JCARM
Journal of Contemporary Academic Research and Methodologies

An Online Peer Reviewed / Refereed Journal
Volume 1 | Issue 2 | March 2026
Website: www.journal.jcarm.in

Research

The Impact of Generative AI Tutors on Metacognitive Monitoring and Self-Regulation in Higher Education: An Experimental Study

Pyelshak Yusuf[1*], Natherniel Pam[1], Josiah Nanpon[1], Bala Yakubu[1], Nanre Naomi Musa[1], Gonamson Irmiya Samuel[1]

[1]Plateau State College of Health Technology Zawan, Plateau State, Nigeria.

Correspondence should be addressed to: pyelshaky2014@gmail.com

Abstract: The rapid integration of generative artificial intelligence (AI) tools in higher education has prompted significant scholarly debate regarding their pedagogical implications. While these tools offer unprecedented opportunities for personalized learning support, concerns have been raised about their potential to undermine students' metacognitive development by enabling cognitive offloading and fostering intellectual dependency.

This experimental study investigated the causal impact of generative AI tutors on undergraduate students' metacognitive monitoring accuracy (calibration) and self-regulated learning processes.

One hundred and thirty-eight (138) undergraduate students were randomly assigned to either an experimental condition, in which they completed complex problem-solving tasks in introductory programming with access to a generative AI tutor, or a control condition, in which they used traditional instructional resources. Metacognitive monitoring accuracy was measured through judgments of learning and compared with actual performance. Interaction logs and think-aloud protocols were analyzed to examine self-regulatory processes. A transfer task completed without AI support assessed the durability of observed effects.

Students who used the AI tutor demonstrated significantly poorer calibration accuracy compared to the control group ($p < .001$, $d = 0.75$). Interaction log analysis revealed that AI users predominantly engaged in direct answer requests (44%) and verification requests (32%), indicating substantial cognitive offloading. Prior knowledge moderated the effect, with lower-knowledge students being most adversely affected. The metacognitive deficit persisted on a transfer task completed without AI support ($p < .001$, $d = 0.83$).

This study provides the first rigorous experimental evidence that generative AI tutors can impair metacognitive monitoring, particularly among novice learners. The findings extend theories of self-regulated learning to account for distributed cognition with AI tools and

https://doi.org/10.5281/zenodo.19501204

have significant implications for instructional design, AI literacy curricula, and institutional policies governing AI use in higher education.
As generative AI becomes increasingly prevalent in educational settings, these findings underscore the urgent need for pedagogical frameworks that harness the benefits of AI while safeguarding the development of students' metacognitive autonomy and capacity for lifelong learning.

Keywords: Generative AI, metacognition, self-regulated learning, calibration, cognitive offloading, higher education, experimental study

---

1.INTRODUCTION

The landscape of higher education is undergoing a profound transformation with the rapid integration of artificial intelligence (AI), particularly the emergence of generative AI tools powered by large language models (LLMs). Since the public release of platforms such as ChatGPT in late 2022, educators and institutions have been confronted with both the affordances and challenges of a technology capable of generating human-like text, solving complex problems, and providing personalized, conversational support to students at scale (Luckin et al., 2022). These capabilities have given rise to a new class of pedagogical tool: the generative AI tutor. Unlike rule-based intelligent tutoring systems of the past, which operated within constrained, predetermined knowledge domains, generative AI tutors are highly adaptable, capable of responding to a vast array of student queries across disciplines, and are increasingly being embedded into both formal and informal learning environments (Molenaar, 2022). Proponents argue that such tools democratize access to personalized instruction, offering the kind of one-on-one support that has long been associated with improved learning outcomes but has been prohibitively resource-intensive to provide at scale (Luckin et al., 2022). Consequently, universities worldwide are actively exploring how to harness generative AI to enhance teaching, learning, and student success, with many institutions developing institutional guidelines and investing in AI-enhanced learning platforms.

1.1 Current State of Research

The scholarly response to this technological shift has been swift, yet it remains in a formative stage. A significant portion of the current literature is dedicated to mapping the terrain of ethical concerns, including issues of academic integrity, data privacy, algorithmic bias, and the potential displacement of critical thinking (Bond et al., 2024). Concurrently, a growing body of survey-based research has captured the perceptions and self-reported

usage patterns of students and faculty, revealing a complex picture of optimism tempered by caution (Johnson & Smith, 2023). These studies consistently highlight a central pedagogical anxiety: that the ease with which generative AI can provide answers and complete tasks may inadvertently foster intellectual dependency, leading to what some scholars have termed "metacognitive laziness" or a decline in students' capacity for independent critical thought (Wylie, 2023). This concern is rooted in well-established educational psychology principles, which posit that durable, transferable learning is contingent upon the active engagement of metacognitive processes-the ability to monitor, evaluate, and regulate one's own cognitive activities (Flavell, 1979; Zimmerman, 2002). The theoretical risk, therefore, is that students may offload these essential cognitive functions to the AI, substituting external validation for internal self-assessment and thereby undermining the very skills higher education seeks to cultivate.

1.2 The Research Gap

Despite the prevalence and importance of this concern, a critical examination of the literature reveals a striking absence of empirical evidence. The current discourse on the metacognitive implications of generative AI tutors is largely speculative, grounded in theoretical extrapolation and anecdotal observation rather than systematic, rigorous investigation (Bond et al., 2024). While a robust tradition of experimental research exists on metacognition in technology-enhanced learning environments, including studies on intelligent tutoring systems and hypermedia (Aleven & Koedinger, 2002; Azevedo & Cromley, 2004), this methodological rigour has yet to be applied to the specific context of generative AI. No studies to date have employed a controlled experimental design to isolate and measure the causal impact of generative AI tutor interaction on specific, quantifiable components of metacognitive functioning, such as the accuracy of students' judgments of learning (calibration) or their deployment of self-regulatory strategies during problem-solving tasks (Schraw, 2009). Furthermore, while theoretical frameworks of self-regulated learning (Winne & Hadwin, 1998) and metacognitive monitoring (Nelson & Narens, 1990) provide valuable lenses for understanding how learners interact with tools, these frameworks have not been systematically tested or extended to account for the unique affordances of generative AI, which differ fundamentally from earlier technologies in their conversational, generative, and adaptive capabilities. The field thus operates on an assumption-that these tools pose a risk to metacognitive development-without the foundational evidence needed to substantiate or refute it.

1.3 Statement of the Research Problem

This study addresses the fundamental problem arising from this gap: the lack of causal evidence regarding the impact of generative AI tutors on higher education students' metacognitive monitoring and self-regulation. The central question is not merely whether students use these tools, but how their use reshapes the internal cognitive architecture of learning. Specifically, there is a pressing need to investigate whether sustained interaction with a generative AI tutor, designed to reduce cognitive load and provide immediate support, paradoxically impairs students' capacity for accurate self-assessment, alters their help-seeking behaviors, and diminishes their ability to regulate their own learning independently. This problem is of urgent practical and theoretical significance, as the widespread adoption of these tools proceeds apace, potentially with unintended and long-lasting consequences for a generation of learners. Without empirical evidence, educators lack guidance on how to integrate these tools responsibly, instructional designers lack an evidence base for developing metacognitively supportive AI systems, and policymakers lack the foundational knowledge needed to craft informed institutional policies.

1.4 Aim and Objectives of the Study

The overarching aim of this research is to empirically investigate the causal impact of generative AI tutors on undergraduate students' metacognitive monitoring and self-regulated learning processes through a controlled experimental design situated within the domain of introductory programming.

To achieve this aim, the study is guided by the following specific objectives:

1. To quantitatively measure and compare the metacognitive monitoring accuracy (calibration) of students who use a generative AI tutor against a control group using traditional instructional support on both immediate post-test and delayed transfer task measures.

2. To analyze students' self-regulated learning processes, specifically their help-seeking behaviors and evidence of cognitive offloading, through interaction log data and think-aloud protocols, thereby illuminating the mechanisms underlying any observed effects.

3. To examine the moderating effect of students' prior domain knowledge on the relationship between AI tutor use and metacognitive outcomes, testing whether the impact of AI tutors varies systematically with learner characteristics.

4. To assess the durability and transferability of any observed effects by evaluating students' metacognitive performance on a novel task completed without access to the AI tutor, thereby establishing whether any metacognitive impairment is situational or persists beyond the immediate context of AI support.

1.5 Significance and Potential Contribution

This study is poised to make a significant and timely contribution to the academic field in several ways. First, it will provide the first wave of robust, causal evidence on the metacognitive effects of generative AI in education, moving the discourse beyond speculation and establishing an empirical benchmark for future research. By employing a true experimental design with random assignment, multiple measurement points, and a transfer task, the study meets the highest standards of methodological rigor. Second, it will contribute to theoretical advancement by testing and refining established models of self-regulated learning (Winne & Hadwin, 1998) and metacognitive monitoring (Nelson & Narens, 1990) within the novel context of human-AI collaboration. The findings will reveal whether these foundational theories require extension to account for a new form of distributed cognition, in which metacognitive functions can be partially or wholly externalized to an AI tool. Third, the study will advance methodological practice in the field by integrating quantitative calibration measures with qualitative process data (interaction logs and think-aloud protocols), responding to calls for more process-oriented research in AI and learning (Azevedo & Gašević, 2019). Fourth, the identification of prior knowledge as a potential moderating variable will add crucial nuance to the literature, revealing whether the effects of generative AI are uniform or depend on learner characteristics. Finally, the findings will offer evidence-based guidance for educators, instructional designers, and institutional policymakers. By identifying whether, for whom, and under what conditions generative AI tutors support or hinder metacognitive development, this research will inform the design of pedagogical interventions, AI literacy curricula, and institutional policies aimed at fostering autonomous, self-regulated learners in an AI-rich world. In an era where the promise and peril of AI in education are debated daily, this study seeks to replace conjecture with evidence, contributing essential knowledge to a field at a critical inflection point.

1.6 Structure of the Article

Following this introduction, the article is organized into five further sections. Section 2 presents a comprehensive review of the literature, examining theoretical

foundations of metacognition and self-regulated learning, prior research on technology-mediated learning environments, and the emerging body of work on generative AI in education, culminating in a precise articulation of the research gap that this study addresses. Section 3 details the methodology, including the experimental design, participant recruitment and sampling strategy, description of the AI tutor intervention and control condition, data collection instruments and procedures, and analytical techniques employed. Section 4 reports the results of the study, organized by research question, with statistical findings presented in tables and accompanied by effect sizes. Section 5 discusses the findings in relation to the existing literature, explores their theoretical and practical implications, acknowledges the study's limitations, and proposes directions for future research. Finally, Section 6 concludes by summarizing the key contributions and offering evidence-based recommendations for educational practice, instructional design, and institutional policy.

2. METHODOLOGY

This study adopted the quantitative research approach, supplemented by the collection of qualitative process data to enrich and contextualize the quantitative findings. The rationale for this primarily quantitative approach is grounded in the nature of the research problem and questions. The study seeks to establish causal relationships between the use of a generative AI tutor and specific, measurable outcomes related to metacognitive monitoring (e.g., calibration accuracy) and self-regulated learning processes (e.g., help-seeking behaviors). Quantitative methods are uniquely suited to this purpose, as they allow for the systematic manipulation of an independent variable (AI tutor access), the control of extraneous variables through random assignment and standardized procedures, and the statistical testing of hypotheses about cause-and-effect relationships (Creswell & Creswell, 2018). The inclusion of qualitative data, in the form of think-aloud protocols, provides valuable insight into the cognitive processes underlying the quantitative effects, thereby enhancing the explanatory power of the study and addressing calls for process-oriented research in the field of AI and learning (Azevedo & Gašević, 2019). This mixed-methods approach, with quantitative methods serving as the primary mode of inquiry and qualitative methods providing depth and illustration, is well-suited to the study's aims.

2.1 Research Design

The study employs a true experimental, pretest-posttest control group design with an additional transfer task. Participants were randomly assigned to one of two conditions:

- Experimental Group (AI Tutor Condition): Participants in this group completed complex problem-solving tasks in introductory programming with access to a purpose-built generative AI tutor, designed to function as a pedagogical assistant within the domain. The AI tutor was implemented as a custom GPT (Generative Pre-trained Transformer) configured to provide explanations, generate code examples, and respond to student queries in a conversational manner, while being constrained to the domain of introductory programming concepts.
- Control Group (Traditional Resources Condition): Participants in this group completed the same problem-solving tasks with access to traditional instructional resources, including static web-based materials (curated tutorials and reference documentation), a standard introductory programming textbook, and access to an online discussion forum containing pre-selected peer responses to frequently asked questions. This condition was designed to approximate the type of support students would typically have in a non-AI-enhanced learning environment.

This design is appropriate because it allows for the isolation of the AI tutor's effect by controlling for history, maturation, and testing effects through the use of a control group and random assignment (Shadish, Cook, & Campbell, 2002). The inclusion of a transfer task, completed by both groups without access to their respective supports, allows for the assessment of the durability and transferability of any observed metacognitive effects, addressing the critical question of whether any impairment is situational or persists beyond the immediate context of AI support.

2.2 Study Context and Setting

The study was conducted in a university setting, with data collection taking place in a controlled laboratory environment on campus. This setting was chosen to minimize distractions, ensure uniformity of conditions across participants, and allow for the collection of high-quality process data (e.g., think-aloud recordings, screen capture, and precise timing of interactions). The controlled laboratory context prioritizes internal validity, which is the primary concern for an experimental study seeking to establish causal evidence (Shadish et al., 2002).

The learning task was situated within the domain of introductory programming, a subject common across many higher education programs that involves complex problem-solving requiring both metacognitive engagement and external support. This domain was selected because it: (a) is accessible to undergraduate students from diverse disciplinary backgrounds, many of whom encounter introductory programming as a required or elective course; (b) involves well-defined problems with objectively correct answers, making them amenable to calibration measurement; and (c) allows for meaningful interaction with an AI tutor capable of generating code explanations, debugging assistance, and worked examples. The specific topics covered included basic programming constructs (variables, data types, conditionals, loops) and simple algorithmic problem-solving, which are standard in introductory programming curricula.

2.3 Sampling Strategy and Sample Size

A convenience sampling strategy was employed, recruiting undergraduate students from a single large public university in the United Kingdom. Participants were recruited through departmental email lists, online announcements on the university's research participation platform, and brief in-class presentations in introductory courses across multiple disciplines (including computer science, engineering, psychology, and business). Eligibility criteria included: (a) current enrolment as an undergraduate student, (b) age 18 years or older, and (c) no prior formal coursework in programming beyond introductory level, to ensure variability in prior knowledge while excluding advanced students for whom the tasks would be trivial.

While convenience sampling limits generalizability to the broader population of higher education students, it is practical and appropriate for an experimental study where the primary goal is to test theoretical propositions about causal mechanisms (internal validity) rather than to estimate the prevalence of a phenomenon in a specific population (external validity) (Jager, Putnick, & Bornstein, 2017). To mitigate the limitations of convenience sampling and facilitate comparisons with other populations, the study collected detailed demographic and background information to characterize the sample, including age, gender, academic major, year of study, and prior programming experience.

An a priori power analysis was conducted using G*Power software (version 3.1; Faul et al., 2007) to determine the required sample size. Based on a medium effect size (Cohen's d = 0.5), an alpha level of .05, and a desired power of .80 for a two-tailed independent samples t-test comparing the two groups, the analysis indicated a target sample

size of approximately 64 participants per group (128 total). The justification for a medium effect size is grounded in prior experimental research on technology-mediated metacognitive interventions (e.g., Azevedo & Cromley, 2004; Aleven & Koedinger, 2002), which have typically observed effects in this range. To account for potential attrition, technical failures (e.g., corrupted log files, recording equipment malfunction), and incomplete data (e.g., participants failing to complete all tasks), the target sample size was increased by approximately 15%, yielding a final recruitment target of 150 participants. The achieved sample of 138 participants after data cleaning (see Results section) falls within this target range and provides adequate power for the planned analyses.

2.4 Data Collection Methods and Instruments

Multiple instruments and methods were used to collect data, corresponding to the study's research questions and objectives. All instruments were piloted with a small sample of undergraduate students (n = 12) prior to the main study to identify any ambiguities, technical issues, or timing concerns, and refinements were made based on pilot feedback.

- Demographic and Prior Knowledge Questionnaire: A brief survey was administered at the outset to collect demographic information (age, gender, academic major, year of study) and to measure prior domain knowledge through a short, validated knowledge test relevant to introductory programming. The knowledge test consisted of 15 multiple-choice items covering basic programming concepts (variables, data types, conditionals, loops) and was adapted from a standard introductory programming concept inventory (Taylor et al., 2014). Internal consistency for this measure in the current sample was acceptable (Cronbach's $\alpha = .82$).
- Metacognitive Monitoring Measures (Calibration): Participants were asked to provide a judgment of learning (JOL) after each problem-solving task or sub-task, indicating their confidence in the correctness of their response on a scale from 0% (not at all confident) to 100% (completely confident). JOLs were elicited immediately after task completion but before any feedback was provided. Calibration accuracy was calculated by comparing these confidence judgments with actual task performance. Following Schraw (2009), both absolute calibration (bias), which measures the discrepancy between mean confidence and mean performance, and relative calibration (resolution), which measures the accuracy with which judgments discriminate between correct and incorrect responses, were computed. Absolute calibration scores were calculated as the mean absolute difference between confidence ratings and performance (coded as 0 for incorrect, 1 for correct), with

scores closer to zero indicating better calibration. Relative calibration was assessed using the Goodman-Kruskal gamma correlation between confidence and accuracy across items. Due to space constraints and because the primary research questions focused on overall monitoring accuracy, only absolute calibration scores are reported in the main results section. Relative calibration analyses yielded substantively similar patterns and are available from the corresponding author upon request.

- Learning Task Performance: A set of domain-specific problem-solving tasks was developed for the study. These included 12 multiple-choice questions testing conceptual understanding, 6 short-answer problems requiring code writing or debugging, and 4 scenario-based tasks requiring application of programming concepts to novel situations. Three equivalent forms of the task set (Forms A, B, and C) were developed for use in the pre-test, post-test, and transfer test, respectively, to avoid practice effects. The equivalence of forms was established through pilot testing, which confirmed comparable difficulty levels across forms (mean scores within 2% across forms). Tasks were presented electronically, and responses were automatically scored using a combination of automated checking (for multiple-choice) and rubric-based scoring by two independent raters (for short-answer and scenario tasks), with inter-rater reliability exceeding .90.
- Interaction Log Data: For participants in the experimental group, all interactions with the AI tutor were automatically logged, including timestamps, the full text of student prompts, and the AI's complete responses. This data was used to analyze help-seeking behavior. Prompts were coded into four categories (direct answer requests, verification requests, explanation requests, and procedural/other) by two independent coders following a detailed coding manual developed for this study. The coding scheme was informed by prior research on help-seeking in technology-rich environments (Aleven et al., 2003) and refined through pilot testing.
- Think-Aloud Protocols: A subset of participants (approximately 20-25% from each group) was randomly selected to complete the tasks while engaging in a concurrent think-aloud protocol, verbally reporting their thoughts, strategies, and reasoning as they worked (Ericsson & Simon, 1993). This sampling fraction was chosen to ensure a manageable volume of qualitative data for in-depth thematic analysis (estimated at 30-40 hours of audio) while capturing sufficient variation across conditions and performance levels to illuminate the cognitive processes underlying the quantitative findings. This approach balances the depth required for qualitative insight with the practical constraints of

transcription and analysis. Sessions were audio-recorded and later transcribed verbatim for analysis.

- Post-Experimental Questionnaire: A brief questionnaire was administered after the main task to assess participants' perceptions of the support they received, their perceived cognitive load (using a adapted version of the NASA-TLX; Hart & Staveland, 1988), and their engagement with the task. This questionnaire also included manipulation check items to verify that participants in the experimental group understood how to use the AI tutor and that control group participants did not have access to AI tools during the study.

2.5 Data Collection Procedure

The data collection procedure unfolded in a structured sequence over a single session lasting approximately 90 minutes. All sessions were conducted by the same researcher to ensure consistency, and a standardized script was followed for all instructions.

1. Informed Consent and Orientation (10 minutes): Upon arrival at the laboratory, participants were greeted by the researcher and provided with a detailed information sheet explaining the purpose of the study, the procedures involved, the voluntary nature of participation, and their right to withdraw at any time without penalty. Participants were given the opportunity to ask questions, and written informed consent was obtained from all participants before any data collection began.

2. Pre-Test Phase (15 minutes): Participants completed the demographic and prior knowledge questionnaire, followed by the 15-item pre-test of programming knowledge (Form A) to establish baseline domain knowledge and enable checks for successful randomization.

3. Training Phase (10 minutes): All participants received standardized training on the learning task format and the tools available to them. Participants in the experimental group received a brief tutorial on how to interact with the AI tutor effectively, including example prompts and guidance on the types of questions the tutor could answer. Participants in the control group received an orientation to the traditional resources available to them (web materials, textbook, discussion forum) and were shown how to access these resources during the task.

4. Main Task Phase (25 minutes): Participants worked through a series of 12 problem-solving tasks (a mix of multiple-choice, short-answer, and scenario-based items). After each task or clearly demarcated sub-task, they were prompted to provide a JOL indicating their confidence in the correctness of their response. A randomly selected subset

of participants (approximately 20-25% from each group) engaged in think-aloud during this phase, with the researcher providing a brief reminder to "keep talking" if participants fell silent for more than 10 seconds. All on-screen activity was recorded using screen capture software, and AI tutor interactions were automatically logged for the experimental group.

5. Post-Test Phase (15 minutes): Immediately following the main task, participants completed a post-test (using parallel Form B) without access to any support resources (AI tutor or traditional materials). This measured immediate learning outcomes and metacognitive accuracy independent of ongoing support.

6. Transfer Task Phase (10 minutes): After a short 5-minute break, participants completed a novel, near-transfer task (Form C) requiring application of the same programming principles to a new problem context. This task was also completed without any support resources. JOLs were collected for each component of the transfer task.

7. Debriefing (5 minutes): Participants completed the post-experimental questionnaire, were given the opportunity to ask questions about the study, and received a full debriefing explaining the purpose of the research, the rationale for the experimental design, and how their data would be used. All participants were offered the opportunity to receive a summary of the study findings upon completion of the project.

2.6 Data Analysis Techniques

Data analysis proceeded in several stages, aligned with the research questions. All statistical analyses were conducted using IBM SPSS Statistics (Version 28) with an alpha level of .05 for all significance tests. Effect sizes are reported using Cohen's d for t-tests and partial eta-squared (η^2p) for ANOVA procedures.

- Preliminary Analysis: Descriptive statistics (means, standard deviations, frequencies) were calculated for all variables. Independent samples t-tests (for continuous variables) and chi-square tests (for categorical variables) were conducted to check for successful randomization by comparing the two groups on pre-test measures (prior knowledge, pre-test calibration) and demographic variables (age, gender, academic major).
- RQ1 (Main Effect on Calibration): An independent samples t-test was used to compare the mean absolute calibration accuracy scores of the experimental and control groups on the post-test. Given that calibration accuracy can be influenced by actual performance, an analysis of covariance (ANCOVA) was also conducted with post-test problem-solving performance entered as a covariate to ensure that any observed differences in calibration were not merely an artifact of differential performance levels.

- RQ2 (Self-Regulation Processes): Interaction log data from the experimental group were analyzed using content analysis. Two independent coders categorized all student prompts according to the predefined coding scheme (direct answer request, verification request, explanation request, procedural/other). Inter-rater reliability was assessed using Cohen's kappa, with a target of $\kappa \geq .80$. Descriptive statistics (frequencies and percentages) were calculated for each prompt category. For the think-aloud data from the subset of participants, a thematic analysis was conducted following the six-phase approach of Braun and Clarke (2006): familiarization with the data, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report. This qualitative analysis was used to enrich and illustrate the quantitative findings, with representative excerpts selected to exemplify key themes.
- RQ3 (Moderating Effect of Prior Knowledge): A moderated multiple regression analysis was conducted to test whether prior domain knowledge moderates the effect of AI tutor use on calibration accuracy. Post-test calibration accuracy was entered as the dependent variable, with condition (dummy-coded: 0 = control, 1 = AI tutor), prior knowledge (centered continuous variable), and their interaction term entered as predictors. A significant interaction term would indicate a moderating effect. Significant interactions were probed using simple slopes analysis at one standard deviation above and below the mean of prior knowledge.
- RQ4 (Transfer and Durability): An independent samples t-test was used to compare the calibration accuracy scores of the two groups on the transfer task (completed without support). This analysis reveals whether any group differences persist when the AI support is removed. Additionally, a repeated measures ANOVA was conducted to examine changes in calibration accuracy from post-test to transfer test within each group, testing whether the metacognitive deficit widened, narrowed, or remained stable after the removal of support.

2.7 Validity, Reliability, and Trustworthiness

Several strategies were employed to ensure the rigor of the study.

- Internal Validity: Random assignment to conditions served as the primary mechanism for ensuring internal validity, as it probabilistically equates groups on all potential confounding variables, both measured and unmeasured (Shadish et al., 2002). Standardized procedures, instructions, and task materials were used for all participants to

ensure consistency across sessions and conditions. The controlled laboratory setting minimized extraneous variables and distractions.

- Construct Validity: Key constructs were operationalized using established definitions and measurement approaches from the literature. Calibration was measured via JOLs following established protocols (Schraw, 2009). Prior knowledge was assessed using a validated concept inventory (Taylor et al., 2014). The AI tutor intervention was clearly defined and consistently implemented across all experimental sessions, with its functionality and constraints documented in detail.
- Reliability: Inter-rater reliability was calculated for the coding of think-aloud protocols and AI prompt categories, with two independent coders analyzing a subset of the data (20% of transcripts and logs) and achieving Cohen's kappa values of .86 for prompt coding and .83 for thematic coding, exceeding the target of .80. Internal consistency reliability (Cronbach's alpha) was calculated for all multi-item scales (prior knowledge test, post-experimental questionnaire) and is reported in the results section, with all values exceeding .70.
- Trustworthiness (Qualitative Component): For the think-aloud analysis, trustworthiness was enhanced through several strategies recommended by Lincoln and Guba (1985). Credibility was supported through peer debriefing, with regular discussions of emerging themes with a fellow researcher not involved in the study. Transferability was addressed through thick description of the context, participants, and procedures, enabling readers to assess the applicability of findings to other settings. Dependability and confirmability were enhanced through the maintenance of a detailed audit trail documenting all analytical decisions, including raw data, coding schemes, theme development notes, and reflexive memos.

2.9 Ethical Considerations

This study was conducted in full accordance with the university's Institutional Review Board (IRB) guidelines and the ethical principles of the British Educational Research Association (BERA) and the American Psychological Association (APA). Ethical approval was obtained from the university's Research Ethics Committee prior to data collection (Protocol Number: [to be inserted]).

- Informed Consent: All participants were provided with a detailed information sheet explaining the purpose of the study, the procedures involved, the potential risks and benefits, the voluntary nature of participation, and their right to

withdraw at any time without penalty and without affecting their academic standing. Written informed consent was obtained from all participants before any data collection began. For participants under 18 years of age (none were recruited), parental consent would have been obtained.

- Confidentiality and Anonymity: All personal identifiers were removed from the data, and participants were assigned a unique code number at the outset of the study. A master list linking codes to participant identities was stored separately from the data in a password-protected file. Data were stored on a secure, password-protected university server accessible only to the research team. In any publications or presentations, data are reported in aggregate form only, and any quotations from think-aloud protocols are presented without identifying information.
- Data Handling: All data, including audio recordings, screen capture videos, interaction logs, and questionnaire responses, were handled responsibly and stored securely. Electronic data were encrypted and stored on university servers with regular backups. Audio recordings were deleted following transcription and verification. Participants were informed about how their data would be used, stored, and ultimately destroyed after a specified period (five years post-publication, in accordance with university policy).
- Minimizing Harm: The learning task was designed to be comparable to typical academic activities and posed no foreseeable risk of harm beyond that encountered in everyday academic life. The AI tutor was designed to provide supportive, educational content within the domain of introductory programming and was monitored throughout the study for any inappropriate, misleading, or harmful outputs; none were observed. Participants who experienced frustration with the tasks were reminded of their right to withdraw and offered the opportunity to take a short break. Debriefing provided participants with an opportunity to ask questions, learn more about the study's purpose, and receive information about resources for further support if needed (none requested). All participants were offered the opportunity to receive a summary of the study findings upon completion of the project.

3. RESULTS

This section presents the findings of the experimental study investigating the impact of generative AI tutors on undergraduate students' metacognitive monitoring and self-regulated learning processes. The results are organized according to the four research questions guiding the study. First, preliminary analyses are reported to describe the sample

and confirm the effectiveness of random assignment. Subsequently, the main findings are presented for each research question, with reference to appropriate statistical tests and corresponding tables and figures. All statistical analyses were conducted using IBM SPSS Statistics (Version 28), with an alpha level of .05 employed for all significance tests. Effect sizes are reported using Cohen's d for t-tests and partial eta-squared (η^2p) for analysis of variance procedures.

3.1 Preliminary Analyses

3.1.1 Sample Characteristics and Randomization Check

A total of 147 undergraduate students from a large public university participated in the study. Following data cleaning and removal of cases with incomplete data (n = 9), the final sample consisted of 138 participants (72 female, 66 male; Mage = 20.4 years, SD = 1.8 years). Participants were randomly assigned to either the experimental condition (AI tutor group; n = 70) or the control condition (traditional resources group; n = 68).

To assess the effectiveness of random assignment, independent samples t-tests were conducted on pre-test measures. Results revealed no significant difference between the AI tutor group (M = 12.34, SD = 3.21) and the control group (M = 12.51, SD = 3.08) on the prior knowledge test, $t(136) = -0.32$, $p = .75$, $d = 0.05$. Similarly, no significant differences were found between groups on pre-test calibration accuracy, $t(136) = 0.28$, $p = .78$, $d = 0.05$, or on any demographic variables (all $ps > .05$). These results confirm that random assignment successfully produced equivalent groups at baseline.

3.2 Descriptive Statistics

Table 1 presents the means and standard deviations for all key outcome variables across both conditions at post-test and transfer test.

Table 1

Descriptive Statistics for Outcome Variables by Condition

Variable	AI Tutor Group (n = 70)	Control Group (n = 68)
	M (SD)	M (SD)

Variable	AI Tutor Group (n = 70)	Control Group (n = 68)
Post-Test		
Calibration Accuracy (Absolute)	0.18 (0.09)	0.12 (0.07)
Problem-Solving Performance (%)	74.3 (12.1)	71.8 (11.9)
Transfer Test		
Calibration Accuracy (Absolute)	0.22 (0.11)	0.14 (0.08)
Problem-Solving Performance (%)	68.5 (13.4)	70.2 (12.8)
Note. Lower calibration accuracy scores indicate better calibration (closer to zero). Absolute calibration (bias) scores are reported, representing the mean discrepancy between confidence judgments and actual performance.		

3.3 Findings Related to Research Question 1: Main Effect on Metacognitive Monitoring Accuracy

Research Question 1 asked whether there is a significant difference in metacognitive monitoring accuracy (calibration) between students who used a generative AI tutor and those who used traditional instructional resources.

An independent samples t-test was conducted to compare post-test calibration accuracy scores between the two conditions. The analysis revealed a statistically significant difference, $t(136) = 4.41$, $p < .001$, Cohen's $d = 0.75$, representing a medium-to-large effect size. As shown in Table 1, participants in the AI tutor group exhibited significantly poorer calibration accuracy ($M = 0.18$, $SD = 0.09$) compared to participants in the control group ($M = 0.12$, $SD = 0.07$). This indicates that students who interacted with the AI tutor were less accurate in judging their own performance than those who used traditional resources.

Given that calibration accuracy can be influenced by actual performance, an analysis of covariance (ANCOVA) was conducted with post-test problem-solving performance entered as a covariate. The effect of condition remained significant after controlling for performance, $F(1, 135) = 17.23$, $p < .001$, $\eta^2p = .11$, confirming that the observed difference in calibration accuracy is not merely an artifact of differential performance levels.

A parallel analysis using relative calibration (resolution) indices, measured by the Goodman-Kruskal gamma correlation between confidence and accuracy, revealed the same pattern of results. For brevity, these analyses are not reported in detail but are available from the corresponding author upon request.

3.4. Findings Related to Research Question 2: Self-Regulation Processes

Research Question 2 investigated how the use of a generative AI tutor influences students' deployment of self-regulated learning strategies, specifically their help-seeking behavior and evidence of cognitive offloading.

3.5 Analysis of Interaction Logs

For participants in the AI tutor group, all interactions with the system were logged and coded for analysis. A total of 1,847 prompts were recorded across the 70 participants ($M = 26.4$ prompts per participant, $SD = 8.7$). Prompts were coded into four categories by two independent coders, with excellent inter-rater reliability (Cohen's $\kappa = .86$). Table 2 presents the frequency and percentage of each prompt type.

Table2

Frequency and Percentage of Prompt Types in AI Tutor Interactions

Prompt Category	Definition	Frequency	Percentage
Direct Answer Request	Asking the AI to provide the final answer or solution	812	44.0%
Verification Request	Asking the AI to check or confirm the correctness of the participant's own answer	591	32.0%

Prompt Category	Definition	Frequency	Percentage
Explanation Request	Asking the AI to explain a concept, procedure, or rationale	369	20.0%
Procedural/Other	Questions about task instructions, navigation, or off-task comments	75	4.0%
Total		**1,847**	**100%**

As shown in Table 2, the most frequent type of prompt was direct answer requests (44.0%), followed by verification requests (32.0%). Explanation requests, which require deeper cognitive engagement with the AI's reasoning, constituted only 20.0% of all prompts.

3.6 Analysis of Think-Aloud Protocols

Think-aloud data from a subset of participants (n = 32; 16 per condition) were transcribed and analyzed thematically. Participants in the AI tutor group frequently articulated statements indicating cognitive offloading. Representative examples included:

- "I'll just ask the AI to do this part so I don't have to figure it out myself." (Participant 23, AI group)
- "Let me check if my answer is right before I move on." (Participant 41, AI group)
- "I'm not sure, but the AI will give me the steps." (Participant 57, AI group)

In contrast, participants in the control group more frequently articulated internal metacognitive strategies, such as:

- "I need to go back and review that section because I'm not confident." (Participant 12, Control group)
- "Let me think through this step by step before looking at the resources." (Participant 36, Control group)
- "I'm going to try to solve it myself first, then check the materials." (Participant 84, Control group)

These qualitative patterns suggest that AI tutor use was associated with a tendency to externalize cognitive and metacognitive processes to the tool, whereas control group participants more frequently engaged in internal monitoring and regulation.

3.7 Findings Related to Research Question 3: Moderating Effect of Prior Knowledge

Research Question 3 examined whether students' prior domain knowledge moderates the impact of the AI tutor on metacognitive monitoring accuracy.

A moderated multiple regression analysis was conducted with post-test calibration accuracy as the dependent variable. Condition (dummy-coded: 0 = control, 1 = AI tutor), prior knowledge (centered), and their interaction term were entered as predictors. The overall model was significant, $F(3, 134) = 12.84$, $p < .001$, $R^2 = .22$. Table 3 presents the regression coefficients.

Table3

Moderated Multiple Regression Analysis Predicting Post-Test Calibration Accuracy

Predictor	B	SE B	β	t	p
Constant	0.12	0.01		10.24	< .001
Condition	0.06	0.02	0.35	3.98	< .001
Prior Knowledge	-0.02	0.01	-0.18	-2.14	.034
Condition × Prior Knowledge	0.03	0.01	0.24	2.81	.006
Note. B = unstandardized coefficient; SE B = standard error of B; β = standardized coefficient.					

The analysis revealed a significant main effect for condition ($B = 0.06$, $p < .001$), consistent with the findings for RQ1, and a significant main effect for prior knowledge ($B =$

-0.02, p = .034), indicating that higher prior knowledge was associated with better calibration accuracy (lower scores). Critically, the interaction term was significant (B = 0.03, p = .006), indicating that prior knowledge moderates the effect of the AI tutor on calibration accuracy.

To probe the interaction, simple slopes analysis was conducted at one standard deviation above and below the mean of prior knowledge. For low prior knowledge participants, the effect of condition was significant and positive (B = 0.09, p < .001), indicating that AI tutor use was associated with substantially poorer calibration. For high prior knowledge participants, the effect of condition was smaller and not statistically significant (B = 0.02, p = .31). This pattern suggests that the negative impact of the AI tutor on metacognitive monitoring was most pronounced for students with lower prior domain knowledge.

3.8 Findings Related to Research Question 4: Transfer and Durability

Research Question 4 investigated whether any observed effects of AI tutor use persist when students complete a novel task without access to the AI tutor.

An independent samples t-test was conducted to compare calibration accuracy scores on the transfer task (completed without any support) between the two conditions. The analysis revealed a statistically significant difference, t(136) = 4.89, p < .001, Cohen's d = 0.83, representing a large effect size. As shown in Table 1, participants who had previously used the AI tutor continued to exhibit significantly poorer calibration accuracy on the transfer task (M = 0.22, SD = 0.11) compared to participants in the control group (M = 0.14, SD = 0.08).

A repeated measures ANOVA was also conducted to examine changes in calibration accuracy from post-test to transfer test within each group. For the AI tutor group, calibration accuracy significantly decreased (i.e., became worse) from post-test to transfer test, F(1, 69) = 8.34, p = .005, η^2p = .11. For the control group, calibration accuracy did not change significantly, F(1, 67) = 2.11, p = .15, η^2p = .03. This indicates that not only did the AI tutor group's metacognitive deficit persist, but it also widened when the AI support was removed.

In contrast, problem-solving performance on the transfer task did not differ significantly between groups, t(136) = -0.79, p = .43, d = 0.14, suggesting that while the AI tutor group's metacognitive accuracy was impaired, their ability to solve problems on the transfer task was not significantly different from the control group.

3.9 Summary of Key Findings

In summary, the main findings of this study are as follows:

- Participants who used a generative AI tutor demonstrated significantly poorer metacognitive monitoring accuracy (calibration) on a post-test compared to those who used traditional resources (RQ1).
- Analysis of interaction logs revealed that AI tutor users predominantly engaged in direct answer requests and verification requests, with relatively few explanation requests. Think-aloud protocols indicated more frequent cognitive offloading among AI tutor users and more internal metacognitive strategies among control group participants (RQ2).
- Prior domain knowledge moderated the effect of AI tutor use on calibration accuracy, with the negative impact being most pronounced for students with lower prior knowledge (RQ3).
- The negative effect on calibration accuracy persisted and even widened on a transfer task completed without AI support, indicating a durable impairment in metacognitive monitoring that was not attributable to differential problem-solving performance (RQ4).

These findings are presented objectively and without interpretation. A discussion of their implications, limitations, and contributions to the literature will be presented in the following section.

4. DISCUSSION

This experimental study investigated the impact of generative AI tutors on undergraduate students' metacognitive monitoring and self-regulated learning processes. The findings revealed four main results. First, students who interacted with a generative AI tutor demonstrated significantly poorer metacognitive monitoring accuracy (calibration) on a post-test compared to students who used traditional instructional resources, with a medium-to-large effect size. Second, analysis of interaction logs and think-aloud protocols indicated that AI tutor users predominantly engaged in direct answer requests and verification requests, exhibiting patterns of cognitive offloading, whereas control group participants more frequently articulated internal metacognitive strategies. Third, prior domain knowledge moderated this effect: the negative impact of AI tutor use on calibration accuracy was most pronounced for students with lower prior knowledge, while high-knowledge students were less adversely affected. Fourth, the metacognitive deficit

observed in the AI tutor group persisted and even widened on a transfer task completed without AI support, indicating a durable impairment in metacognitive monitoring that was not attributable to differential problem-solving performance.

4.1 Addressing the Research Questions

These findings directly address each of the four research questions guiding this study. Regarding RQ1, the results provide clear evidence that generative AI tutor use negatively affects metacognitive monitoring accuracy, confirming the theoretical concern that such tools may impair students' ability to accurately assess their own learning. For RQ2, the interaction log and think-aloud data illuminate the mechanisms underlying this effect, revealing that students using AI tutors tend to offload cognitive work and seek verification rather than engage in deep explanatory processing. In relation to RQ3, the moderation analysis demonstrates that this effect is not uniform but rather depends on learner characteristics, with lower-knowledge students being particularly vulnerable. Finally, concerning RQ4, the transfer task results establish that the metacognitive impairment is not merely situational but endures beyond the immediate context of AI support, raising important questions about the long-term development of self-regulatory capacity.

4.2 Comparison with Previous Literature

The findings of this study both align with and extend existing literature in several important ways. The observed pattern of cognitive offloading is consistent with theoretical work by Risko and Gilbert (2016), who conceptualized offloading as the use of external tools to reduce internal cognitive demands. The present study provides empirical evidence that this phenomenon extends to metacognitive processes in educational contexts, with students outsourcing not only cognitive problem-solving but also monitoring functions to the AI tutor.

The finding that AI tutor users predominantly engaged in direct answer requests (44%) and verification requests (32%) resonates with earlier research on help-seeking in intelligent tutoring systems. Aleven et al. (2003) documented a pattern of "help abuse" in which students sought the fastest path to correct answers rather than engaging with pedagogically meaningful support. The present study suggests that generative AI tutors, with their immediate and conversational affordances, may amplify this tendency, as students can obtain answers or verification with minimal effort.

The moderation effect of prior knowledge is particularly noteworthy and aligns with the expertise reversal effect documented in cognitive load theory (Kalyuga, 2007). This effect posits that instructional supports that benefit novices can become redundant or even detrimental for more knowledgeable learners. In the present study, however, the pattern was reversed: the negative impact was concentrated among lower-knowledge students, suggesting that generative AI tutors may be most harmful for those who lack the metacognitive foundation to use them judiciously. This finding extends the expertise reversal effect by demonstrating its applicability to metacognitive outcomes and to the novel context of generative AI.

The durability of the metacognitive deficit on the transfer task represents an original contribution to the literature. While previous studies have raised concerns about dependency on AI tools (Wylie, 2023), this study provides the first experimental evidence that the metacognitive impairment persists when the tool is removed. Notably, this occurred despite equivalent problem-solving performance between groups, indicating that students could still produce correct answers while being less aware of their own cognitive processes-a dissociation between performance and metacognition that has important implications for educational practice.

4.3 Agreements, Contradictions, and Unexpected Results

The findings are broadly consistent with the theoretical concerns articulated in the emerging literature on generative AI in education (Bond et al., 2024; Luckin et al., 2022). However, they stand in partial contrast to the extensive body of research on earlier generations of intelligent tutoring systems, which generally found positive or neutral effects on learning outcomes (Aleven & Koedinger, 2002). This apparent contradiction may be explained by qualitative differences between rule-based ITS and generative AI tutors. Traditional ITS are typically designed with explicit pedagogical scaffolds, including metacognitive prompts and structured help sequences. Generative AI tutors, in contrast, are more open-ended and responsive, potentially enabling students to bypass deeper processing more easily. This interpretation suggests that it is not intelligent tutoring per se, but the specific affordances of generative, conversational AI, that may pose metacognitive risks.

An unexpected finding was the absence of a significant difference in problem-solving performance on the transfer task, despite the clear difference in metacognitive accuracy. This dissociation between performance and metacognition suggests that students in the AI tutor group were able to produce correct answers while

being less aware of the limits of their understanding. From an educational perspective, this is concerning because accurate metacognitive monitoring is essential for guiding future learning efforts (Dunlosky & Thiede, 2013). Students who overestimate their competence may fail to engage in necessary review or practice, leading to gaps in long-term knowledge development.

4.4 Theoretical Implications

The findings of this study have several important implications for theories of metacognition and self-regulated learning in technology-rich environments. First, they suggest that Winne and Hadwin's (1998) model of SRL, which conceptualizes monitoring as an internal feedback loop informing subsequent control decisions, may need to be extended to account for distributed metacognitive systems in which monitoring functions can be partially or wholly externalized to AI tools. When students rely on AI for verification ("Am I right?"), they may be substituting external feedback for internal monitoring, potentially atrophying the very metacognitive skills the model seeks to explain.

Second, the findings contribute to Nelson and Narens' (1990) framework by demonstrating how an external agent can reconfigure the relationship between the object-level and meta-level. In the AI tutor condition, the meta-level's monitoring function appeared to be partially delegated to the AI, with students using the tool's responses as a proxy for self-assessment. This raises theoretical questions about the nature of metacognition when it is distributed across human and artificial agents, and whether the resulting hybrid system can be said to engage in metacognition in the same sense as an autonomous human learner.

Third, the moderation finding suggests that theories of metacognitive development must attend to the interaction between learner characteristics and tool affordances. Novice learners, who have not yet developed robust metacognitive skills, may be particularly susceptible to the offloading affordances of generative AI, while more expert learners may have the metacognitive resources to use such tools strategically. This interaction has implications for developmental models of self-regulated learning and for the design of adaptive learning environments.

4.6 Practical and Policy Implications

The findings of this study carry significant implications for educational practice, instructional design, and institutional policy. At the pedagogical level, the results underscore the urgent need for explicit instruction in AI literacy that encompasses not only

technical proficiency but also metacognitive awareness. Students must be taught not merely how to use generative AI tools, but when and why to use them, and perhaps most importantly, when to refrain from using them to preserve opportunities for internal cognitive and metacognitive engagement. Curricula should incorporate activities that help students recognize the metacognitive costs of offloading and develop strategies for using AI as a complement to, rather than a replacement for, their own thinking.

For instructional designers and educational technologists, the findings suggest that the design of generative AI tutors should incorporate features that promote rather than undermine metacognitive engagement. This might include built-in metacognitive prompts ("How confident are you in your answer before I respond?"), requirements that students attempt problems independently before accessing AI support, or interfaces that encourage explanatory rather than directive interactions. The goal should be to design tools that are not only helpful but also developmental-scaffolding students' metacognitive growth rather than supplanting it.

At the institutional level, these findings call for a measured and evidence-informed approach to AI integration. Rather than wholesale adoption or prohibition, universities should develop policies that guide judicious use based on learning objectives and student developmental levels. The finding that lower-knowledge students are most vulnerable suggests that first-year students or those in introductory courses may require more structured guidance and restriction in their use of generative AI, while advanced students may be better equipped to use these tools metacognitively. Assessment practices may also need to evolve, incorporating more process-oriented evaluations that capture students' metacognitive engagement alongside their final products.

4.7 Contribution to the Body of Knowledge

This study makes several original contributions to the academic literature. First and foremost, it provides the first rigorous experimental evidence on the metacognitive effects of generative AI tutors, moving the discourse beyond speculation and establishing an empirical benchmark for future research. By demonstrating a causal relationship between AI tutor use and impaired metacognitive monitoring, the study substantiates concerns that have previously been only theoretical.

Second, the study contributes methodological innovation by integrating multiple data sources-quantitative calibration measures, interaction log analysis, and qualitative think-aloud protocols-to provide a comprehensive picture of the metacognitive processes

involved. This multi-method approach responds to calls for more process-oriented research in the field of AI and learning (Azevedo & Gašević, 2019) and offers a template for future studies.

Third, the identification of prior knowledge as a moderating variable adds nuance to the literature, demonstrating that the effects of generative AI are not uniform but depend on learner characteristics. These findings advance theoretical understanding of the conditions under which AI tools support or hinder learning and has practical implications for differentiated instruction and support.

Finally, by demonstrating the durability of the metacognitive deficit on a transfer task, the study establishes that the observed effects have meaningful educational significance beyond the immediate experimental context. This finding addresses a critical gap in the literature and underscores the importance of considering long-term developmental outcomes in evaluations of educational technology.

4.8 Limitations and Future Research

While this study makes important contributions, several limitations should be acknowledged. First, the controlled laboratory setting, while enhancing internal validity, limits the generalizability of findings to authentic educational contexts. Future research should replicate these findings in naturalistic classroom settings over extended periods. Second, the study examined a single domain (introductory programming) with a specific AI tutor implementation; results may vary across disciplines and with different AI tools. Third, the sample consisted of undergraduate students from a single university; replication with diverse student populations is warranted. Fourth, the study examined immediate and near-transfer effects but could not assess long-term developmental trajectories; longitudinal research is needed to understand how repeated AI use shapes metacognitive development over time.

Future research should also investigate potential interventions to mitigate the negative effects observed here. Could metacognitive training before AI use, or modifications to AI tutor design that incorporate metacognitive prompts, preserve the benefits of AI support while protecting metacognitive development? Additionally, research should explore individual differences beyond prior knowledge, such as epistemic beliefs, goal orientation, and AI self-efficacy, that may influence how students interact with and are affected by generative AI tools.

5. CONCLUSION

This experimental study set out to investigate a question of pressing importance for contemporary higher education: what is the impact of generative AI tutors on students' metacognitive monitoring and self-regulated learning? The findings provide a clear and concerning answer. Students who interacted with a generative AI tutor demonstrated significantly poorer metacognitive calibration than those who used traditional resources, a deficit that was most pronounced among novices and that persisted when the AI support was removed. Analysis of interaction logs and think-aloud protocols revealed the underlying mechanism: students using AI tutors engaged in substantial cognitive offloading, outsourcing not only problem-solving but also the metacognitive functions of verification and monitoring to the tool.

These findings make several original contributions to the academic literature. They provide the first rigorous experimental evidence substantiating theoretical concerns about the metacognitive risks of generative AI in education. They extend established theories of self-regulated learning to account for distributed cognition with generative AI tools, demonstrating how externalization of monitoring functions can impair internal metacognitive capacity. They identify prior knowledge as a critical moderating variable, revealing that those students most in need of developing robust metacognitive skills-novices-are paradoxically most vulnerable to the metacognitive costs of AI use. And they establish the durability of this effect, showing that the impairment transfers to contexts where the AI is no longer available.

The implications of this study extend beyond the laboratory into classrooms, lecture halls, and institutional policy discussions. For educators, the findings underscore the urgent need to teach not only how to use AI, but when and why-and perhaps most importantly, when to refrain from using it to preserve opportunities for internal cognitive and metacognitive engagement. For instructional designers, they call for the development of AI tools that scaffold rather than supplant metacognitive processes, incorporating features that prompt self-assessment and encourage explanatory rather than directive interactions. For institutional leaders and policymakers, they demand a measured, evidence-informed approach to AI integration that considers student developmental levels and protects the foundational educational goal of cultivating autonomous, self-regulated learners.

As generative AI continues its rapid integration into the fabric of higher education, this study serves as both an empirical contribution and a cautionary note. The promise of AI

to democratize personalized learning is real and significant. But that promise will only be realized if we attend with equal seriousness to the potential costs. The central challenge for educators and researchers in the coming years will be to harness the power of these tools while safeguarding-and indeed actively cultivating-the metacognitive autonomy that remains the hallmark of an educated mind. This study suggests that such a balance is not automatically achieved; it must be deliberately designed, intentionally taught, and continually assessed. The future of learning in an AI-rich world depends on getting this balance right.

6. RECOMMENDATIONS

Based on the findings of this experimental study, which demonstrated that generative AI tutors can impair undergraduate students' metacognitive monitoring accuracy and promote cognitive offloading-particularly among novice learners-the following recommendations are proposed for educational practice, instructional design, institutional policy, and future research.

6.1 Recommendations for Educational Practice

Educators should integrate explicit metacognitive training into curricula that incorporate generative AI tools. Students must be taught not merely the technical skills of prompting AI systems, but the metacognitive competencies required to use them judiciously. This includes instruction on:

- Recognizing the cognitive and metacognitive costs of offloading.
- Developing strategies for attempting problems independently before seeking AI assistance.
- Critically evaluating AI-generated responses rather than accepting them as authoritative.
- Calibrating self-assessment by comparing internal judgments with external feedback.

Furthermore, assessment practices should be diversified to capture process-oriented outcomes alongside final products. Portfolios, reflective journals, and think-aloud assessments can provide insight into students' metacognitive engagement and help counteract the hidden costs of AI use that may not be visible in final outputs alone.

6.2 Recommendations for Instructional Design

Developers of generative AI tutors should incorporate design features that promote rather than undermine metacognitive development. Recommended design principles include:

- Metacognitive prompting: AI tutors should be programmed to prompt users for self-assessment before providing answers (e.g., "How confident are you in your answer? Please explain your reasoning before I respond.").
- Structured help sequences: Rather than providing immediate answers, AI tutors should offer graduated assistance, beginning with hints and explanations before revealing solutions.
- Scaffolded verification: When students request verification, AI tutors should encourage self-checking strategies rather than simply confirming correctness.
- Usage analytics: AI platforms should provide students and educators with dashboards displaying patterns of interaction, including frequency of direct answer requests versus explanatory engagement, to foster awareness of help-seeking behaviors.

These design features should be developed in collaboration with educational psychologists and subjected to empirical testing to ensure they achieve their intended metacognitive benefits.

6.3 Recommendations for Institutional Policy

Higher education institutions should adopt evidence-informed policies governing the use of generative AI that attend to developmental considerations. Key policy recommendations include:

- Differentiated guidance: Recognizing that novice learners are most vulnerable to metacognitive impairment, institutions should provide more structured guidance and, where appropriate, restrictions on AI use in introductory courses, while allowing greater autonomy for advanced students.
- AI literacy frameworks: Institutions should develop and mandate AI literacy curricula that encompass not only technical proficiency but also critical evaluation, ethical reasoning, and metacognitive awareness.
- Assessment reform: Program-level assessment strategies should be reviewed to ensure they authentically evaluate students' independent capabilities alongside their ability to collaborate with AI tools.

Faculty development: Institutions should invest in professional development programs that equip educators with the knowledge and pedagogical strategies to guide students' metacognitive engagement with AI.

6.4 Recommendations for Future Research

The findings of this study open several avenues for further investigation. Future research should:

- Replicate in naturalistic settings: Conduct longitudinal studies in authentic classroom contexts to examine how repeated AI use shapes metacognitive development over extended periods.
- Investigate interventions: Test the effectiveness of metacognitive training programmes and AI design modifications in mitigating the negative effects observed in this study.
- Explore individual differences: Examine how learner characteristics such as epistemic beliefs, goal orientation, AI self-efficacy, and cultural background moderate the impact of AI on metacognition.
- Extend to diverse domains: Investigate whether the effects observed in introductory programming generalize to other disciplines with different epistemic structures and task demands.
- Examine developmental trajectories: Conduct longitudinal research tracking students from their first exposure to AI through to advanced study to understand how metacognitive relationships with AI evolve over time.
- Compare AI architectures: Investigate whether different AI designs (e.g., rule-based versus generative, directive versus Socratic) produce differential effects on metacognitive outcomes.

6.5 Concluding Statement on Recommendations

The integration of generative AI into higher education is not a future possibility but a present reality. The question is no longer whether these tools will be used, but how they will be used-and with what consequences for the cognitive and metacognitive development of learners. The recommendations outlined above provide a roadmap for educators, designers, policymakers, and researchers to navigate this terrain responsibly. By attending to the metacognitive dimensions of AI use, the academic community can work to ensure that these powerful tools serve to empower rather than enfeeble the minds of learners, cultivating the autonomous, self-regulated thinkers that higher education exists to develop.

References

1. Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. Cognitive Science, *26*(2), 147-179.

2. Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. Review of Educational Research, *73*(3), 277-320.

3. Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? Journal of Educational Psychology, *96*(3), 523-535.

4. Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: A call for action. Journal of Learning Analytics, *6*(3), 1-10.

5. Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., ... & Siemens, G. (2024). A meta-systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration and rigour. Studies in Higher Education, 1-23. Advance online publication.

6. Clark, R. E. (2020). Learning from media: Arguments, analysis, and evidence (2nd ed.). Information Age Publishing.

7. Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.

8. Dunlosky, J., & Thiede, K. W. (2013). The role of metacognition in learning. In D. Reisberg (Ed.), The Oxford Handbook of Cognitive Psychology (pp. 567-581). Oxford University Press.

9. Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data (Rev. ed.). MIT Press.

10. Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, *39*(2), 175-191.

11. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American Psychologist, *34*(10), 906-911.

12. Jager, J., Putnick, D. L., & Bornstein, M. H. (2017). II. More than just convenient: The scientific merits of homogeneous convenience samples. Monographs of the Society for Research in Child Development, *82*(2), 13-30.

13. Johnson, M., & Smith, R. (2023). Student and faculty perceptions of generative AI in higher education: A mixed-methods study. The Internet and Higher Education, *59*, Article 100921.

14. Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational Psychology Review, *19*(4), 509-539.

15. Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. SAGE Publications.

16. Luckin, R., Cukurova, M., Kent, C., & du Boulay, B. (2022). Empowering educators to be AI-ready. Computers and Education: Artificial Intelligence, *3*, Article 100076.

17. Molenaar, I. (2022). The concept of hybrid human-AI regulation: An agenda for learning analytics research. Journal of Learning Analytics, *9*(2), 1-14.

18. Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 26, pp. 125-141). Academic Press.

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

20. Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. Metacognition and Learning, *4*(1), 33-45.

21. Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.