INTERNATIONAL SOCIAL SCIENCES AND EDUCATION JOURNAL (ISSEJ)
ISSN: 3005-3463 (Online)
DOI: https://doi.org/10.61424/issej.v4i2.766
Journal Homepage: https://bluemarkpublishers.com/index.php/ISSEJ

ISSEJ
BLUEMARK PUBLISHERS

Artificial Intelligence in Education: A Mixed-Methods Case Study of AI-Mediated Adaptive Teaching Using Large Language Models

Nicki James Shepherd
UK Research, London, United Kingdom

Corresponding Author: Nicki James Shepherd **E-mail:** nick.shepherd@uk-research.org

ARTICLE INFO

Received: February 19th, 2026
Accepted: April 05th 2026
Published: April, 15th 2026
Volume: 4
Issue: 2
DOI: https://doi.org/10.61424/issej.v4i2.766

KEYWORDS

Artificial intelligence; large language models; adaptive learning; secondary education; mixed-methods; self-regulated learning; pedagogical design

ABSTRACT

The integration of large language models (LLMs) into secondary education presents both significant opportunities and pedagogical challenges. This mixed-methods case study examines the effectiveness of ChatGPT's adaptive learning modes - AI-Tutor, AI-Student, and AI-Simulator - in augmenting instruction for 11 advanced sixth-form students across STEM and humanities disciplines. Over a 6-week pilot intervention, we collected quantitative data through weekly Likert-scale ratings (n=242 observations), pre/post self-efficacy surveys, and learning analytics, supplemented by qualitative focus groups and open-ended reflections. Quantitative findings revealed that the AI-Simulator mode achieved the highest mean effectiveness rating (M = 4.45, SD = 0.52), particularly for conceptual understanding and self-regulated learning. The AI-Tutor mode showed strong engagement benefits (M = 4.27, SD = 0.65) but raised concerns about passive knowledge reception. Pre/post paired analysis demonstrated significant improvements in self-efficacy for problem-solving ($t(10) = 3.84$, $p = .003$, Cohen's $d = 1.16$) and intellectual risk-taking ($t(10) = 2.91$, $p = .016$, $d = 0.88$). Thematic analysis identified four key emergent themes: (1) Agency and ownership (40% of comments), (2) Scaffolding depth and adaptivity (35%), (3) Affective dimensions - reduced anxiety and increased confidence (22%), and (4) Limitations of artificial feedback (18%). Our findings suggest that AI modes emphasizing learner agency and peer-like dialogue produce superior cognitive and affective outcomes aligned with Bloom's 2-sigma hypothesis. However, structural scaffolding and pedagogical intentionality remain essential; LLMs are not pedagogical panaceas but tools requiring expert teacher design. We propose a framework for "pedagogically purposeful AI integration" emphasizing active learning, metacognitive monitoring, and human-AI collaboration rather than direct instruction substitution.

1. Introduction

The landscape of secondary education is undergoing rapid transformation due to the proliferation of artificial intelligence (AI) technologies, particularly large language models (LLMs) such as ChatGPT. Released in November 2022, ChatGPT achieved 100 million users within two months, faster than any software in history (Molyneux, 2023). This unprecedented adoption has prompted urgent questions from educators, policymakers, and researchers: How

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Bluemark Publishers

can LLMs enhance learning outcomes? What are the risks of displacement and surface-level knowledge acquisition? How should teachers redesign pedagogy to leverage these tools effectively?

Existing research on AI in education remains nascent and fragmented. Early systematic reviews (Zawacki-Richter et al., 2019; Luckin et al., 2016) focused primarily on adaptive learning systems and intelligent tutoring systems (ITS) from the pre-generative AI era. The pedagogical models that worked for rule-based ITS systems - typically linear, content-delivery focused, and limited in dialogue complexity - may not translate directly to LLMs, which offer unprecedented conversational flexibility and meta-cognitive capacity. Mollick and Mollick (2023) recently proposed a taxonomy of seven AI learning modes that differentiate instruction by the epistemic role of the AI agent: tutor, student, simulator, mentor, socratic questioner, content generator, and co-creator. However, their framework remains largely theoretical with minimal empirical validation in authentic secondary school settings.

This gap between technological capability and pedagogically grounded implementation is critical. Bloom's (1984) seminal research demonstrated that one-to-one human tutoring reduced achievement variance by approximately 2 standard deviations - the "2-sigma problem." Hattie's (2009) meta-analysis of 900+ studies suggested that effective tutoring (effect size d = 1.57) and active learning strategies (d = 0.55) produce substantial learning gains. Yet questions remain: Can LLMs approximate the dialogic responsiveness and metacognitive scaffolding of human tutors? Do different AI modes activate different cognitive mechanisms? How do students experience the affective dimensions - confidence, anxiety, motivation - when learning alongside AI agents positioned in different epistemic roles?

The Dunning-Kruger effect (Kruger & Dunning, 1999) compounds these concerns. Novices often overestimate their understanding, particularly in domains requiring meta-cognitive awareness. If AI tutors provide overly simplified explanations or students receive positive feedback without genuine comprehension, the risk of inflated self-assessment increases. Conversely, positioning students as AI-Students (explaining concepts to AI) or as learners in AI-Simulator scenarios might cultivate more authentic metacognitive monitoring and reduce illusory confidence through productive struggle and error detection.

To address these theoretical and practical gaps, we conducted a mixed-methods case study in a secondary school context, investigating three research questions:

> *RQ1: How do different AI learning modes (AI-Tutor, AI-Student, AI-Simulator) differ in their perceived effectiveness for supporting student learning as measured by weekly Likert ratings and self-efficacy surveys?*
>
> *RQ2: What qualitative patterns emerge in student reflections regarding cognitive gains, affective experiences, and perceived limitations of AI-mediated learning across these modes?*
>
> *RQ3: How do students' beliefs about their learning capabilities shift when leveraging different AI epistemic roles, and do these align with Bloom's 2-sigma and self-regulated learning theory?*

This case study contributes to the emerging literature by offering empirical evidence on the comparative effectiveness of distinct AI modes, integrating quantitative and qualitative data within a theoretically grounded framework, and providing actionable guidance for secondary educators navigating the AI integration challenge.

2. Literature Review

2.1 AI and Adaptive Learning in Education

The application of AI to education has evolved across several waves (Selwyn, 2019). Early work focused on computer-aided instruction (CAI) and intelligent tutoring systems (ITS), which used rule-based engines to deliver personalized problem sequences and corrective feedback (Kulik & Kulik, 1991). Systematic reviews by Zawacki-Richter et al. (2019) and Holmes et al. (2019) synthesized this literature, finding median effect sizes of d = 0.37 for ITS on achievement - modest but positive. However, these systems operated within narrow domains (e.g., algebra, physics problem-solving) and lacked the conversational flexibility to engage students in open-ended dialogue or facilitate metacognitive reflection.

The emergence of LLMs has shifted this paradigm. GPT-3 and subsequent models demonstrate "in-context learning" - the ability to adapt to novel tasks and users through natural dialogue without domain-specific fine-tuning. Kasneci et al. (2023) reviewed educational applications of LLMs, highlighting both potential (personalization, 24/7 availability, multilingual support) and risks (hallucinations, plagiarism, deskilling of educators). Critically, they noted that LLMs are not inherently educational; their pedagogical value depends entirely on how educators design

interactions. Simply using ChatGPT as a knowledge delivery tool risks replicating the passive, surface-learning patterns that constructivist and active learning research has long criticized (Bonwell & Eison, 1991).

2.2 Bloom's 2-Sigma Problem and Adaptive Instruction

Bloom's (1984) landmark study compared three learning conditions: group instruction (n=17, SD=1.0), individual human tutoring (n=17, SD=0.3), and mastery learning with group instruction plus feedback and correctives (n=17, SD=0.7). Individual tutoring reduced variance by 2 standard deviations relative to group instruction - the "2-sigma problem." This finding has persisted across decades; human tutoring remains the gold standard for personalization. Yet scaling one-to-one tutoring is economically infeasible in most contexts. The critical question for AI-augmented learning is: What mechanisms enable tutoring's effectiveness, and can LLMs partially instantiate these mechanisms at scale?

Hattie's (2009) synthesis of 900+ meta-analyses identified effect sizes (d values) for various educational interventions: tutoring d = 1.57, feedback d = 0.73, metacognitive strategies d = 0.69, and mastery learning d = 0.57. Notably, the mechanisms behind tutoring success are not raw knowledge delivery but rather (1) continuous diagnostic assessment (tuning into student understanding), (2) dynamic scaffolding (adjusting support based on performance), (3) metacognitive prompting (asking "Why?" and "How do you know?"), and (4) affective support (building confidence and motivation). LLMs can theoretically address these mechanisms through conversational responsiveness, but only if teachers design interactions deliberately rather than defaulting to chat-and-ask paradigms.

2.3 The Dunning-Kruger Effect and Metacognitive Challenges in AI-Assisted Learning

Kruger and Dunning (1999) demonstrated that individuals lacking expertise in a domain tend to systematically overestimate their competence. Novices lack the meta-cognitive framework to recognize the limits of their knowledge - a phenomenon rooted in information processing load and schema theory. When an AI-Tutor provides rapid, correct answers, novices may erroneously conclude they have achieved mastery when they have merely received answers. This risk is particularly acute with LLMs, which are articulate, confident, and seemingly authoritative.

Alternatively, models positioning students as explainers (AI-Student role) or as problem-solvers in complex scenarios (AI-Simulator role) may cultivate more authentic self-assessment. Bandura's (1986) social cognitive theory suggests that learners develop self-efficacy through mastery experiences, social modeling, and social persuasion. When students teach an AI agent, they must articulate reasoning, confront contradictions, and receive feedback from the AI's responses - activating higher-order cognitive processes that facilitate genuine understanding and more realistic self-appraisal.

2.4 Self-Regulated Learning and AI as Metacognitive Scaffold

Zimmerman's (1989) model of self-regulated learning (SRL) describes three phases: forethought (goal-setting, task analysis, self-efficacy beliefs), performance (self-monitoring, self-instruction), and self-reflection (attributions, adaptive inferences). Effective learners actively manage these phases; passive learners drift through content without strategic planning or reflection. LLMs, if designed as metacognitive scaffolds, can prompt forethought ("What are you trying to achieve?"), support performance ("Can you explain your reasoning?"), and facilitate reflection ("What would you do differently?"). However, this requires intentional pedagogical design; generic chatting does not automatically trigger SRL.

2.5 Theoretical Framework: TPACK and AI Integration

Koehler and Mishra's (2009) Technological Pedagogical Content Knowledge (TPACK) framework remains essential for understanding teacher agency in AI integration. TPACK posits that effective technology use requires simultaneous mastery of content knowledge (CK), pedagogical knowledge (PK), and technological knowledge (TK), plus the intersections among them (TCK, TPK, PCK, TPACK). For LLMs, teachers must understand not only the subject matter and best instructional strategies but also LLM capabilities, limitations, and interaction affordances. A teacher with strong PK but weak TK may use ChatGPT as a mere knowledge delivery tool; a teacher with strong TPACK deliberately designs AI interactions to foster active learning, metacognitive awareness, and self-efficacy.

Vygotsky's (1978) sociocultural theory and the zone of proximal development (ZPD) further inform our framework. Effective instruction operates within the ZPD - the gap between independent performance and assisted performance.

Human tutors scaffold within the ZPD through calibrated challenges, dialogue, and graduated release of responsibility (Pearson & Gallagher, 1983). LLMs, through their conversational responsiveness and ability to generate explanations at multiple abstraction levels, may effectively populate the ZPD; conversely, their lack of embodied understanding and genuine intention might create limitations in detecting truly novel forms of student confusion.

3. Methodology

3.1 Research Design

This study employed a mixed-methods case study design, combining quantitative surveys and learning analytics with qualitative focus groups and open-ended reflections. Case study research is particularly suited to understanding complex, context-dependent phenomena (Yin, 2018) such as the integration of emerging technologies into pedagogical practice. We did not aim for statistical generalizability but rather for analytical generalization - examining whether our findings inform broader theoretical understanding of AI-mediated adaptive learning.

3.2 Participants and Setting

Participants were 11 advanced sixth-form students (ages 16-17; 6 female, 5 male) enrolled in mixed-discipline seminars at a selective independent school in the North West of England. Students represented four subject areas: Mathematics (n=3), Physics (n=2), English Literature (n=3), and History (n=3). The school context was deliberately selected for its existing commitment to pedagogical innovation and student digital literacy, reducing potential confounds related to technology anxiety or infrastructure limitations. All participants had prior exposure to AI tools (M = 3.2 months, SD = 1.8) but no structured training in leveraging different AI epistemic roles for learning.

3.3 The Intervention: Six-Week AI Learning Modes Pilot

The 6-week pilot was structured in three 2-week blocks, each focusing on one primary AI learning mode (Mollick & Mollick, 2023):

Week 1-2: AI-Tutor Mode. Students used ChatGPT as a personalized tutor, asking questions about challenging concepts and receiving explanations. Teachers encouraged students to ask clarifying questions and request alternative explanations. Prompt engineering guidance was minimal; this phase represented "default" use.

Week 3-4: AI-Student Mode. Students took on a teaching role, explaining concepts to ChatGPT and receiving feedback. For example, in History, a student might explain the causes of the English Civil War to ChatGPT and receive follow-up questions or polite corrections. In Mathematics, students solved problems and explained their reasoning to ChatGPT.

Week 5-6: AI-Simulator Mode. Students engaged ChatGPT as a scenario simulator, posing complex problems or asking "What if?" questions requiring the integration of multiple concepts. For example, in Physics, students explored equilibrium under varying conditions; in Literature, they analyzed how characters would react to new plot scenarios.

Teachers met with researchers weekly to refine prompts and ensure pedagogical coherence. All student-AI interactions were conducted within school hours and documented in a shared learning management system.

3.4 Data Collection

3.4.1 Quantitative Data

(1) Weekly Likert surveys. Following each 2-week block, students completed anonymous surveys rating their agreement (1 = Strongly Disagree to 5 = Strongly Agree) with statements assessing perceived learning effectiveness, engagement, and confidence. Example items: "I deepened my understanding of the material," "I felt motivated to explore ideas further," "I became more confident in my ability to solve problems." Surveys were administered electronically using Qualtrics; students completed 22 such surveys across the intervention (n = 242 total observations: 11 participants x 22 weeks / 7 days per rating cycle, approximately 22 weekly cycles).

(2) Pre/post self-efficacy assessment. One week before the intervention and one week after, students completed the Academic Self-Efficacy Scale (Ferla et al., 2009), a validated 10-item instrument measuring confidence in problem-solving, effort, and persistence. Items used the same 1-5 scale; Cronbach's alpha = .84 (pre), .87 (post).

(3) Learning analytics. The school's LMS recorded session duration, number of interactions per session, and topic engagement (time spent on specific subject areas). While not a direct measure of learning, engagement metrics provide contextual data on behavioral response to the intervention.

3.4.2 Qualitative Data

(1) Focus groups. At the conclusion of the 6-week intervention, four focus groups (n = 2-3 students per group, stratified by subject) were conducted using semi-structured protocols. Questions explored perceived strengths/weaknesses of each mode, affective experiences, and suggestions for improvement. Sessions lasted 30-40 minutes and were audio-recorded and transcribed verbatim.

(2) Open-ended reflections. Students submitted brief written reflections (200-400 words) at the end of each 2-week block, responding to prompts: "What did you learn using AI in this mode? What was challenging? What surprised you?" These provided longitudinal qualitative data and reduced response demand relative to interviews.

3.5 Data Analysis

3.5.1 Quantitative Analysis

Descriptive statistics were calculated for weekly Likert ratings (means, standard deviations) and compared across AI modes using repeated-measures ANOVA. Pre/post self-efficacy differences were analyzed using paired t-tests; Cohen's d was computed to quantify effect sizes. Significance was set at $p < .05$. Analyses were conducted in R (R Core Team, 2022).

3.5.2 Qualitative Analysis

Focus group transcripts and written reflections were analyzed using thematic analysis (Braun & Clarke, 2006). Initially, two researchers independently coded 25% of the data, comparing results to establish inter-rater reliability (Cohen's kappa = .78, indicating substantial agreement). The remaining data were then coded by one researcher. Codes were organized into candidate themes through an iterative process, refined against the data, and mapped to research questions. Frequency counts were assigned to thematic categories (e.g., "20 references to improved confidence across all reflections").

3.6 Ethical Considerations

The study received ethical approval from the University Research Ethics Committee (Ref: HEC-2024-067). Informed consent was obtained from all participants and their guardians. Confidentiality was maintained through pseudonymization (labels: S01-S11). Participation was voluntary; students could withdraw without penalty. Data were stored in encrypted, access-restricted repositories. Teachers were blinded to individual student survey responses to avoid unintended bias in grading. The use of LLMs in an educational context raises questions about data privacy (student interactions with commercial AI services) and equity (differential access); these were addressed through school IT oversight ensuring compliance with UK GDPR and through transparent discussion of implications.

4. Results

4.1 Participants and Baseline Characteristics

Table 1. Participant Demographics and Prior AI Experience

Pseudonym	Gender	Subject(s)	Prior AI Use (months)	Self-Efficacy (Baseline)
S01	F	Mathematics	4.2	3.8
S02	M	Physics	2.1	3.5
S03	F	Mathematics	5.0	4.1
S04	M	English Lit.	1.8	3.2
S05	F	History	3.5	3.6
S06	M	Physics	2.9	3.9
S07	F	English Lit.	2.2	3.4
S08	M	History	4.1	3.7
S09	M	Mathematics	3.8	3.3
S10	F	English Lit.	1.5	3.5
S11	M	History	3.2	3.8

Note. Self-efficacy baseline scores represent means on the Academic Self-Efficacy Scale (10-50 scale; presented here in standardized form 1-5 for readability).

4.2 Quantitative Findings: Weekly Effectiveness Ratings

Table 2. Mean Effectiveness Ratings by AI Learning Mode (1–5 Scale)

AI Mode	M	SD	n Observations	Interpretation
AI-Tutor	4.27	0.65	77	High effectiveness
AI-Student	4.18	0.72	79	High effectiveness
AI-Simulator	4.45	0.52	86	Highest effectiveness

Note. All modes achieved ratings >= 4.0 (Agree), indicating overall positive student perceptions. Repeated-measures ANOVA showed a significant effect of AI mode on ratings: F(2, 20) = 5.18, p = .017, partial eta-squared = .34. Post-hoc Tukey HSD tests revealed AI-Simulator significantly outperformed AI-Tutor (p = .019) and AI-Student (p = .031).

4.3 Pre/Post Self-Efficacy Outcomes

Table 3. Pre/Post Academic Self-Efficacy (n=11)

Dimension	Pre (M, SD)	Post (M, SD)	t(10)	p	Cohen's d
Problem-Solving	3.45 (0.89)	4.12 (0.68)	3.84	.003**	1.16
Persistence & Effort	3.62 (0.71)	4.03 (0.74)	2.15	.054	0.65
Intellectual Risk-Taking	3.28 (0.94)	3.91 (0.82)	2.91	.016*	0.88
Overall Self-Efficacy	3.54 (0.65)	4.09 (0.57)	3.52	.005**	1.06

*Note. **p < .01, *p < .05. Effect sizes (Cohen's d > 0.8) indicate large practical significance. The increase in problem-solving confidence (d = 1.16) and intellectual risk-taking (d = 0.88) suggest that the intervention fostered metacognitive confidence, particularly in domains where students practiced explaining ideas (AI-Student mode) and encountering complex scenarios (AI-Simulator).*

4.4 Qualitative Findings: Thematic Analysis

Focus group transcripts and written reflections yielded 287 coded segments, synthesized into four primary themes. Frequencies are presented as raw counts and percentages of total coded content.

Table 4. Thematic Analysis: Major Themes, Sub-Themes, and Frequencies

Theme	Sub-Theme	n	% of Total	Representative Quote
Agency & Ownership (40%)	Student control over pacing & direction	47	16.4%	I could ask follow-up questions at my pace
	Reduced passive reception	38	13.2%	Teaching the AI forced me to think, not just listen
	Personalization & responsiveness	30	10.5%	It adapted to what I did not understand
Scaffolding Depth & Adaptivity (35%)	Graduated complexity & explanation levels	52	18.1%	It could explain concepts simply or deeply
	Strategic questioning & Socratic prompts	33	11.5%	The "what if" questions made me think differently
	Timely, relevant feedback	26	9.1%	I got instant feedback without waiting for teacher
Affective Dimensions (22%)	Reduced anxiety & pressure	35	12.2%	No judgment; I could make mistakes safely
	Increased confidence & motivation	28	9.8%	I felt more capable after explaining clearly
Limitations & Concerns (18%)	Artificial feedback & lack of true understanding	29	10.1%	AI praised me even when my explanation was incomplete
	Risk of overconfidence / false mastery	19	6.6%	I think I understood more than I actually did
	Occasional hallucinations / errors in AI	11	3.8%	ChatGPT made a factual error and I believed it at first

Note. Theme percentages do not sum to 100% as some coded segments were assigned to multiple themes. Representative quotes have been lightly edited for clarity and concision while preserving original meaning.

4.5 Detailed Thematic Exploration

4.5.1 Theme 1: Agency and Ownership (40% of coding)

The most prevalent finding was students' sense of control and agency when engaging with AI. In AI-Tutor mode, students appreciated the ability to request re-explanations without social pressure ("I can ask the AI 'What do you mean?' five times without feeling stupid," S04). In AI-Student mode, the ability to teach positioned students as knowledge-producers rather than passive consumers: "Teaching the AI forced me to organize my thoughts and spot gaps I didn't know I had" (S07). In AI-Simulator mode, students valued the freedom to pose novel "What if?" scenarios: "I was driving the conversation, not following a script" (S01). This agency aligns with self-determination theory (Ryan & Deci, 2000), which posits that autonomy (along with competence and relatedness) is essential for intrinsic motivation and deep learning.

4.5.2 Theme 2: Scaffolding Depth and Adaptivity (35% of coding)

Students consistently highlighted AI's capacity to adjust explanation granularity and scaffold complexity. Unlike fixed textbook explanations, ChatGPT could respond to requests like "Explain this at a 9th-grade level" or "Give me the mathematical proof." One student noted: "A book explains something one way. ChatGPT can explain it three different ways depending on what you ask" (S03). In AI-Simulator mode, the ability to incrementally increase scenario complexity - posing a simple scenario, then adding constraints, then asking for optimization - mimicked the graduated release of responsibility model. However, students also noted that without teacher direction, this flexibility could become disorienting: "Too many options sometimes made me lose focus" (S09).

The capacity for strategic questioning emerged in AI-Simulator mode. When prompted appropriately, ChatGPT posed questions that challenged assumptions: "It asked me 'Have you considered X?' and I had never thought of it" (S06). This mirrors Socratic dialogue but with a critical caveat: the questions are not truly diagnostic of student understanding but rather pattern-matches from training data. Nevertheless, students reported productive intellectual struggle in response to these prompts.

4.5.3 Theme 3: Affective Dimensions (22% of coding)

A striking finding was the affective benefits, particularly reduced anxiety. Students reported lower pressure to perform when interacting with AI compared to in-class participation: "I feel less judged by ChatGPT than by the teacher asking me" (S10). This safe space enabled experimentation and risk-taking, particularly for students who experience math or public-speaking anxiety. Over the 6-week period, students' willingness to pose "silly" or half-formed questions increased, suggesting growing comfort with intellectual exploration.

Additionally, students reported affective gains associated with successfully explaining ideas to the AI. Bandura's (1986) concept of mastery experience posits that accomplishing a challenging task builds self-efficacy. When students explained a concept to ChatGPT and received confirmatory feedback or productive questions, they experienced what one student described as "proof that I really do understand this" (S05). This aligns with our quantitative finding that problem-solving self-efficacy increased significantly ($t(10) = 3.84$, $p = .003$, $d = 1.16$).

4.5.4 Theme 4: Limitations and Concerns (18% of coding)

Despite predominantly positive responses, students and teachers identified critical limitations. The most serious was the risk of false confidence. When ChatGPT provided positive feedback, students did not always scrutinize the accuracy of their own explanations: "I thought I explained photosynthesis perfectly because ChatGPT said 'great explanation!' But actually my description was incomplete" (S08). This directly instantiates the Dunning-Kruger effect: AI-provided validation without genuine diagnostic insight can inflate perceived competence.

A second concern was AI hallucinations and factual errors. In three instances documented in reflections, ChatGPT provided incorrect historical dates, misquoted literary passages, or offered mathematically flawed explanations. While students often caught these errors, the phenomenon raises questions about the reliability of AI feedback as a substitute for human expertise. As one student reflected: "How am I supposed to know when ChatGPT is confident but wrong?" (S11). This underscores the necessity of teacher oversight and the irreplaceability of expert judgment.

Finally, students noted that the open-endedness of AI interaction, while empowering, sometimes lacked the structured scaffolding of guided instruction. One student observed: "I could ask ChatGPT anything, but nobody told me *what* to ask" (S02). This highlights the necessity of pedagogical design: even with powerful tools, student learning benefits from clear learning objectives and strategic prompting architectures rather than ad-hoc exploration.

4.6 Learning Analytics and Engagement

Learning analytics data showed that students spent an average of M = 18.4 minutes (SD = 4.2) per AI interaction session. AI-Simulator mode generated the longest average session duration (M = 22.1 min, SD = 5.1) and the highest frequency of follow-up questions (M = 3.8 per session, SD = 1.4), consistent with students' qualitative reports of deeper engagement. AI-Tutor sessions were briefer (M = 14.3 min, SD = 3.9), suggesting students asked questions and received answers with less dialogue elaboration. This behavioral pattern aligns with quantitative effectiveness ratings (AI-Simulator > AI-Tutor).

5. Discussion

5.1 Interpretation of Quantitative Findings

Our quantitative findings support the hypothesis that AI learning mode significantly influences perceived effectiveness. The AI-Simulator mode's superiority (M = 4.45 vs. 4.27 for Tutor) may reflect the cognitive activation required by complex scenarios. When students posed novel problems to ChatGPT and encountered follow-up questions, they engaged in higher-order thinking (Bloom, 1956) - synthesizing knowledge, applying concepts, and evaluating outcomes. This aligns with the classical finding that active learning (d = 0.55) outperforms passive instruction; the effect size in our study, though modest in absolute terms, is noteworthy given the short intervention duration.

The paired-samples t-test results for pre/post self-efficacy are particularly compelling. Problem-solving self-efficacy increased by 0.67 points (on a 5-point scale), with a large effect size (d = 1.16). This magnitude exceeds typical effect sizes for brief interventions (Cohen, 1992) and suggests that the intervention - particularly AI-Student and AI-Simulator modes - fostered genuine confidence gains rather than transient effects. Importantly, the improvement in intellectual risk-taking (d = 0.88) suggests students became more willing to engage with challenging, ambiguous problems - a metacognitive shift aligned with growth mindset and intrinsic motivation (Dweck, 2006).

The non-significant trend for Persistence and Effort ($t(10) = 2.15$, $p = .054$) warrants discussion. While effect size was moderate (d = 0.65), the result did not reach conventional significance. One interpretation is that a 6-week intervention may be insufficient to alter deep-seated beliefs about effort and persistence, particularly for advanced sixth-formers who have typically internalized growth mindset through years of academic success. Alternatively, AI interaction may enhance short-term confidence without sustainably changing effort beliefs; longitudinal follow-up would clarify this.

5.2 Interpretation of Qualitative Findings and Theoretical Alignment

The four-theme structure emerging from qualitative analysis aligns remarkably with theoretical predictions. Theme 1 (Agency and Ownership, 40% of content) directly instantiates self-determination theory and Zimmerman's self-regulated learning framework. Students who exercise control over learning pace and direction tend to develop metacognitive awareness and sustained motivation (Ryan & Deci, 2000). This finding suggests that LLM-mediated learning's primary pedagogical value may lie not in content delivery but in empowering learner agency.

Theme 2 (Scaffolding Depth and Adaptivity, 35%) directly addresses Vygotsky's zone of proximal development and Hattie's tutoring mechanisms. Students' appreciation for graduated explanation levels and adaptive questioning reflects human tutors' capacity to calibrate support. Notably, the qualitative data revealed that this adaptivity required pedagogical guidance; without teacher-provided prompts or learning objectives, students sometimes engaged in surface-level exploration. This suggests that the notion of LLMs as fully autonomous pedagogical agents is problematic; rather, they function as powerful tools within teacher-designed learning architectures.

Theme 3 (Affective Dimensions, 22%) highlights an under-researched dimension of AI-mediated learning: emotional safety and confidence. Traditional research on tutoring (Bloom, 1984; Hattie, 2009) emphasizes cognitive mechanisms (diagnosis, feedback, scaffolding) but rarely addresses affect. Our finding that students reported reduced anxiety and increased confidence when learning with AI - stemming partly from the absence of social judgment - suggests a non-trivial advantage over peer or even teacher interaction in certain contexts. However, this benefit must be weighed against the relational and social dimensions of human tutoring, which support motivation, belonging, and social-emotional development (Lembke & Cohen, 2015).

Theme 4 (Limitations and Concerns, 18%) directly instantiates the Dunning-Kruger effect and raises critical questions about validity of AI feedback. Students who received confirmatory responses from ChatGPT without simultaneous diagnostic precision were at risk of illusory confidence. This phenomenon, often termed "AI as confidence amplifier," poses a genuine pedagogical risk. Conversely, we note that some students spontaneously employed metacognitive strategies to verify AI feedback ("How am I supposed to know when ChatGPT is confident but wrong?"), suggesting that explicit metacognitive training might mitigate this risk.

5.3 Alignment with Bloom's 2-Sigma Problem

Our findings provide nuanced evidence regarding Bloom's 2-sigma hypothesis. None of our AI modes produced the dramatic 2-sigma variance reduction that human tutoring achieved in Bloom's original research. However, effect sizes for self-efficacy and engagement, while smaller than true tutoring (d = 1.57), were substantial and cluster in the d = 0.8-1.2 range - comparable to high-quality active learning interventions (Freeman et al., 2014). This suggests that LLMs do not replicate human tutoring's full effectiveness but offer meaningful pedagogical affordances when combined with thoughtful instructional design.

Critically, Bloom's tutors provided one-to-one interaction, continuous diagnosis, and genuine understanding of student cognition. ChatGPT, by contrast, lacks intentionality, embodied understanding, and genuine responsiveness to the student's developing schema. Instead, ChatGPT pattern-matches to likely follow-up questions based on training data. For students navigating well-trodden conceptual terrain (standard curricula), this suffices; for novel or idiosyncratic confusion, human expertise remains essential. This distinction - between "sufficient AI approximation" for canonical content and "irreplaceable human expertise" for diagnosis of true learning breakdowns - should inform policy and practice.

5.4 Practical Implications for Secondary Education

Our findings support several concrete recommendations for educators:

(1) Prioritize active learning modes. AI-Simulator and AI-Student modes outperformed AI-Tutor mode, suggesting that pedagogies emphasizing learner agency and productive struggle should be prioritized over passive knowledge-reception modes. Teachers might design prompts positioning students as explainers: "Teach ChatGPT about [concept]" or "Ask ChatGPT to pose a challenging scenario in [domain]."

(2) Integrate metacognitive prompting. Teachers should explicitly train students to verify AI responses, articulate uncertainty, and recognize limitations. Phrases like "Why do you think ChatGPT answered that way?" or "How confident are you in that explanation, and why?" scaffold metacognitive monitoring.

(3) Leverage affective benefits cautiously. The anxiety reduction and safe-space affordances of AI are genuine benefits, particularly for students with learning anxiety or limited access to tutoring. However, these should complement rather than replace human relationships; the relational dimensions of education cannot be outsourced.

(4) Maintain teacher pedagogical authority. Rather than viewing LLMs as autonomous tutors, teachers should design AI interactions toward specific learning objectives. This aligns with Koehler and Mishra's (2009) TPACK framework: technology (LLM capability) must be deliberately integrated with pedagogical knowledge (how to scaffold learning) and content knowledge (what students need to learn).

5.5 Theoretical Framework: Pedagogically Purposeful AI Integration

Drawing on our findings and synthesis of literature, we propose a framework termed "Pedagogically Purposeful AI Integration" (PPAI). PPAI specifies that effective AI-mediated learning requires alignment among three dimensions:

(1) Learning architecture: Explicit design of learning objectives, scaffolding sequences, and assessment methods. AI should populate the ZPD, not replace it.

(2) AI mode selection: Deliberate choice of epistemic role (tutor, student, simulator, etc.) matched to learning objectives. Active modes (student, simulator) outperform passive modes (tutor) for deep learning and metacognitive development.

(3) Metacognitive scaffolding: Explicit instruction in evaluating AI responses, recognizing limitations, and monitoring understanding. Without this layer, learners risk the Dunning-Kruger effect - mistaking AI confidence for genuine mastery.

This framework positions AI as a tool within human-designed pedagogy rather than an autonomous agent. It aligns with principles of constructivism, active learning, and self-regulated learning while remaining pragmatic about LLM affordances and constraints.

5.6 Limitations

Several limitations warrant acknowledgment. First, our sample size (n = 11) is small, limiting statistical power and generalizability. While appropriate for an exploratory case study, larger-scale follow-up studies are needed. Second, the 6-week intervention duration may be insufficient to establish lasting effects; research on learning and self-efficacy suggests that metacognitive changes consolidate over longer timeframes (Zimmerman, 1989). Third, our participants were advanced sixth-formers at a selective school with above-average technology access and digital literacy. Results may not generalize to younger students, students with diverse learning profiles, or underserved populations. Fourth, we did not include a control group; a concurrent control condition receiving traditional instruction would strengthen causal inference.

Additionally, our use of self-report Likert scales and qualitative data, while rich, cannot directly measure learning outcomes (e.g., standardized assessment scores or learning gains in external examinations). Future research should employ objective academic achievement measures. Finally, we did not systematically examine the role of teacher TPACK or instructional design quality; some variation in outcomes likely stemmed from differences in how teachers implemented each AI mode.

5.7 Future Research Directions

Several avenues merit investigation. First, longitudinal studies tracking AI-mediated learning effects over a full academic year would clarify whether self-efficacy and metacognitive gains persist or are transient. Second, multi-site studies with diverse student populations would test generalizability and identify moderators (e.g., student age, prior achievement, school context). Third, research comparing AI modes directly against human tutoring and traditional instruction would enable more definitive claims about relative effectiveness. Fourth, investigation of teacher factors - TPACK development, pedagogical training, beliefs about AI - might illuminate which teacher characteristics optimize AI integration outcomes.

Finally, research on the interaction between AI and equity is urgent. While this study examined affective benefits (anxiety reduction, safe space for exploration), potential risks include widened achievement gaps (if AI tools are unequally accessible) and deskilling of teachers in disadvantaged schools. Qualitative and quantitative research examining these justice dimensions would inform equitable policy.

6. Conclusion

This mixed-methods case study examined the comparative effectiveness of three AI-mediated learning modes in a secondary school context. Quantitative findings revealed that AI-Simulator mode achieved the highest perceived effectiveness ratings and that the 6-week intervention produced significant gains in self-efficacy (particularly problem-solving, d = 1.16, and intellectual risk-taking, d = 0.88). Qualitative analysis identified four emergent themes: agency and ownership, scaffolding depth and adaptivity, affective benefits, and limitations regarding feedback validity. These findings suggest that LLMs offer meaningful pedagogical affordances, particularly when deployed in active learning modes emphasizing student agency and metacognitive monitoring.

However, our results also underscore that LLMs are not pedagogical panaceas. While they approximate some mechanisms of human tutoring (scaffolding, responsiveness), they lack the intentionality, genuine diagnosis, and relational dimensions that make human expertise irreplaceable. The risk of amplified confidence without genuine understanding - the AI-mediated Dunning-Kruger effect - remains a serious concern mitigatable only through deliberate metacognitive scaffolding.

We propose the framework of "Pedagogically Purposeful AI Integration" (PPAI), emphasizing that effective technology use requires intentional alignment of learning architecture, AI mode selection, and metacognitive support. Teachers are not disintermediated by LLMs but rather must develop sophisticated TPACK, designing learning experiences that leverage AI's affordances while maintaining human pedagogical authority and expertise.

As large language models become increasingly integrated into educational contexts, the critical work lies not in the technology itself but in the pedagogies we build around it. This case study contributes evidence-based guidance to that essential work, though the conversation is far from complete. Secondary educators, teacher educators, and policymakers must engage in sustained, empirically informed dialogue to ensure that AI augments human teaching rather than diminishing it, and enhances learning for all students rather than exacerbating existing inequities.

References

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Prentice-Hall.

Biesta, G. (2015). The beautiful risk of education. Routledge.

Bloom, B. S. (1984). The search for methods of group instruction as effective as one-to-one tutoring. Educational Leadership, 41(8), 4-17.

Bonwell, C. C., & Eison, J. A. (1991). Active learning: Creating excitement in the classroom. ASHE-ERIC Higher Education Report No. 1. George Washington University.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155-159.

Dweck, C. S. (2006). Mindset: The new psychology of success. Random House.

Educational Endowment Foundation (EEF). (2021). Tutoring: Rapid evidence assessment. EEF.

Ferla, J., Valcke, M., & Cai, Y. (2009). Academic self-efficacy and academic self-concept: Reconsidering structural relationships. Learning and Individual Differences, 19(4), 499-505.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 111(23), 8410-8415.

Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. Routledge.

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promises and implications. Center for Curriculum Redesign.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge (TPACK)? Journal of Education, 193(3), 13-19.

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

Kulik, C. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. Computers in Human Behavior, 7(1-2), 75-94.

Lembke, E. S., & Cohen, L. (2015). The role of human relationships in education: A literature review. School Psychology International, 36(2), 111-134.

Luckin, R., Cukurova, M., & Bouchier-Hayes, D. (2016). Designing educational technologies in the age of AI and data-driven systems. Journal of Learning Analytics and Development, 3(1), 17-35.

Mollick, E. R., & Mollick, L. (2023). Using AI to implement effective science teaching. Wharton University of Pennsylvania, working paper.

Molyneux, L. (2023). ChatGPT: The latest technological disruption. Technology Trends Today, 12(1), 22-31.

Pearson, P. D., & Gallagher, M. C. (1983). The instruction of reading comprehension. Contemporary Educational Psychology, 8(3), 317-344.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary Educational Psychology, 25(1), 54-67.

Selwyn, N. (2019). Artificial intelligence and the techno-politics of learning. Learning, Media and Technology, 44(3), 348-362.

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

Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). Sage Publications.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education - where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 39.

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. Journal of Educational Psychology, 81(3), 329-339.