, a single subject, or a short intervention. By sampling across levels and fields and testing moderation explicitly, the study will indicate where concern is warranted and where it is not. It will also add evidence on the divergence between students' and teachers' perceptions, which has direct implications for how institutions interpret survey-based findings and for how teachers should interpret their own impressions of student work.

D. Practical and Policy Contribution

For teachers and institutions, the framework offers a vocabulary and a diagnostic. Instead of asking whether a student has used AI, a teacher can ask what role it played. This shifts policy from blanket prohibition or unrestricted permission toward design of tasks and rules that favour scaffold and collaborator roles. It also supports concrete assessment decisions, such as including components that require unassisted demonstration of understanding. Section VI develops these implications into a model of responsible use.

E. Alignment of Questions, Propositions, and Methods

Table V shows how each research question maps onto the propositions, data sources, and analytic procedures. Presenting the alignment explicitly makes it possible to see that every question is addressed by at least two independent sources of evidence, and that every proposition has a corresponding test.

TABLE V
ALIGNMENT OF RESEARCH QUESTIONS, PROPOSITIONS, AND METHODS

RQ	Propositions	Evidence	Analysis
RQ1	P1, P2, P3	Archival cohort comparison; experiment	ANCOVA; between-student similarity indices
RQ2	P1, P8	Student and teacher surveys; interviews	Descriptive statistics; thematic analysis
RQ3	P4	Paired student-teacher ratings; task scores	Correlation; kappa and ICC; difference scores
RQ4	P5, P6, P7	All strands, stratified by level and field	Interaction terms; multilevel models; spline terms
RQ5	P8	Experiment; interviews; prompt-log coding	Mediation; joint displays

VI. TOWARD A MODEL OF RESPONSIBLE ARTIFICIAL INTELLIGENCE USE IN LEARNING

RQ5 asks how AI can support learning without weakening problem-solving and originality. The framework implies that responsible use is not a matter of limiting the amount of AI but of protecting the student's performance of the target cognitive operation. We propose five principles, each linked to a mechanism, and each to be tested in the planned study.

A. Five Principles

Attempt first. Students make a genuine independent attempt before consulting AI. This preserves productive struggle and the retrieval and generation processes that support learning. A simple practice is to record an initial approach, however incomplete, before any prompt is entered. This principle addresses the practice mechanism.

Ask for support, not for answers. Students prompt for hints, questions, counter-examples, and feedback on their own work rather than for finished products. Tutor-style configurations make this default, but students can also be taught to prompt in this way. This principle also addresses the practice mechanism and encourages the scaffold role.

Verify and justify. Students check AI output against course materials, independent sources, or their own reasoning, and explain in their own words why an output is right or wrong. Verification converts a passive reading of output into an active evaluation. This principle addresses the monitoring mechanism and is particularly important for novices, who need explicit instruction in how to evaluate output they cannot yet judge by intuition.

Own the idea. Students are expected to make and defend the key creative and conceptual decisions. Practices include generating several ideas independently before consulting AI, deliberately departing from the AI's first suggestion, and being able to explain the work orally. This principle addresses the motivation and convergence mechanisms.

Disclose and reflect. Students state how AI was used and reflect on what they learned from the process. Disclosure supports academic integrity, and reflection strengthens metacognition. This principle also gives teachers the information needed to distinguish roles.

B. Implications for Teachers and Institutions

Responsible use depends on conditions that individual students cannot supply alone. At the level of task design, teachers can build assignments that require process evidence, personal data or context, oral defence, or in-class unassisted components, so that the artefact alone is not sufficient evidence of learning. At the level of assessment, a balanced approach combines AI-permitted tasks, which assess the ability to work effectively with tools, with AI-free tasks, which assess the underlying capability. At the level of instruction, AI literacy should be taught explicitly, including the limitations of language models, the possibility of errors and fabricated content, and the principles above. At the level of policy, institutions should state clear and differentiated expectations by course and task, since a rule that is sensible for an introductory programming course may be inappropriate for a design studio. At the level of equity, institutions should ensure that access to well-configured tools does not depend on students' ability to pay, and that language and disability support uses of AI are recognised as legitimate.

C. Implications for Developers

The framework also has implications for the design of educational AI. Systems intended for learning can embed scaffolding by default, require an attempt before revealing solutions, fade support as competence grows, and surface uncertainty and sources so that students can monitor reliability. Features that give teachers visibility of how a tool was used would make role-based assessment practicable. The evidence on tutor-style configuration [15] suggests that such design choices matter.

D. Illustrations Across Disciplines

The five principles take different concrete forms in different subjects. The examples below are illustrative, and they are intended as hypotheses to be examined in the study and not as validated recommendations.

In an introductory programming course, the target operation is typically problem decomposition and debugging. A responsible pattern is for students to write and test their own first solution, then ask an assistant to explain an error message or suggest a test case, and finally to refactor their own code rather than replace it. Assessments might include a short unassisted coding exercise and a viva in which students explain a line of code they submitted.

In academic writing, the target operations are forming an argument and organising evidence. Students might draft a thesis and outline by hand, then ask for critique of the argument's weakest point, and record how they responded. Formative feedback is among the most effective instructional practices [32], and AI can extend its availability, provided students remain the authors of the revisions. In-class timed writing can provide an unassisted baseline.

In mathematics and the sciences, the target operations include setting up a problem, selecting a method, and checking plausibility. Tutor-style assistance that gives a hint after an attempt preserves this work. Retrieval practice, in which students recall and apply knowledge without prompts, is strongly supported by research on test-enhanced learning [33] and can be built into routines in which AI generates practice questions but students answer them unaided.

In design, art, and architecture, the target operations include exploring a wide space of possibilities and exercising judgement. A potential risk is convergence on a common visual vocabulary. A responsible pattern asks students to produce several divergent concepts before consulting any generative tool, to document how they transformed AI-generated material, and to defend their choices in critique. Here, the originality dimension of the rubric is particularly relevant.

In professional and clinical programmes, the target operations include judgement under uncertainty, and the consequences of uncritical acceptance of erroneous output are serious. Verification and source-checking should be treated as core competencies, and assessment should include scenarios in which the AI output is subtly wrong, so that students practise detecting errors.

E. An Implementation Roadmap

Institutions seeking to apply the model can work through three phases. In the diagnostic phase, programme teams identify the cognitive operations that each course is intended to develop, audit existing assignments to see which operations an AI could perform on a student's behalf, and gather baseline information on how students and staff currently use AI. This phase converts a general worry about AI into a specific list of vulnerable tasks.

In the design phase, teams redesign vulnerable tasks and set differentiated rules. Redesign might add process evidence, oral explanation, or personalised data, or move part of the assessment into supervised unassisted conditions. Rules should say clearly what is permitted, what is prohibited, and what must be disclosed for each task, and should be accompanied by instruction in AI literacy and in the five principles. Staff development is essential, because teachers need the confidence to design and to judge AI-era work.

In the evaluation phase, institutions monitor outcomes using a mixture of unassisted assessment results, student and staff feedback, and indicators of equity, and they revise policy in light of the evidence. Because the tools change quickly, review should be periodic and not one-off. The study described in this paper is intended to supply part of the evidence base for such reviews.

VII. DISCUSSION

A. Reconciling Optimism and Concern

A central aim of the paper is to show that the polarised public debate over GenAI in education rests partly on a confusion between two different questions. The first is whether AI helps students complete tasks. On the evidence reviewed, it plainly does. The second is whether AI helps students become more capable thinkers and creators. The answer to this second question is conditional, and is likely to depend on the cognitive role of the AI. Optimists are usually answering the first question and pessimists the second. The framework provides a way of holding both views at once, and suggests that the most productive educational research programme is one that identifies the conditions under which the benefits of AI as a tool and the aims of education as a developmental project are aligned.

B. Anticipated Challenges to the Framework

Several challenges should be anticipated. One is that the roles may be difficult to distinguish in practice. A student may move between roles within a single session, and a single prompt can be read as either a request for a hint or a request for an answer. The study addresses this by combining behavioural coding with self-report and by treating role as a continuum rather than a strict category. A second challenge is that the relevant cognitive operation is not always obvious. In a literature review assignment, the target operation might be searching, evaluating, or synthesising, and AI might appropriately take over the first while the student retains the others. Teachers need to specify what the task is meant to develop, which is itself a valuable discipline. A third challenge is that some skills may become less necessary as tools improve, just as long division became less central after calculators. The framework does not claim that all offloading is harmful. It claims that offloading harms learning when the offloaded operation is one that the learner still needs, and it leaves open the empirical and normative question of which skills those are.

C. Interpreting Patterns of Results

Because the study measures both assisted and unassisted performance, it can produce several distinct patterns, each with a different interpretation. If assisted performance is high and unassisted performance is no lower than that of non-users, AI is functioning as a productivity tool without apparent cost to capability, at least in the short term. If assisted performance is high but unassisted performance is lower, the performance-learning gap predicted by the offloading and desirable difficulties accounts is supported, and the next question is whether the gap is confined to substitute and shortcut roles. If unassisted performance is higher among tutor-condition participants than among control participants, the result would support the claim that well-designed AI can enhance learning and not just preserve it.

In the perception data, close agreement between students and teachers, and between each group and task performance, would suggest that surveys are a serviceable proxy for outcomes. Divergence would suggest that survey-based studies, which dominate the literature, may give a distorted picture, and would argue for greater reliance on performance measures. For originality, a pattern of higher average ratings but lower between-student diversity would support the convergence mechanism, while an absence of such convergence would suggest that students transform AI output sufficiently to maintain diversity. Null findings are informative in each case, since they would indicate where concern about AI is not warranted, and they will be reported with the same prominence as positive findings.

D. The Question of Originality

Originality deserves particular attention because it is both highly valued in education and difficult to measure. The convergence mechanism suggests that a student who relies on AI for ideas may produce competent but predictable work. However, it is equally plausible that AI can serve as a springboard, exposing students to possibilities they would not otherwise have considered and that they then transform. The difference may lie in what the student does after receiving the idea. A student who treats an AI suggestion as raw material

to be challenged, combined, and extended is exercising creative processes, whereas one who adopts it unchanged is not. Measuring both the product and the process, as this study intends to do, is therefore preferable to measuring either alone.

E. Directions for Future Research

Beyond the planned study, several directions are important. Longitudinal studies that follow students over multiple semesters would show whether short-term effects accumulate or fade. Studies of younger children would address developmental questions that the present design, focused on older learners, cannot. Comparative research across countries and education systems would examine how assessment culture, language, and access shape the effects. Finally, as AI systems become more capable of agentic, multi-step work, the boundary between assistance and delegation will shift, and the framework will need to be re-examined in light of new forms of use.

F. Ethical and Societal Considerations

Research and policy on this topic raise ethical questions that extend beyond individual learning. Equity is the first. Students differ in their access to capable tools, in their digital skills, and in the support they receive for using them well. If AI use is rewarded in some settings and penalised in others, existing inequalities may be amplified, and students who rely on AI for legitimate language or accessibility support should not be treated as cheating. Privacy is the second. Collecting prompt logs and process data is analytically valuable, but it exposes sensitive information about how students think and struggle, and it must be governed by consent, minimisation, and secure storage.

Surveillance and detection raise a third set of concerns. Automated detectors of AI-generated text are error-prone, and wrongly accusing a student can have serious consequences, particularly for multilingual writers. Institutions are therefore better advised to invest in assessment design and in dialogue with students than in detection alone. Finally, there is a broader question about which human capabilities a society wishes to preserve, as tools take over more cognitive work. The framework does not settle that question, but it provides a way of asking it precisely, by identifying which operations a given course is intended to develop and by protecting the opportunity to practise them.

VIII. CONCLUSION

Generative AI has made it possible to produce acceptable academic work with little personal cognitive effort. This capability does not in itself determine whether learning is helped or harmed. The argument developed in this paper is that the decisive factor is the cognitive role the AI plays: whether it substitutes for the student's thinking or scaffolds it. Drawing on cognitive, motivational, and creativity theory, the paper has set out the research foundation for comparing student learning and creativity before and during the AI era, identified the gaps that current evidence leaves open, and proposed a framework, a set of propositions, and a mixed-methods design capable of testing them. Its contributions are a role-based theory of AI in learning, a triangulated approach to the comparison problem, an agenda for examining

moderation and perception gaps, and a practice model of responsible use.

The educational stakes are considerable. If AI is allowed to quietly replace the effortful practice through which students develop judgement, problem-solving ability, and original thought, education may produce impressive artefacts and weaker people. If it is used to extend access to feedback and explanation while preserving that practice, it may help close long-standing inequalities. Which of these futures is realised will depend on choices made by students, teachers, institutions, and developers, and those choices should be informed by evidence rather than by enthusiasm or alarm. The study outlined here is intended to provide some of that evidence.

DECLARATIONS

Data availability. This paper is a conceptual and protocol contribution and reports no primary data. All empirical claims attributed to other studies should be checked against the original sources before submission or citation.

Use of AI assistance. This manuscript was drafted with the assistance of a generative AI system. The authors are responsible for reviewing, verifying, and revising all content, confirming every reference against its original source, and complying with the disclosure requirements of the institution or journal to which it is submitted.

REFERENCES

- [1] N. C. Soderstrom and R. A. Bjork, "Learning versus performance: An integrative review," *Perspect. Psychol. Sci.*, vol. 10, no. 2, pp. 176–199, 2015.
- [2] J. Kruger and D. Dunning, "Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments," *J. Pers. Soc. Psychol.*, vol. 77, no. 6, pp. 1121–1134, 1999.
- [3] E. F. Risko and S. J. Gilbert, "Cognitive offloading," *Trends Cogn. Sci.*, vol. 20, no. 9, pp. 676–688, 2016.
- [4] B. Sparrow, J. Liu, and D. M. Wegner, "Google effects on memory: Cognitive consequences of having information at our fingertips," *Science*, vol. 333, no. 6043, pp. 776–778, 2011.
- [5] E. L. Bjork and R. A. Bjork, "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning," in *Psychology and the Real World*, M. A. Gernsbacher et al., Eds. New York, NY, USA: Worth, 2011, pp. 56–64.
- [6] M. Kapur, "Productive failure," *Cogn. Instr.*, vol. 26, no. 3, pp. 379–424, 2008.
- [7] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA, USA: Harvard Univ. Press, 1978.
- [8] E. L. Deci and R. M. Ryan, "The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior," *Psychol. Inq.*, vol. 11, no. 4, pp. 227–268, 2000.
- [9] T. M. Amabile, "The social psychology of creativity: A componential conceptualization," *J. Pers. Soc. Psychol.*, vol. 45, no. 2, pp. 357–376, 1983.
- [10] T. M. Amabile, *Creativity in Context*. Boulder, CO, USA: Westview Press, 1996.
- [11] B. J. Zimmerman, "Becoming a self-regulated learner: An overview," *Theory Into Pract.*, vol. 41, no. 2, pp. 64–70, 2002.
- [12] Y. Fan et al., "Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance," *Br. J. Educ. Technol.*, 2024.
- [13] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, no. 6654, pp. 187–192, 2023.

- [14] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individ. Differ.*, vol. 103, Art. no. 102274, 2023.
- [15] H. Bastani et al., "Generative AI without guardrails can harm learning: Evidence from high school mathematics," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 122, no. 26, Art. no. e2422633122, 2025.
- [16] N. Kosmyna et al., "Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task," *arXiv preprint*, 2025.
- [17] H.-P. Lee et al., "The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2025.
- [18] M. Gerlich, "AI tools in society: Impacts on cognitive offloading and the future of critical thinking," *Societies*, vol. 15, no. 1, p. 6, 2025.
- [19] M. Urban et al., "ChatGPT improves creative problem-solving performance in university students: An experimental study," *Comput. Educ.*, vol. 215, Art. no. 105031, 2024.
- [20] A. R. Doshi and O. P. Hauser, "Generative AI enhances individual creativity but reduces the collective diversity of novel content," *Sci. Adv.*, vol. 10, no. 28, Art. no. eadn5290, 2024.
- [21] S. Kalyuga, P. Ayres, P. Chandler, and J. Sweller, "The expertise reversal effect," *Educ. Psychol.*, vol. 38, no. 1, pp. 23–31, 2003.
- [22] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cogn. Sci.*, vol. 12, no. 2, pp. 257–285, 1988.
- [23] M. T. H. Chi and R. Wylie, "The ICAP framework: Linking cognitive engagement to active learning outcomes," *Educ. Psychol.*, vol. 49, no. 4, pp. 219–243, 2014.
- [24] J. W. Creswell and V. L. Plano Clark, *Designing and Conducting Mixed Methods Research*, 3rd ed. Thousand Oaks, CA, USA: SAGE, 2018.
- [25] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum, 1988.
- [26] P. Tierney and S. M. Farmer, "Creative self-efficacy: Its potential antecedents and relationship to creative performance," *Acad. Manage. J.*, vol. 45, no. 6, pp. 1137–1148, 2002.
- [27] J. P. Guilford, "Creativity," *Amer. Psychol.*, vol. 5, no. 9, pp. 444–454, 1950.
- [28] E. P. Torrance, *Torrance Tests of Creative Thinking: Norms-Technical Manual*. Princeton, NJ, USA: Personnel Press, 1966.
- [29] T. M. Amabile, "Social psychology of creativity: A consensual assessment technique," *J. Pers. Soc. Psychol.*, vol. 43, no. 5, pp. 997–1013, 1982.
- [30] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [31] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, 2006.
- [32] P. Black and D. Wiliam, "Assessment and classroom learning," *Assess. Educ.: Princ., Policy Pract.*, vol. 5, no. 1, pp. 7–74, 1998.
- [33] H. L. Roediger III and J. D. Karpicke, "Test-enhanced learning: Taking memory tests improves long-term retention," *Psychol. Sci.*, vol. 17, no. 3, pp. 249–255, 2006.