

USING CHATGPT “TEACHER MODE” TO TEACH: A DESIGN-BASED RESEARCH PROTOCOL AND IMPLEMENTATION FRAMEWORK

Bharat Khushalani

Cypher Codes, Hebbal, Bangalore, India

e-mail: bharat.khushalani@gmail.com

Abstract: The unstructured use of generative artificial intelligence (AI) in education may encourage passive answer-seeking rather than active learning. This study introduces ChatGPT Teacher Mode, a structured framework designed to position generative AI as a guided instructional partner that promotes critical thinking, metacognition, and evidence-based reasoning. Using a design-based research approach, the study develops a protocol for implementing Teacher Mode across mathematics, history, civics, and geography. The framework incorporates discipline-specific learning activities, including articulating mathematical reasoning, evaluating primary sources, explaining constitutional principles, and interpreting climate maps. It integrates pedagogical scaffolding, formative assessment, verification routines, and transparent AI-use policies to support responsible implementation. The proposed framework targets four outcomes: improved learning through structured reasoning, enhanced metacognitive awareness, higher-quality feedback aligned with assessment rubrics, and greater transparency in AI use. Potential challenges, including inaccurate outputs, over-reliance, and inequitable access, are addressed through verification strategies, AI-assisted and AI-free assessments, and institutional guidelines. This study contributes a socio-technical framework, an empirically testable research protocol, and practical guidance for educators integrating generative AI into teaching and learning. Teacher Mode offers a structured approach to leveraging generative AI while maintaining academic integrity and supporting meaningful student engagement. Its effectiveness, however, requires empirical validation across educational contexts.

Keywords: academic integrity; chatgpt; formative assessment; generative ai; Teacher Mode.

INTRODUCTION

Generative AI tools such as ChatGPT provide immediate explanations and examples, but their unstructured use can allow students to submit answers without practicing the reasoning needed for durable learning. Responses to this challenge have ranged from restricting AI to exploring its educational potential. A more productive approach is to design learning activities

that require effort, evidence, reflection, and transparent use rather than treating AI as either a universal threat or an unrestricted answer engine (Tan & Maravilla, 2024; Chan, 2023).

This paper uses “Teacher Mode” to describe a proposed instructional configuration, not a claim that a single, universally available platform switch guarantees these behaviors. It combines available platform features such as study-oriented interaction and custom



instructions with teacher-designed routines and institutional safeguards. Its purpose is to position AI as a guided partner: students first attempt a task, explain their thinking, receive targeted prompts, verify information, and revise their work.

The research gap is the need for a replicable model connecting technology, pedagogy, and policy. AI can support feedback and access to explanations, but risks include hallucinations, bias, academic misconduct, and unclear accountability (Lee et al., 2024; Gruenhagen et al., 2024). Responsible integration therefore requires more than technical access; it requires explicit learning goals, assessment design, and governance aligned with practice (Alazmi & Alazmi, 2022).

The paper has two objectives: (1) conceptualize Teacher Mode as a socio-technical instructional framework; and (2) propose a DBR protocol for examining its implementation and potential effects. Four questions guide the protocol: (a) how does structured AI use relate to learning outcomes compared with usual instruction or unstructured AI use? (b) how does it shape metacognitive planning, monitoring, and revision? (c) how does its feedback compare with instructional rubrics and teacher feedback? and (d) how do students and instructors perceive usefulness, trust, fairness, and academic integrity?

The framework has three connected layers: platform, pedagogy, and policy. Platform features make guided interaction possible; pedagogy defines the learning sequence and discipline-specific prompts; policy sets boundaries for disclosure, privacy, and assessment. None is sufficient alone. A well-written instruction cannot guarantee factual accuracy, and a tool cannot

replace teacher judgment or institutional responsibility.

At the platform layer, custom instructions can ask the system to begin with diagnostic questions, provide hints rather than complete answers, flag uncertainty, and encourage verification. For example, a mathematics prompt may ask which operation isolates a variable before showing a worked step. Such instructions are design intentions, not technical guarantees; outputs still require monitoring and verification.

At the pedagogical layer, interactions follow a cycle of attempt, feedback, revision, and reflection. Students explain their reasoning before receiving help and assess whether feedback improves their work. This approach aligns with formative assessment, which uses evidence of learning to guide next steps (Carney et al., 2022), and with research emphasizing scaffolded, reliable AI support for self-regulated learning (Qian et al., 2025).

At the policy layer, teachers define permitted, conditional, and prohibited uses. Students disclose AI assistance and, where appropriate, submit selected conversation excerpts and a brief reflection. Policies should specify how records are used, who can access them, and how privacy is protected. Transparency is intended to assess learning processes, not to replace substantive evaluation of the student's own work.

Research question 1 concerns learning outcomes: do students show improved conceptual understanding, problem-solving accuracy, or transfer? Subject tests and performance tasks can compare baseline and post-intervention performance.

Research question 2 concerns metacognition: do students increasingly

plan, monitor, explain, and revise their thinking? Weekly reflection journals and a structured rubric can capture these behaviors without assuming that self-report alone proves learning.

Research question 3 concerns feedback quality. AI responses can be rated for specificity, accuracy, actionability, and alignment with course rubrics, then compared with teacher feedback on sampled tasks.

Research question 4 concerns legitimacy and trust. Student surveys and interviews with students and instructors can explore usefulness, confidence, concerns about dependence, perceived fairness, and the clarity of institutional rules.

METHOD

The study proposes a design-based research approach because Teacher Mode is an instructional innovation to be designed, implemented, evaluated, and refined in authentic classrooms. DBR links practical problems with iterative design and evidence from implementation. The protocol is prospective: the sample, measures, and outcomes below describe a plan, not completed research findings.

Participants would comprise approximately 120–240 secondary or post-secondary students across mathematics, history, civics, and geography, subject to institutional access, consent, and feasibility. Within each subject, one class could use the structured framework and a comparison class could receive usual instruction or unstructured ChatGPT access. Because intact classes may differ at baseline, analyses should account for prior achievement and report group allocation and attrition

transparently.

The intervention spans eight weeks. Week 1 introduces the framework, responsible-use rules, and discipline-specific instructions. Weeks 2–6 embed guided AI activities in regular coursework. Weeks 7–8 use capstone tasks to assess whether students can apply reasoning practices to integrated or unfamiliar problems. Teachers document deviations and refine prompts between cycles, while keeping core design principles stable enough for comparison.

Activities reflect disciplinary practice. In civics, students explain a constitutional principle before receiving feedback and then revise terminology or application. In geography, they interpret maps or graphs before proposing causal explanations. Mathematics tasks require a proposed solution and justification of steps; history tasks require claims supported by primary or secondary evidence. These examples are adaptable templates, not identical interventions across subjects.

Capstone work asks students to use AI as a guide rather than an author. History students may develop a comparative argument and use feedback to examine evidence; mathematics students may solve multi-step problems and explain each transformation. An AI-free task or oral explanation can help determine whether learning transfers beyond assisted practice.

Data collection aligns with the four research questions. Pre- and post-tests assess learning; essays and problem solutions are scored with subject-specific rubrics; weekly journals document attempts and revisions; sampled AI feedback is rated for quality; and surveys/interviews examine trust and perceived usefulness. Instruments should be piloted, and raters trained using shared

scoring guidance.

Feedback quality is rated on four dimensions: specificity (identifies the issue), accuracy (is factually and conceptually sound), actionability (offers a feasible next step), and alignment (matches learning objectives and rubrics). Independent raters should score a subset of responses, with inter-rater agreement reported. AI feedback should not be presumed equivalent to teacher feedback before comparison.

Quantitative analysis may use ANCOVA or an appropriate alternative to compare post-test scores while accounting for baseline performance, alongside effect sizes and confidence intervals. Assumptions and class-level clustering should be considered; where sample size or design does not support ANCOVA, the analysis plan should be adjusted transparently. Qualitative reflections and interviews can be coded thematically, with coding procedures and discrepant cases documented. Integration of both strands can explain not only whether outcomes differ but how students use feedback.

Ethical procedures include informed consent, guardian consent where required, de-identification of transcripts, restricted access, data minimization, and clear retention and deletion rules. Personal or sensitive information should not be entered into AI systems. AI use supplements rather than replaces instruction and assessment; teachers retain responsibility for evaluating work and addressing errors. The protocol requires institutional review where applicable.

RESULTS AND DISCUSSION

Implementation Framework

Implementation of ChatGPT Teacher Mode requires coherent instructional design rather than reliance on a single technical setting. Teachers should begin by identifying intended learning outcomes and the cognitive processes students are expected to demonstrate. These outcomes inform the configuration of AI prompts, the design of repeatable learning activities, the selection of assessment evidence, and the establishment of boundaries for responsible AI use. This approach is consistent with Liao et al. (2024), who investigated a ChatGPT-based scaffolding system for computational thinking. Their findings indicated improvements in computational thinking, although improvements in problem-solving skills were not established. This distinction suggests that AI-supported scaffolding should be designed around explicit learning objectives and evaluated against specific competencies rather than assumed to improve all dimensions of learning.

At the technical level, custom instructions should define the AI's instructional role, limitations, and response strategies. The system should ask students to attempt a task before providing assistance, use guiding questions and graduated hints, request evidence for claims, identify uncertainty, encourage verification against credible sources, and avoid completing graded assignments. Teachers should test these instructions using representative student questions and revise them when responses are inaccurate, excessively directive, or misaligned with lesson objectives. Such iterative refinement is important because the educational value of generative AI depends not only on its technical capabilities but also on how those capabilities are incorporated into

instructional design.

The pedagogical layer should establish a consistent but flexible learning routine: students predict or attempt a solution, explain their reasoning, request feedback, verify the feedback, revise their work, and reflect on what has changed. This sequence makes the learning process visible and supports the development of self-monitoring skills. Evidence synthesized by Deng et al. (2025), in a systematic review and meta-analysis of experimental studies, indicates that ChatGPT interventions can improve academic performance and higher-order thinking outcomes. However, the authors also identified methodological limitations and substantial variation across studies, emphasizing the need for rigorous assessment and cautious interpretation. Accordingly, Teacher Mode should be evaluated through observable learning outcomes rather than students' perceptions of usefulness alone.

Feedback quality is another essential component of the framework. Rather than offering generic statements such as “incorrect” or “good job,” Teacher Mode should identify specific reasoning gaps and suggest actionable next steps. In mathematics, this may involve checking sign changes in an equation; in history, it may require connecting an argument to a relevant primary source. The importance of this distinction is supported by Steiss et al. (2024), whose study compared human and ChatGPT feedback on student writing. Although ChatGPT offered potential benefits in terms of efficiency, human feedback performed better on several dimensions of formative feedback. Similarly, Mahapatra (2024) reported positive effects of ChatGPT-based formative feedback on the academic writing skills of undergraduate

English-as-a-second-language students. Taken together, these findings support the use of AI for timely feedback while underscoring the need for teacher oversight, explicit assessment criteria, and verification of feedback accuracy.

Process documentation should provide evidence of how students engage with AI and use feedback to improve their work. Depending on the task, this evidence may include selected prompts, relevant AI responses, revised work, and a brief reflection explaining which suggestions were accepted, modified, or rejected. Documentation should be proportionate to the learning objectives and assessment requirements. Requiring complete transcripts for every activity may create unnecessary workload and privacy risks. The framework therefore prioritizes purposeful documentation that supports assessment and reflection while limiting the collection of unnecessary personal information.

The policy layer should clarify acceptable AI use through a context-sensitive traffic-light system. Green activities may include practice questions, brainstorming, and requesting hints. Yellow activities may include AI-assisted drafting or language feedback when disclosure and reflection are required. Red activities may include submitting AI-generated work as one's own or using AI in assessments where it is expressly prohibited. These categories must be adapted to course objectives, assessment conditions, and institutional policies. Zhang and colleagues' (2024) systematic review of ChatGPT use in K–12 education identified both educational opportunities and concerns about output quality and academic integrity. These concerns reinforce the need for explicit rules, transparent expectations, and teacher supervision rather than

unrestricted or indiscriminate AI adoption.

Finally, Teacher Mode should remain adaptable across disciplines while maintaining consistent instructional principles. Mathematics requires valid procedures and justified solutions; history emphasizes contextualized evidence and source evaluation; civics requires conceptual precision and reasoned application; and geography emphasizes the interpretation of spatial data and relationships. Lo et al. (2024), in a systematic review of 72 studies, found that ChatGPT-supported learning could produce both engagement and disengagement, highlighting the importance of instructional context and further evaluation. Therefore, implementation should be monitored through student performance, the quality of revisions, feedback accuracy, and learner reflections. The proposed framework provides a structured basis for implementation, but its effectiveness across disciplines and educational settings remains an empirical question to be investigated through design-based research.

Table 1. presents a planning matrix linking learner action, AI response, and evidence of learning. It is a proposed implementation guide; actual prompts and rubrics should be adapted and validated in each context.

Discipline	Initial learner action	Teacher Mode response	Evidence of learning
Mathematics	Show an attempted solution and explain the first step.	Ask a diagnostic question, identify a specific reasoning gap, and provide a hint before revealing a worked solution.	Accuracy of intermediate steps, explanation of procedures, and performance on a new problem.
History	State a claim and identify a relevant primary or secondary source.	Ask how the source supports the claim; prompt attention to context, perspective, and corroboration.	Quality of argument, relevance and interpretation of evidence, and historical reasoning.
Civics	Explain a principle in the student’s own words and give an example.	Identify imprecise terminology and request justification or application to a concrete case.	Conceptual accuracy, use of appropriate terminology, and application to a new scenario.
Geography	Interpret a map, graph, or spatial pattern before explaining it.	Ask the student to describe the evidence, test a causal interpretation, and consider consequences.	Accuracy of data interpretation, quality of causal explanation, and evidence-based inference.

Implementation Reporting Checklist
A replicable report should describe the instructional context, learner level, subject objectives, duration, and the

rationale for selecting Teacher Mode. It should identify the AI product and version where this information is available, the instructions used, the types

of tasks assigned, and the role of the teacher during each activity. If system settings or model behavior change during the study, researchers should document the change and consider whether it affects comparability. This information allows readers to assess whether a result is likely to transfer to their own setting.

The importance of systematic intervention reporting is supported by Upsher et al. (2025), who introduced the CheckList Of Standards of Reporting in Education Research (CLOSER) and the Checklist for Intervention Description of Education Research (CIDER). These complementary checklists were developed to address limitations in the reporting of educational interventions and to improve the transparency, replicability, and practical interpretation of research findings. CLOSER addresses the reporting of the overall research process, whereas CIDER focuses on describing intervention characteristics in sufficient detail. Applied to the proposed Teacher Mode framework, these principles suggest that reporting should extend beyond identifying the AI tool to documenting the instructional procedures, implementation conditions, teacher involvement, and adaptations made during the intervention.

The report should distinguish planned procedures from procedures actually delivered. For each cycle, researchers can summarize the design decision, the evidence that prompted it, the change made, and the consequence observed in the next iteration. A concise design log is especially valuable in design-based research (DBR) because the intervention evolves through engagement with classroom realities. Without such a record, readers may be unable to determine whether an outcome reflects the original design, a later revision, or

local teacher adaptation. Consistent with the emphasis on transparent intervention reporting, the design log should record both intended procedures and departures from the original plan, including the reasons for those changes.

A useful reporting matrix links each research question to its data source, analysis, and interpretive limit. Learning outcomes may be examined through tests and scored tasks, but a score difference alone does not explain which interaction produced the difference. Metacognitive journals can show how learners describe monitoring and revision, but they do not independently establish durable behavior. Feedback ratings can characterize response quality within the sampled tasks, but cannot guarantee accuracy in all future responses. Interviews can clarify perceptions, while remaining sensitive to social desirability and participants' prior expectations. Triangulation strengthens interpretation when these limits are stated rather than concealed.

Recent evidence on ChatGPT interventions reinforces the need to connect outcome measures with clearly specified research questions. Deng et al. (2025), in a systematic review and meta-analysis of experimental studies, reported generally positive effects of ChatGPT interventions on academic performance and higher-order thinking propensities. However, they also identified methodological concerns, including insufficient attention to statistical power and limitations in post-intervention assessment. Their findings suggest that evaluation should not rely solely on immediate achievement scores or learners' perceptions. For the proposed framework, researchers should specify in advance how learning, metacognitive awareness, feedback quality, and responsible AI use will be measured, and

should distinguish outcomes assessed with AI assistance from those assessed independently.

Anticipated Outcomes

The first anticipated outcome is improved learning, especially where students must articulate reasoning before receiving feedback. In mathematics, the approach may support transfer by requiring justification of each step; in history, it may improve argument–evidence alignment. These expectations are consistent with Deng et al. (2025) and Urban et al. (2024), whose systematic review and meta-analysis of experimental studies found an overall positive effect of ChatGPT interventions on academic performance and higher-order thinking propensities.

A second expected outcome is stronger metacognitive awareness. Prompts to predict, monitor, and revise may help learners recognize gaps in their understanding. Reflection journals can show whether students explain changes in strategy, but claims about durable self-regulation require performance evidence and follow-up after AI scaffolding is withdrawn (Ng et al., 2024).

A third expected outcome is more actionable formative feedback. A consistent rubric can reveal where AI feedback is specific and useful and where it is inaccurate, generic, or misaligned. Comparison with teacher feedback is necessary before drawing conclusions about quality, efficiency, or workload (Steiss et al., 2024).

A fourth expected outcome is greater transparency and trust when rules are clear and AI assistance is disclosed. Transcript excerpts and reflections may help teachers assess process, but students' perceptions may vary and documentation alone does not guarantee integrity.

Surveys and interviews will test these possibilities (Lo, 2024).

Taken together, these outcomes define an evaluation agenda rather than a set of findings. The examples illustrate plausible classroom use; they do not demonstrate causal effectiveness or generalizability.

Design Principles For Responsible Use

Several design principles follow from the framework. First, cognitive responsibility should remain with the learner: AI should ask for an initial attempt and make the reasoning visible before offering support. Second, support should be graduated. A diagnostic question or small hint is preferable to a complete solution when the learning objective is to practice problem-solving. Direct explanation may still be appropriate when the teacher determines that a learner needs explicit instruction, but the choice should be intentional and aligned with the objective.

Verification should be taught as a routine rather than appended as a warning. Learners should compare factual claims with authoritative sources, check calculations independently, and explain why they accept or reject a suggestion. In history, this may involve checking a quotation against a primary document; in geography, comparing an interpretation with the map legend and data; and in mathematics, substituting a result into the original equation. The evidence of learning is not simply that students receive feedback, but that they evaluate it and make defensible decisions.

The framework should preserve teacher judgment and learner agency. The AI is not the final authority on subject content, grading, or student support needs. Teachers select tasks, validate materials, decide when to intervene, and

interpret evidence in context. Students should be allowed to challenge a response and explain why they reject it. Such disagreement can itself become an instructional opportunity, particularly when the AI produces an incomplete or incorrect response.

Assessment should separate supported practice from independent demonstration. AI-assisted tasks can reveal how learners use feedback and revise their work, whereas AI-free tasks can indicate what they can do without immediate assistance. Neither alone provides a complete picture. A balanced assessment plan uses both and explains their distinct purposes, avoiding the assumption that polished AI-supported work is equivalent to independent competence.

Risks, Limitations, And Mitigations

Teacher Mode does not remove the limitations of generative AI. Its value depends on verification, teacher oversight, equitable access, and alignment with the task. The proposed protocol should record implementation problems as part of DBR rather than treating them as incidental.

Accuracy and hallucination are persistent risks. An AI system may misstate historical facts, misread data, or provide flawed explanations while sounding confident. Students should verify factual claims against textbooks, primary sources, peer-reviewed research, or authoritative datasets. Instructions to cite sources can help but do not guarantee that sources or citations are genuine (Darvishi, 2024).

Over-reliance may weaken independent problem-solving if students request hints for every step. The design therefore combines AI-supported practice with AI-free assessments, oral

explanations, and tasks requiring transfer. Teachers should gradually reduce scaffolding and evaluate whether students can apply skills independently.

Access inequity may arise where students have different devices, connectivity, or permitted accounts. Institutions can provide supervised access, accessible alternatives, and equivalent low-tech tasks. Participation and access barriers should be documented so that observed differences are not mistakenly attributed to the instructional design alone.

Students and teachers may disagree about what constitutes legitimate assistance. Explicit green/yellow/red categories, examples modeled in class, and consistent course-level communication can reduce ambiguity. Policies should distinguish acceptable support from undisclosed substitution and provide a fair process for resolving concerns.

AI may also constrain creativity or lack disciplinary nuance, especially in open-ended writing, arts, or advanced scientific inquiry. It should support foundational reasoning and revision while teachers and peers retain central roles in interpretation, creative judgment, and collaborative discussion.

Inconsistent implementation can undermine comparability and trust. Teachers should share core instructions, document adaptations, and use common rubrics where appropriate. The framework also has methodological limits: proposed participants and duration may not capture long-term learning, and results may vary across institutions, age groups, languages, and access conditions. These limitations make iterative testing essential.

Contributions

The principal theoretical contribution is to define Teacher Mode as a socio-technical framework linking platform affordances, instructional routines, and governance. This shifts attention from tool capability alone to the conditions under which AI use can support learning.

The methodological contribution is a testable DBR protocol connecting research questions to outcomes, instruments, and iterative refinement. Pre/post measures, rubric-based scoring, reflections, feedback ratings, and interviews allow the framework to be examined from cognitive, pedagogical, and ethical perspectives (Qian et al., 2026).

The practical contribution is a reusable implementation sequence: configure prompts, require an initial attempt, provide guided feedback, verify outputs, revise work, document use proportionately, and evaluate learning independently. Subject-specific examples help educators adapt the sequence without abandoning disciplinary standards.

The governance contribution is to treat transparent AI use as part of assessment design. Disclosure and clear boundaries can support accountability, provided that privacy is protected and responsibility for evaluation remains with educators.

Because the sequence can be adapted to multiple disciplines, Teacher Mode offers a cross-curricular design proposition. Its transferability, however, must be demonstrated through empirical work rather than assumed from conceptual breadth.

Implications And Future Work

For teaching, the framework

positions educators as designers of learning processes who decide when AI is useful, what reasoning students must show, and how feedback is evaluated. For curriculum and assessment, it encourages attention to both the final product and the process by which students construct and revise understanding.

Professional development should include prompt design, source verification, rubric alignment, privacy protection, and strategies for reducing dependence. Teachers need opportunities to test AI responses and discuss cases where the system is misleading or unsuitable.

Institutions should publish clear, proportionate policies defining disclosure, permitted uses, data handling, and accountability. Policy should support responsible experimentation while ensuring consistent expectations across courses and equitable access.

Future studies should test the framework across additional subjects and educational contexts, compare alternative forms of scaffolding, and examine how implementation fidelity relates to outcomes. Researchers should report null or negative findings as well as positive results.

Longitudinal research should assess whether students retain reasoning and self-regulation skills after support is removed. AI-free tasks, delayed tests, and transfer activities can help distinguish temporary performance support from durable learning.

Further policy research should examine effects on teacher workload, assessment practices, equity, and student perceptions. Research should also investigate which forms of transcript documentation are educationally useful without imposing disproportionate privacy and administrative burdens.

CONCLUSION

This paper proposes ChatGPT Teacher Mode as a socio-technical framework that integrates platform features, pedagogical routines, and institutional policy to support active, evidence-based, and transparent learning. Its DBR protocol outlines an eight-week implementation across mathematics, history, civics, and geography and proposes mixed-method evaluation of learning outcomes, metacognition, feedback quality, and perceptions of integrity. The framework emphasizes student attempts before assistance, verification, revision, proportionate documentation, and AI-free assessment. Accuracy errors, dependence, inequitable access, and inconsistent policies remain material risks requiring teacher oversight and institutional safeguards. Because the paper presents anticipated outcomes rather than completed empirical findings, future studies must test effectiveness, implementation conditions, and long-term transfer before strong claims about impact can be made.

REFERENCES

- Alazmi, A. A., & Alazmi, H. S. (2022). Closing the gap between research and policy-making to better enable effective educational practice: a proposed framework. *Educational Research for Policy and Practice*, 22.
- Carney, E. A., Zhang, X., Charsha, A., Taylor, J. N., & Hoshaw, J. P. (2022). Formative Assessment Helps Students Learn Over Time: Why Aren't We Paying More Attention to It? *Intersection: A Journal at the Intersection of Assessment and Learning*, 4(1).
- Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20(1), 1–25.
- Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, 210, 104967.
- Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, 227, 105224.
- Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence*, 6, 100198.
- Holzmann, U., Anand, S., & Payumo, A. Y. (2025). The ChatGPT fact-check: Exploiting the limitations of generative AI to develop evidence-based reasoning skills in college science courses. *Advances in Physiology Education*, 49(1).
- Lee, D., Arnold, M., Srivastava, A., Plastow, K., Strelan, P., Ploeckl, F., Lekkas, D., & Palmer, E. (2024). The Impact of Generative AI on Higher Education Learning and Teaching: A Study of Educators' Perspectives. *Computers and Education*.

- Artificial Intelligence, 6(100221), 100221.
- Liao, J., Zhong, L., Zhe, L., Xu, H., Liu, M., & Xie, T. (2024). Scaffolding computational thinking with ChatGPT. *IEEE Transactions on Learning Technologies*, 17, 1628-1642.
- Lo, C. K., Hew, K. F., & Jong, M. S.-Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education*, 219, 105100.
- Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, 11.
- Mai, D. T. T., Da, C. V., & Hanh, N. V. (2024). The use of ChatGPT in teaching and learning: A systematic review through SWOT analysis approach. *Frontiers in Education*, 9.
- Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology*, 55(4), 1328-1353.
- Qian, K., Liu, S., Li, T., Mladen Raković, Li, X., Guan, R., Molenaar, I., Nawaz, S., Swiecki, Z., Yan, L., & Dragan Gašević. (2025). Towards reliable generative AI-driven scaffolding: Reducing hallucinations and enhancing quality in self-regulated learning support. *Computers & Education*, 105448–105448.
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894.
- Tan, M. J. T., & Maravilla, N. M. A. T. (2024). Shaping integrity: why generative artificial intelligence does not have to undermine education. *Frontiers in Artificial Intelligence*, 7.
- Upsher, R., Dommett, E., Carlisle, S., Conner, S., Codina, G., Nobili, A., & Byrom, N. C. (2025). Improving reporting standards in quantitative educational intervention research: introducing the CLOSER and CIDER checklists. *Journal of New Approaches in Educational Research*, 14(1), 2.
- Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education*, 215, 105031.
- Yin, J., Goh, T.-T., & Hu, Y. (2024). Using a chatbot to provide formative feedback: A longitudinal study of intrinsic motivation, cognitive load, and learning performance. *IEEE Transactions on Learning Technologies*, 17, 1378–1389.
- Zhang, P., & Tur, G. (2024). A systematic review of ChatGPT use in K-12 education. *European Journal of Education*, 59(2), e12599.