IJCER
p-ISSN 2549-9947 | e-ISSN 2614-1426
doi: 10.20885/ijcer.vol9.iss2.art4
p-ISSN 2549-9947
e-ISSN 2614-1426

AI-Powered Chatbots in Science Education: Meta-Analysis of Pedagogical Integration and Student Achievement

Jermae B. Dizon [a,*], Maricar S. Prudente [b]

[a, b] *Department of Science Education, College of Education, De La Salle University- Manila, Philippines*

** Corresponding author: jermae_dizon@dlsu.edu.ph*

Received: June 4, 2025 ; Accepted: July 19, 2025; Published: October 25, 2025

ABSTRACT. The integration of artificial intelligence (AI) in education has accelerated, yet its pedagogical impact remains uneven and theoretically underexplored, particularly in science education. Existing studies often emphasize technical features or user satisfaction, with limited focus on how instructional design and learning context shape learning outcomes. This meta-analysis evaluated the effectiveness of AI-powered chatbots in improving student achievement in science education and identified key moderating factors influencing their impact. Using PRISMA guidelines, 26 empirical studies published between 2020 and 2024 were systematically reviewed and analyzed with a random-effects model. The overall effect size was statistically significant and moderate (Hedges' g=0.610, p< 0.001), suggesting that chatbot-supported instruction outperformed traditional methods in many cases. However, substantial heterogeneity was observed (I^2=96.58%), indicating that effectiveness varied significantly based on socio-economic context, subject area, pedagogical design, and learner experience. Chatbots were most effective in lower-middle-income countries and in subjects like computer science and natural sciences, especially when implemented through scaffolded or personalized learning strategies. Gains in engagement and satisfaction were common, while effects on self-efficacy and navigation were mixed. These findings challenge uniform assumptions about AI's role in education and call for theory-informed, context-sensitive integration strategies. Importantly, this study extends existing learning theories by showing that AI-driven dialogue systems act not merely as tools but as active mediators of both cognitive and affective processes. Future research should pursue longitudinal designs, hybrid human–AI teaching models, and ethical frameworks to guide equitable and sustainable implementation across educational contexts.

Keywords: artificial intelligence, chatbots, science education, learning outcomes, student achievement

INTRODUCTION

Artificial Intelligence (AI) is transforming the educational landscape by enabling more personalized, adaptive, and efficient learning experiences [1]. Among its most promising applications are AI-powered virtual assistants and chatbots, which are increasingly used in science education today. These tools aim to address persistent challenges in science teaching, such as providing individualized support, accommodating diverse learning paces, and delivering timely feedback, by simulating human-like interactions through natural language processing (NLP) [2,3].

Historically, AI in education began with systems focused on automating routine tasks and offering scalable tutoring solutions [4]. Early implementations, such as intelligent tutoring systems (ITS), provided structured, adaptive feedback to students [5]. Over the past decade, advancements in NLP and machine learning have enabled more dynamic AI tools such as ChatGPT, capable of real-time conversation and contextual understanding [6]. These tools represent a shift from static content delivery to interactive, dialogic learning environments. This evolution aligns with socio-constructivist and activity

theory perspectives, which emphasize guided interaction, contextual learning, and learner-tool mediation [7].

AI-powered chatbots are now used across various instructional models, including flipped classrooms, gamified learning, scaffolded instruction, and formative assessment [7,9]. Their impact on learning outcomes varies widely depending on how they are integrated into pedagogy and what learner contexts they serve [2,6]. When aligned with strong pedagogical frameworks and tailored to learner needs, chatbots can enhance engagement, understanding, and self-directed learning. However, poorly integrated or contextually misaligned implementations may limit their effectiveness.

Despite growing interest, many studies remain confined to specific contexts, most commonly in higher education which leaves gaps in our understanding of chatbot influence on primary and secondary learners [10,11]. Moreover, existing literature often emphasizes technical functionalities or user satisfaction without thoroughly examining instructional design, duration, or frequency of use as factors shaping learning outcomes [12,13]. These gaps present not only empirical but also theoretical challenges. Specifically, the pool o research lack a comprehensive synthesis that explains how chatbot-mediated learning interacts with core educational theories, such as cognitive load theory, scaffolding within Vygotsky's Zone of Proximal Development, or learner regulation in constructivist and activity theory frameworks.

This absence of synthesis restricts our ability to position chatbots meaningfully within or against existing learning paradigms. Without understanding how these tools mediate cognition, support or overload working memory, and either empower or constrain learner autonomy, we risk reducing AI to a technical novelty rather than recognizing its broader pedagogical and theoretical implications. A richer theoretical inquiry is needed to explore how chatbot use reshapes the cognitive, emotional, and social dimensions of science learning.

The objective of this research was to examine the effect of AI-powered chatbots on student achievement in science education. In particular, this study sought to answer two key questions: (1) What is the overall effect of chatbots on learning outcomes? and (2) What moderating factors, such as context, pedagogy, duration, or learner characteristics, influence their effectiveness? These questions guide the meta-analysis and help identify when and how AI tools are most beneficial in science classrooms. These questions allow us to move beyond a simple yes or no and instead ask: when, how, and for whom do chatbots work best?

RESEARCH METHODS

Research Design

This study employed a meta-analytic research design to quantitatively synthesize findings from 26 empirical studies on AI-powered chatbots in science education. Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, a random-effects model was used to account for variability across educational contexts. Effect sizes were calculated using Hedges' g to adjust for small sample bias. Moderator analyses were conducted to explore how factors such as pedagogy, context, duration, and user experience influenced learning outcomes.

Criteria for Inclusion and Exclusion

To ensure methodological rigor and relevance to the study objectives, studies had to meet the following inclusion criteria: (1) focus specifically on AI-powered chatbots or virtual assistants used in educational contexts; (2) evaluate learning outcomes using empirical data from quantitative or mixed-methods research designs; and (3) be published in English in peer-reviewed journals or reputable conference proceedings between 2020 and 2024.

The exclusion criteria included: (1) studies describing chatbot development without assessing learning impact; (2) purely anecdotal or qualitative accounts lacking measurable outcomes; (3) studies unrelated to science or STEM education; (4) duplicate publications or non-English texts; and (5) insufficient methodological transparency.

This selection strategy may introduce language and publication bias, as studies in other languages or those not indexed in the selected databases were excluded. This limitation is acknowledged as a potential constraint on the generalizability of findings, especially from underrepresented regions.

Data Collection

Data were collected systematically using Publish or Perish 8 software to search four major academic databases: Scopus, Google Scholar, Crossref, and Semantic Scholar. Boolean search strings (e.g., "AI chatbots AND science education AND learning outcomes") were applied, and results were filtered to

include studies published in English between 2020 and 2024. Duplicate entries were removed using Mendeley, and remaining records were screened by title and abstract based on predefined inclusion criteria. Full-text reviews were then conducted to determine eligibility. This process followed PRISMA guidelines to ensure transparency and replicability [16].

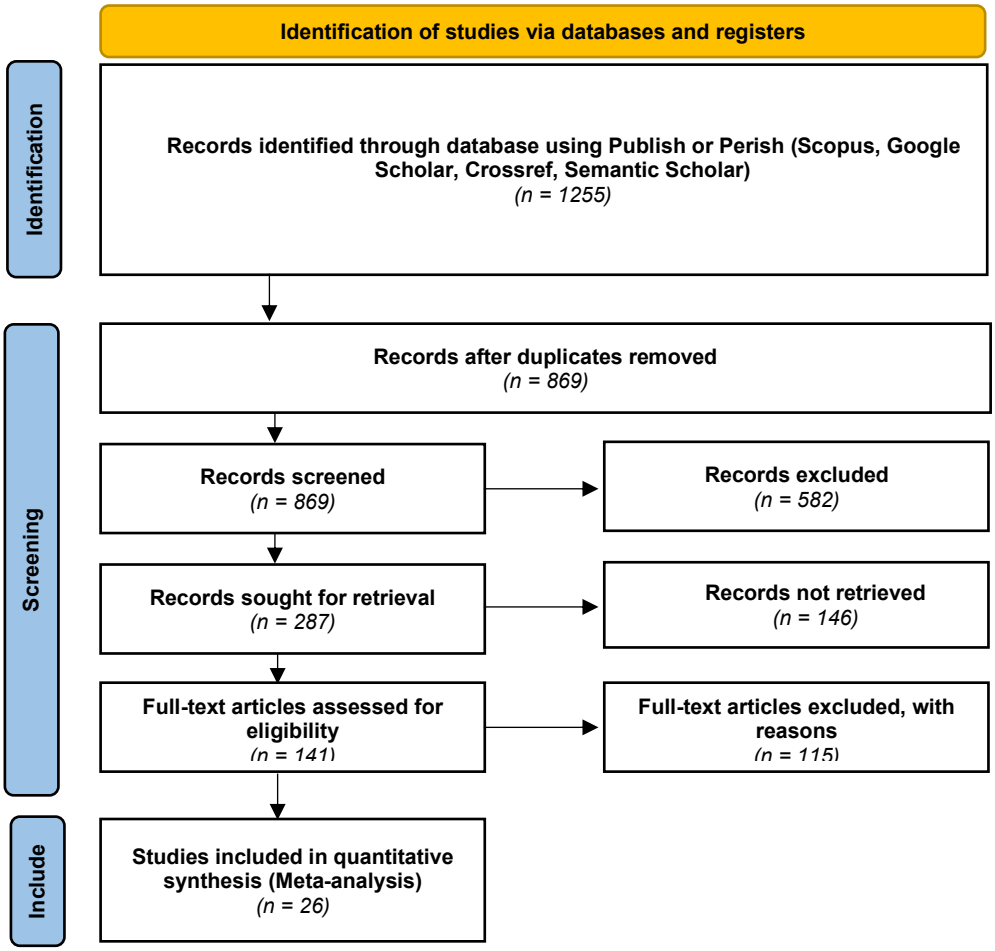


FIGURE 1. Flow diagram of article selection

Risk of Bias Assessment

To assess the internal validity of included studies, this review applied the Cochrane Risk of Bias tool [18]. Each study was evaluated across key domains, including selection bias, performance bias, detection bias, attrition bias, and reporting bias. Ratings were assigned as "low," "unclear," or "high" risk for each domain. While no studies were excluded solely based on high risk, bias levels were recorded and considered during interpretation and subgroup analysis. Sensitivity analysis was conducted to check whether high-risk studies disproportionately influenced the overall effect size.

Coding Validation

To ensure reliable and consistent data extraction, a coding manual was developed prior to the review process. The first author performed the initial coding of study features, including sample size, chatbot design, educational level, outcome measures, and pedagogical approach. Validation was conducted by a master teacher and two science teachers, minimizing subjectivity and enhancing coding consistency. These reviewers independently cross-checked the coded data, and inter-coder agreement was assessed through percent agreement. Any discrepancies were resolved through discussion until full consensus was reached. This collaborative process helped ensure the accuracy and reliability of the categorized variables used in the moderator analysis.

Data Analysis Technique

Data analysis was performed using Comprehensive Meta-Analysis for Windows, Version 3.0 (Biostat, Inc., Englewood, NJ, USA). Hedges' g was calculated for all included studies to standardize

mean differences in learning outcomes between chatbot and control groups. A random-effects model was used due to the anticipated heterogeneity among studies. Homogeneity was assessed using Q statistics and I^2, with I^2 values exceeding 50% interpreted as substantial heterogeneity.

To assess publication bias, the Classic Fail-Safe N test was used. Additionally, a funnel plot was generated to visually examine symmetry, and Egger's regression test was performed to detect potential bias due to small-study effects. These results are reported in the Results section and used to gauge the reliability of the overall effect size.

RESULT AND DISCUSSION

Current State of AI Chatbots in Science Education

Table 1 summarizes the 26 studies included in this meta-analysis, representing a total of 3,287 participants, 1,845 in experimental groups (56.13%) and 1,442 in control groups (43.87%). These studies, published between 2020 and 2024, reflect a growing body of work examining the use of AI-powered chatbots in science education. The sharp rise in studies published in 2024 (57.69%, n = 15) is likely driven by advances in free, publicly accessible tools such as ChatGPT, as well as increasing institutional interest in scalable digital learning solutions [4].

While the rise in research is encouraging, it remains heavily skewed toward high-income (46.15%) and upper-middle-income countries (38.46%), with fewer studies from lower-middle-income regions (15.38%). Interestingly, the greatest positive effect sizes were found in lower-middle-income countries, where chatbots were leveraged to compensate for chronic under-resourcing [14]. This finding supports existing notions of the digital divide and extends equity-focused educational theories by showing how context-sensitive AI interventions can function as tools for leveling educational access disparities [13]. This invites a critical pedagogical lens, in which chatbots are not just technical innovations but redistributive supports that temporarily mitigate inequities in teacher presence and material access [22, 23].

In terms of chatbot design, the majority (80.77%) utilized advanced natural language processing (NLP) with adaptive learning features [10]. These were often deployed in higher education settings where learners were more autonomous and technical infrastructure supported scalability. Interactive chatbots (19.23%) that incorporated videos and diagrams were found to be more engaging but required greater development effort, making them less scalable [24, 9].

Subject-wise, most studies focused on computer science and technology (50%), followed by natural sciences (26.92%), medical and health sciences (15.38%), and mathematics (11.54%). The lower and even negative effects in health sciences likely reflect the situated and embodied nature of clinical learning, which chatbots, primarily text-based tools, currently fail to replicate. Theoretical frameworks such as situated learning theory emphasize the need for context-rich, hands-on interaction in knowledge development [4, 48], which suggests that current chatbot designs may not align well with the competencies required in medical education. Future iterations may need to integrate simulation or multi-modal feedback systems [40] to bridge this gap.

Finally, most studies (73.08%) were conducted in higher education, with far fewer at the secondary (23.08%) and primary (3.85%) levels [4]. This reflects greater flexibility and digital capacity in universities, but also highlights a critical research gap. Younger learners, especially those developing early science literacy, stand to benefit from conversational AI, but more studies are needed to explore developmentally appropriate chatbot integration at foundational levels [11, 13].

Implementation strategies used in AI Chatbots

The reviewed studies implemented a variety of strategies across pedagogical models, instructional integration, duration and frequency of use, user experience, assessment types, and noncognitive outcomes.

Scaffolded learning was the most common strategy (30.77%), emphasizing guided instruction to support cognitive development and problem-solving skills [25, 12]. This supports Vygotsky's Zone of Proximal Development (ZPD) [26] but also expands its interpretation. In AI-enhanced learning environments, the "more knowledgeable other" is not a person but an algorithm capable of dynamic adaptation. These findings suggest that chatbots can serve as asynchronous cognitive mediators, a theoretical extension of Vygotsky's original human-to-human scaffolding model into the realm of AI-human interaction.

TABLE 1. Summary of included studies

Author	Year	Country	Category	Chatbot Design	AI Tool Used	Subject Area	Educational Level	N	EG (CG)
Al Kahf et al. [30]	2023	France	HI	Interactive	Chatprogress	Medical	HE	356	104 (252)
Alneyadi and Wardat [8]	2023	UAE	HI	Advanced	ChatGPT	Physics	SE	122	58 (64)
Beltozar-Clemente et al. [10]	2024	Peru	UMI	Advanced	ChatGPT	Physics	HE	188	98 (90)
Bhatia et al. [34]	2024	India	LMI	Advanced	ChatGPT	Anatomy	HE	100	50 (50)
Challapalli and Leddo [31]	2024	USA	HI	Advanced	ChatGPT	Calculus	SE	30	15 (15)
Çiçek et al. [28]	2024	Turkey	UMI	Advanced	ChatGPT	Anatomy	HE	115	56 (59)
Ekukinam et al. [14]	2024	Nigeria	LMI	Advanced	Chatbot AI	Biology	SE	600	359 (241)
Essel et al. [20]	2022	Ghana	LMI	Advanced	KNUSTbot	Programming	HE	68	34 (34)
Essel et al. [25]	2024	Ghana	LMI	Advanced	VoiceBot	Multimedia Programming	HE	65	33 (32)
Fırat and Kuleli [15]	2024	Turkey	UMI	Advanced	ChatGPT	Programming	HE	374	223 (151)
Graefen and Fazali [26]	2023	USA	HI	Advanced	ChatGPT	Public Health Education	HE	200	100 (100)
Hakiki et al. [35]	2023	Indonesia	UMI	Advanced	ChatGPT	Technology education	HE	62	31 (31)
Huesca et al. [36]	2024	Mexico	UMI	Advanced	ChatGPT	Programming	HE	356	140 (114)
Koć-Januchta et al. [37]	2020	Sweden	HI	Advanced	AI-enriched digital book	Biology	HE	16	6 (10)
Kumar et al. [13]	2024	Canada	HI	Advanced	LLM-based chatbot assistant	Computer Science	HE	218	145 (73)
Kusuma et al. [38]	2024	Indonesia	UMI	Advanced	AAIL MiClima Chatbot	Physics	SE	64	32 (32)
Li [6]	2023	China	UMI	Advanced	ChatGPT-FLGA	Educational Technology	HE	81	42 (39)

Author	Year	Country	Category	Chatbot Design	AI Tool Used	Subject Area	Educational Level	N	EG (CG)
Lin and Ye [39]	2023	Taiwan	HI	Interactive	Biology chatbot system	Biology	SE	34	17 (17)
Liu et al. [29]	2024	UK	HI	Advanced	ChatGPT	Educational Technology	HE	30	15 (15)
Mellado-Silva et al. [40]	2020	Chile	HI	Advanced	Tribuchat	Taxation	HE	50	26 (24)
Pardos and Bhandari [31]	2024	USA	HI	Advanced	ChatGPT	Mathematics	HE	188	98 (90)
Topal et al. [11]	2021	Turkey	UMI	Interactive	Dialogflow	General Science	PE	41	20 (21)
Wu et al. [42]	2023	Taiwan	HI	Advanced	Peer Assessment with ChatGPT	Programming	HE	61	31 (30)
Xu et al. [12]	2024	China	UMI	Interactive	Digital game-based AI chatbot	Information Technology	SE	77	38 (39)
Xue et al. [13]	2024	USA	HI	Advanced	ChatGPT	Computer Science	HE	48	23 (33)
Yin et al. [19]	2020	China	UMI	Interactive	Chatbot-based micro-learning	Basic Computer Science	HE	99	51 (48)

Legend: IH, High-income; UMI, Upper-middle income; LMI, Lower-middle income; HE, Higher Education; SE, Secondary Education; PE, Primary Education; EG, Experimental group sample size; CG, Control group sample size

TABLE 2. Summary of Implementation Strategies

Author	Main Pedagogical Approach	Instructional Integration	Duration of Intervention	Frequency	User Experience	Assessment Type	Other Outcome Measured
Al Kahf et al. [30]	Gamified Learning	Supplementary	Long-term	Varying	Novice	HOTS (Medical reasoning)	Satisfaction and Perception
Alneyadi and Wardat [8]	Personalized Learning	Supplementary	Very Short-Term	Varying	Novice	HOTS (Mathematical reasoning test)	Satisfaction and Perception
Beltozar-Clemente et al. [10]	Gamified Learning	Supplementary	Long-term	Weekly	Not specified	HOTS (Problems and practical application)	Multiple (Satisfaction and Perception, Self-Efficacy

Author	Main Pedagogical Approach	Instructional Integration	Duration of Intervention	Frequency	User Experience	Assessment Type	Other Outcome Measured
							and Learning Performance)
Bhatia et al. [34]	Scaffolded Learning	Supplementary	Single Session	Single Session	Novice	LOTS (Anatomical landmarks)	No Measured Outcome
Challapalli and Leddo [31]	Personalized Learning	Primary tool	Single Session	Not specified	Novice	LOTS (Procedural fluency to solve problems)	No Measured Outcome
Çiçek et al. [28]	Inquiry-based Learning	Supplementary	Very Short-Term	Daily	Novice	HOTS (Clinical reasoning skills)	No Measured Outcome
Ekukinam et al. [14]	Inquiry-based Learning	Primary tool	Not specified	Varying	Novice	HOTS (Biology Performance Test)	Equity and Demographics
Essel et al. [20]	Scaffolded Learning	Supplementary	Long-term	Varying	Experienced	MIX (Objective-type questions and practical examination)	Satisfaction and Perception
Essel et al. [25]	Scaffolded Learning	Supplementary	Medium-Term	Daily	Novice	MIX (Objective-type questions and practical examination)	Engagement and Motivation
Fırat and Kuleli [15]	Autonomous Learning (Self-Directed)	Primary tool	Single Session	Single Session	Varying	LOTS (JavaScript functionality test)	Usefulness and Navigation
Graefen and Fazali [27]	Personalized Learning	Primary tool	Not specified	Varying	Novice	LOTS (Knowledge and understanding of medical terminology)	Usefulness and Navigation
Hakiki et al. [35]	Personalized Learning	Primary tool	Not specified	Not specified	Novice	LOTS (Knowledge-Based test)	Engagement and Motivation
Huesca et al. [36]	Flipped Classroom	Supplementary	Medium-Term	Weekly	Novice	MIX (Objective-type questions and conceptual and code analysis)	No Measured Outcome

Author	Main Pedagogical Approach	Instructional Integration	Duration of Intervention	Frequency	User Experience	Assessment Type	Other Outcome Measured
Koć-Januchta et al. [37]	Inquiry-based Learning	Supplementary	Very Short-Term	Varying	Novice	MIX (Retention and comprehension)	Engagement and Motivation
Kumar et al. [13]	Autonomous Learning (Self-Regulated)	Supplementary	Very Short-Term	Varying	Varying	HOTS (Metacognition, self-reflection, and conceptual understanding)	Self-Efficacy and Learning Performance
Kusuma et al. [38]	Autonomous Learning (Independent)	Primary tool	Not specified	Varying	Novice	LOTS (Knowledge and understanding)	No Measured Outcome
Li [6]	Flipped Classroom	Supplementary	Very Short-Term	Daily	Novice	HOTS (Project performance)	Multiple (Self-Efficacy and Learning Performance, Engagement and Motivation)
Lin and Ye [39]	Scaffolded Learning	Supplementary	Very short-term	Varying	Novice	LOTS (Knowledge-Based Assessments)	Multiple (Engagement and Motivation, Satisfaction and Perception)
Liu et al. [29]	Online Collaborative learning	Supplementary	Not specified	Not specified	Novice	HOTS (Knowledge activation, OCL performance, and critical thinking)	Self-Efficacy and Learning Performance
Mellado-Silva et al. [40]	Scaffolded Learning	Supplementary	Not specified	Varying	Novice	LOTS (Recognize, identify, procedural application)	No Measured Outcome
Pardos and Bhandari [31]	Scaffolded Learning	Supplementary	Single Session	Varying	Novice	LOTS (Procedural application)	No Measured Outcome
Topal et al. [11]	Inquiry-Based Learning	Supplementary	Very Short-Term	Daily	Novice	LOTS (Concept mastery)	Engagement and Motivation

Author	Main Pedagogical Approach	Instructional Integration	Duration of Intervention	Frequency	User Experience	Assessment Type	Other Outcome Measured
Wu et al. [42]	Scaffolded Learning	Supplementary	Short-term	Not specified	Experienced	HOTS (Critical thinking, problem-solving, and creativity)	Satisfaction and Perception
Xu et al. [12]	Gamified Learning	Primary tool	Single Session	Single Session	Experienced	HOTS (Problem-solving, computational thinking, and creativity)	Engagement and Motivation
Xue et al. [13]	Scaffolded Learning	Supplementary	Single Session	Varying	Novice	MIX (Comprehension and implementation of Object-Oriented Programming)	Satisfaction and Perception
Yin et al. [19]	Micro-learning	Primary tool	Single Session	Single Session	Novice	LOTS (Students' mastery of knowledge)	Engagement and Motivation

Legend: HOTS=Higher order thinking skills; LOTS=Lower order thinking skills; MIX= HOTS and LOTS

Personalized and inquiry-based learning, each used in 15.38% of studies [8, 27, 28], showed strong alignment with constructivist learning theory, where learners build knowledge through exploration and self-paced progression. Gamified and autonomous learning (11.54% each), flipped learning (7.69%) [6], and online collaborative learning and microlearning (3.85% each) [19, 29] were less common, though they offer flexible formats. Microlearning in particular showed limited effectiveness, which may reflect cognitive load theory limitations—i.e., fragmented learning chunks may not support deep schema integration [46]. Most studies (69.23%) used chatbots as supplementary tools [20, 29, 30], while 30.77% employed them as primary instructional agents [12, 14, 31]. Supplementary roles allowed for reinforcement and practice, whereas primary use reflected a push toward autonomous learning environment. These findings support distributed cognition theory, suggesting that optimal outcomes occur when AI tools are embedded in blended learning systems that distribute cognitive responsibility between humans and machines [32, 33]. Most implementations were short-term (5 to 8 weeks), or single-session (one-time intervention lasting a few hours or less), while only 23.1% (n = 6) used long-t erm durations of 13–16 weeks [44]. This limits the ability to assess the impact of sustained feedback loops, which are crucial for cognitive apprenticeship and deep learning. Long-term exposure is more likely to support the gradual development of transferable knowledge and self-regulation skills [45]. Flexible usage was most common (42.3%, n = 11), while daily and single-session uses were reported in 15.4% of studies each. Only one study used weekly interaction. These trends reinforce the value of self-regulated learning (SRL) frameworks, in which learners choose when and how often to engage with digital tools [4, 47]. However, findings also suggest that structured engagement routines may be necessary for students unfamiliar with AI tools [46].

The majority of studies (76.92%) involved novice users, with fewer involving experienced learners (11.54%) or varied experience levels (7.69%). Novices still benefited, but experienced users showed stronger outcomes—likely due to decreased extraneous cognitive load and more efficient interaction with the tool [45, 47]. This finding aligns with cognitive tool theory, which emphasizes the importance of fluency in tool use for deeper learning [32].

Assessment approaches included lower-order thinking skills (LOTS, 42.3%) [4], higher-order thinking skills (HOTS, 38.5%), and mixed methods (19.2%) [47, 32]. The strongest effects were associated with HOTS and mixed assessments, reinforcing the role of chatbots in supporting critical thinking, conceptual understanding, and transfer—hallmarks of higher-level cognition [13].

Beyond academic performance, several studies reported improvements in student engagement and motivation (30.77%, n = 8), satisfaction (26.92%), self-efficacy (15.38%), and navigation usefulness (7.69%) [25, 12, 30, 42, 14]. One study (3.85%) reported enhanced perceptions of equity, particularly in under-resourced contexts [14]. These findings suggest that chatbots may serve not only as cognitive tools but also as affective and motivational supports, particularly in inclusive education contexts. However, the limited effect on self-efficacy highlights a need for better design of feedback mechanisms that affirm learner competence and encourage persistence [48].

Overall Effect Size of Using AI Chatbots

Figure 2 presents the forest plot of effect sizes of individual studies using Hedges' g, with a pooled effect size of 0.610 (95% CI: 0.550–0.670, Z = 19.929, p < 0.001), indicating a statistically significant moderate positive effect of AI-powered chatbots on learning outcomes. This suggests that students using chatbots performed better than those in control groups. According to Cohen's guidelines, this effect is meaningful and aligns with previous findings on the benefits of AI in promoting engagement and self-regulated learning [4, 13].

Publication Bias Analysis

To ensure the internal quality and validity of the synthesized evidence, this study included both a publication bias analysis and an explicit risk of bias assessment. The Cochrane Risk of Bias tool [18] was used to evaluate each of the 26 included studies across multiple domains such as selection bias, performance bias, detection bias, attrition bias, and reporting bias. Studies were rated as low, unclear, or high risk in each domain. While no studies were excluded based on these ratings, the results were considered during subgroup interpretation and sensitivity analysis. The overall pattern of findings did not suggest that studies with higher risk of bias disproportionately influenced the effect size, supporting the reliability of the results.

For publication bias, the Classic Fail-Safe N test estimated that 2,319 additional studies with null results would be needed to raise the overall p-value above the significance threshold of 0.05. This high threshold indicates that the observed effect size (Hedges' g = 0.610) is highly robust and unlikely to be nullified by unpublished negative results [13]. In addition, a funnel plot of standard error by Hedges' g was generated (Figure 3) to visually inspect potential asymmetry. The plot showed mild asymmetry, suggesting that smaller studies with non-significant or negative effects may be underrepresented. However, Egger's regression test produced an intercept of 1.372 (p = 0.467), indicating no statistically significant evidence of small-study effects

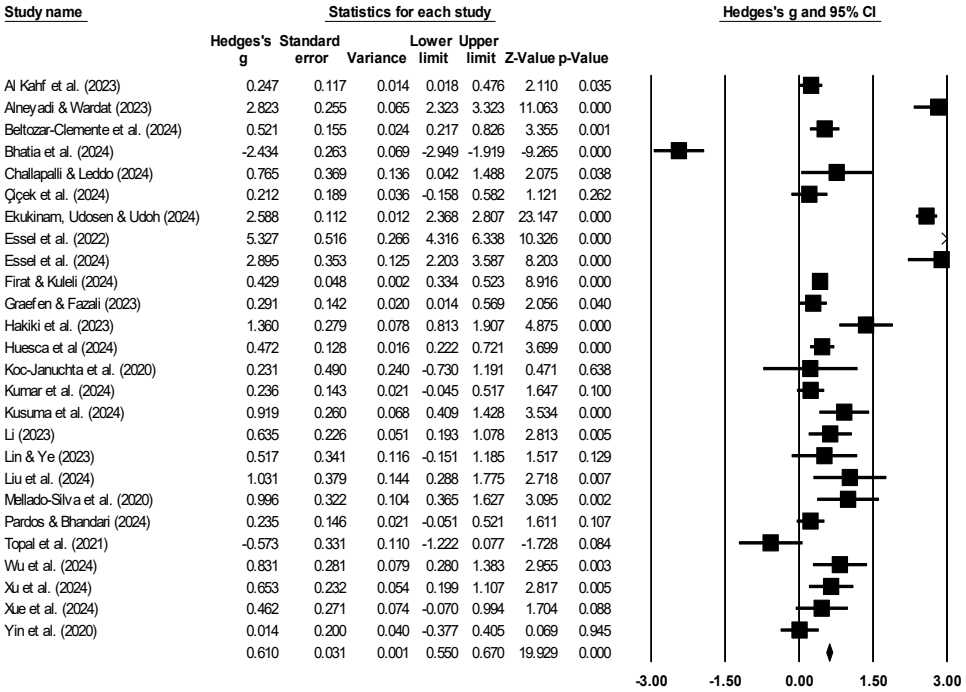
TABLE 3. Classic Fail-Safe N

Metric	Value
Z-value for observed studies	18.61227
P-value for observed studies	0
Alpha	0.05
Tails	2
Z for alpha	1.95996
Number of observed studies	26
Number of missing studies that would bring p-value to > alpha	2319

Heterogeneity Results

Table 4 presents the heterogeneity analysis, revealing significant variation in effect sizes across studies. The Q-value (730.963, p < 0.001) and I^2 statistic (96.58%) indicate considerable heterogeneity, suggesting that nearly all observed differences stem from real variations in study characteristics rather than random error [13]. These results imply that the impact of AI-powered

chatbots varies significantly across contexts and implementations. Therefore, further subgroup analyses are warranted to explore moderators that influence chatbot effectiveness and provide a more nuanced understanding of their impact in science education.



Study name	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Al Kahf et al. (2023)	0.247	0.117	0.014	0.018	0.476	2.110	0.035
Alneyadi & Wardat (2023)	2.823	0.255	0.065	2.323	3.323	11.063	0.000
Beltozar-Clemente et al. (2024)	0.521	0.155	0.024	0.217	0.826	3.355	0.001
Bhatia et al. (2024)	-2.434	0.263	0.069	-2.949	-1.919	-9.265	0.000
Challapalli & Leddo (2024)	0.765	0.369	0.136	0.042	1.488	2.075	0.038
Çiçek et al. (2024)	0.212	0.189	0.036	-0.158	0.582	1.121	0.262
Ekukinam, Udosen & Udoh (2024)	2.588	0.112	0.012	2.368	2.807	23.147	0.000
Essel et al. (2022)	5.327	0.516	0.266	4.316	6.338	10.326	0.000
Essel et al. (2024)	2.895	0.353	0.125	2.203	3.587	8.203	0.000
Firat & Kuleli (2024)	0.429	0.048	0.002	0.334	0.523	8.916	0.000
Graefen & Fazali (2023)	0.291	0.142	0.020	0.014	0.569	2.056	0.040
Hakiki et al. (2023)	1.360	0.279	0.078	0.813	1.907	4.875	0.000
Huesca et al (2024)	0.472	0.128	0.016	0.222	0.721	3.699	0.000
Koc-Januchta et al. (2020)	0.231	0.490	0.240	-0.730	1.191	0.471	0.638
Kumar et al. (2024)	0.236	0.143	0.021	-0.045	0.517	1.647	0.100
Kusuma et al. (2024)	0.919	0.260	0.068	0.409	1.428	3.534	0.000
Li (2023)	0.635	0.226	0.051	0.193	1.078	2.813	0.005
Lin & Ye (2023)	0.517	0.341	0.116	-0.151	1.185	1.517	0.129
Liu et al. (2024)	1.031	0.379	0.144	0.288	1.775	2.718	0.007
Mellado-Silva et al. (2020)	0.996	0.322	0.104	0.365	1.627	3.095	0.002
Pardos & Bhandari (2024)	0.235	0.146	0.021	-0.051	0.521	1.611	0.107
Topal et al. (2021)	-0.573	0.331	0.110	-1.222	0.077	-1.728	0.084
Wu et al. (2024)	0.831	0.281	0.079	0.280	1.383	2.955	0.003
Xu et al. (2024)	0.653	0.232	0.054	0.199	1.107	2.817	0.005
Xue et al. (2024)	0.462	0.271	0.074	-0.070	0.994	1.704	0.088
Yin et al. (2020)	0.014	0.200	0.040	-0.377	0.405	0.069	0.945
	0.610	0.031	0.001	0.550	0.670	19.929	0.000

FIGURE 2. Forest plot of effect sizes

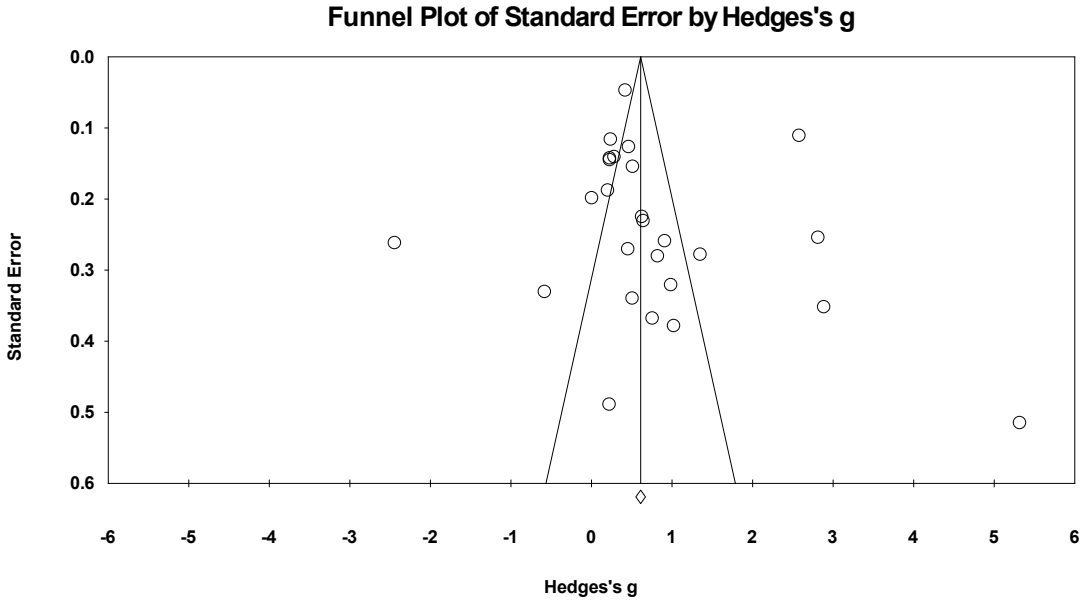


Figure 3. Funnel plot of standard error by Hedges'g

TABLE 4. Testing for heterogeneity

Effect Model	Point Estimate	Q Value	df	p-value	I^2
Fixed Effect Model	0.610	730.963	25	0.000	96.580
Random Effect Model	0.798				

Moderator Analysis

This study offers a novel contribution by conducting a moderator-rich meta-analysis of 26 empirical studies, uncovering contextual, pedagogical, cognitive, and affective factors that shape the effectiveness of AI-powered chatbots in science education. Unlike previous syntheses that only report aggregate effects, this analysis disaggregates findings across multiple layers of educational practice, revealing not only whether chatbots work but when, how, and for whom they work best. The insights are discussed below using four interrelated thematic clusters.

a. Social Context and Infrastructure

One of the most significant findings of this study is the differential effectiveness of chatbots by economic context. The highest effect sizes were found in lower-middle-income countries ($g = 1.919$, $p < 0.001$), followed by high-income and then upper-middle-income nations. In these low-resource environments, AI chatbots often acted as compensatory learning agents which helps bridging gaps in teacher availability, instructional materials, and individualized feedback [14]. These findings echo earlier work on the digital divide and educational technology in resource-constrained settings, which suggests that the marginal utility of AI is greater where baseline support is weaker [13]. From a theoretical perspective, this aligns with critical pedagogy and equity-driven frameworks, in which digital tools are not neutral add-ons but redistributive instruments that can balance access inequalities [22, 23]. In contrast, students in high-income contexts, where digital infrastructures and qualified educators are more widely available, tended to experience more incremental gains, which suggests that chatbots serve a more supplementary or optimization role in those environments [6].

This has strong implications for global education policy where AI integration strategies must be sensitive to local infrastructure, pedagogical needs, and learner profiles. One-size-fits-all models risk reinforcing inequalities if they fail to adapt to contextual realities of the learners and systems they aim to serve.

b. Pedagogical Strategies and Intervention Duration

Pedagogical design significantly influenced chatbot effectiveness. Scaffolded learning and personalized learning yielded the strongest results ($g > 1.0$), supporting the need for structured, adaptive guidance in AI-supported instruction [26, 45]. These findings do more than confirm existing learning theories but they extend Vygotsky's Zone of Proximal Development (ZPD) into AI-mediated environments. In this context, the "more knowledgeable other" becomes a chatbot capable of offering real-time, context-aware support that helps learners advance from what they can do independently to what they can achieve with guidance [26]. This suggests a reconfiguration of social constructivist models, wherein conversational AI can function as a mediating agent, not replacing teachers, but simulating aspects of pedagogical interaction in resource-limited or large-scale settings. The strong performance of personalized learning also affirms learner-centered paradigms, which argue that content tailored to individual needs improves motivation, retention, and transfer [8, 27].

Moderate effects were found for inquiry-based and gamified learning strategies [28, 48]. While these promote engagement, their open-ended nature may offer less structure than scaffolded approaches, making them more suitable for advanced learners. Flipped learning, autonomous learning, and microlearning were less effective, possibly due to insufficient scaffolding or fragmented content delivery [19].

Equally important is the duration of intervention. Studies using long-term (13–16 weeks) or medium-term (9–12 weeks) chatbot implementation showed stronger results than short-term or single-session interventions, which often fail to support feedback cycles, knowledge consolidation, and metacognitive reflection [44, 46]. These findings resonate with cognitive apprenticeship models, which emphasize repeated, authentic practice over time. In terms of frequency, flexible and student-driven use showed the highest impact [4, 48]. This aligns with self-regulated learning theory, where learners benefit from control over pacing and content sequencing. However, novice learners may still require initial guidance to maximize these flexible systems [45].

c. Cognition vs. Affection in Chatbot Use

Another critical insight from this meta-analysis is that chatbots influence both cognitive and noncognitive learning outcomes. On the cognitive side, the strongest effect sizes were found in studies using higher-order thinking skills (HOTS) assessments ($g = 0.974$) or mixed assessments ($g = 1.778$) that combine LOTS and HOTS measures [47]. These findings suggest that chatbots can support complex learning goals such as critical thinking, problem-solving, and conceptual transfer, particularly when integrated into instruction designed to elicit those outcomes.

This affirms and extends Bloom's revised taxonomy by demonstrating that AI tools are capable of targeting not just knowledge recall but analysis, synthesis, and evaluation, especially in well-designed learning environments [13, 47]. Chatbots are evolving beyond basic Q&A functions into cognitive tutors that can support metacognitive reflection [29].

On the affective side, chatbots were found to enhance student satisfaction, engagement, and perceived equity, particularly among students in marginalized or underserved settings [4, 12, 14]. These outcomes are critical for student retention and long-term learning success, especially in STEM education. The personalization and immediacy offered by chatbots may help humanize digital instruction, particularly in large or impersonal classrooms.

However, mixed results were found for self-efficacy and navigation experience. This highlights a critical challenge where learners may enjoy using chatbots but not yet feel confident in their ability to learn independently with them. This gap points to the importance of affective feedback systems and chatbot design features that build learner agency and competence, not just compliance [48].

d. Risks, Limitations, and Ethical Implications

While the findings support the educational value of chatbots, this study also reveals critical boundaries and risks. The lowest effect sizes were found in medical and health sciences, where hands-on clinical judgment and contextual reasoning are central [4, 48]. This reflects a fundamental theoretical misalignment between chatbot-based learning and situated learning theory, which emphasizes that knowledge is socially and physically contextualized.

Similarly, while chatbots performed well in technology-rich fields like computer science, they were less effective in mathematics, possibly due to the abstract-symbolic nature of the subject or lack of visual problem-solving support. These discipline-specific mismatches highlight the need for tailored chatbot designs that reflect epistemological differences across domains. The findings also raise concerns about technological bias. Users with prior experience benefited more than novice users, suggesting that chatbot effectiveness partly depends on digital fluency [45]. This poses an equity challenge, especially in contexts where students have unequal access to or familiarity with technology. Without intentional onboarding and universal design, chatbots may widen rather than close learning gaps. Finally, most studies did not address ethical concerns such as data privacy, algorithmic bias, or over-reliance on automated instruction. As chatbot integration deepens, researchers and educators must grapple with questions about teacher displacement, data surveillance, and decision-making transparency. These are not merely technical issues but pedagogical and moral ones, demanding interdisciplinary collaboration and policy attention.

e. Overall

This moderator analysis reveals that chatbot effectiveness in science education is shaped by multiple, interdependent factors, from socio-economic context and pedagogical design to learner cognition and emotional experience. The findings affirm existing learning theories while also extending them into AI-human interaction spaces, suggesting that chatbots are no longer merely support tools but can function as active mediators of learning. They are particularly impactful when used in low-resource settings to promote equity, designed for scaffolded and personalized learning, and implemented over sustained durations that allow for learner autonomy and deeper cognitive engagement.

However, the analysis also serves as a caution against technological determinism. Chatbots are not universally effective; their success is highly contingent on thoughtful integration aligned with specific subject matter, the readiness and experience of learners, and ethical considerations in design and use. Discipline-specific needs, variations in student digital fluency, and the potential for data privacy and bias issues all require deliberate attention. As developers, educators, and researchers continue to shape the future of AI in education, a critical question remains, can personalization through AI ever truly replace pedagogical presence?

TABLE 5. Summary of moderator analysis

Moderator					95% CI			
Random Effects Model	*k*	ES	SE	σ^2	Lower	Upper	*Z*	*p*
A. Study Location								
High-income	12	0.717	0.257	0.066	0.214	1.22	2.793	0.005
Lower-middle	4	1.919	0.452	0.204	1.032	2.805	4.243	0.000
Upper-middle	10	0.466	0.275	0.076	-0.073	1.005	1.695	0.09
Overall	26	0.794	0.173	0.03	0.454	1.134	4.581	0.000
B. Subject Area								
Computer	12	1.097	0.257	0.066	0.593	1.601	4.267	0.000
Mathematics	3	0.649	0.518	0.268	-0.366	1.663	1.253	0.210
Medical and	4	-0.391	0.434	0.189	-1.242	0.46	-0.901	0.368
Natural	7	1.054	0.339	0.115	0.389	1.718	3.109	0.002
Overall	26	0.794	0.174	0.030	0.453	1.136	4.555	0.000
C. Pedagogical Approach								

Moderator					95% CI			
Random Effects Model	***k***	**ES**	**SE**	**σ²**	**Lower**	**Upper**	***Z***	***p***
Autonomous	3	0.523	0.643	0.414	-0.738	1.783	0.813	0.416
Collaborative	1	1.031	1.164	1.355	-1.250	3.313	0.886	0.376
Flipped	2	0.552	0.789	0.622	-0.994	2.099	0.700	0.484
Gamified	3	0.472	0.643	0.414	-0.789	1.733	0.733	0.464
Inquiry-based	4	0.656	0.571	0.326	-0.464	1.775	1.148	0.251
Micro-learning	1	0.014	1.119	1.251	-2.179	2.206	0.012	0.99
Personalized	4	1.308	0.567	0.321	0.137	2.418	2.307	0.021
Scaffolded	8	1.021	0.405	0.164	0.227	1.816	2.519	0.012
Overall	26	0.789	0.248	0.062	0.302	1.276	3.176	0.001
D. Instructional Integration								
Primary	8	0.879	0.356	0.127	0.182	1.576	2.470	0.013
Supplementary	18	0.768	0.241	0.058	0.295	1.240	3.184	0.001
Overall	26	0.803	0.200	0.040	0.411	1.194	4.022	0.000
E. Intervention Duration								
Single Session	7	0.018	0.387	0.150	-0.741	0.777	0.046	0.963
Very Short Term	7	0.594	0.393	0.154	-0.176	1.364	1.511	0.131
Short-term	1	0.831	1.036	1.073	-1.199	2.862	0.802	0.422
Medium-Term	2	1.622	0.729	0.531	0.193	3.05	2.225	0.026
Long-term	3	1.801	0.601	0.361	0.622	2.979	2.995	0.003
Not specified	6	1.206	0.421	0.177	0.381	2.031	2.864	0.004
Overall	26	0.803	0.203	0.041	0.406	1.200	3.966	0.000
F. User Frequency								
Single Session	4	-0.318	0.540	0.291	-1.376	0.740	-0.590	0.555
Daily	4	0.775	0.549	0.301	-0.3	1.851	1.413	0.158
Weekly	2	0.496	0.757	0.573	-0.987	1.979	0.656	0.512
Varying	12	1.196	0.317	0.100	0.575	1.817	3.774	0.000
Not specified	4	0.999	0.555	0.308	-0.089	2.088	1.799	0.072
Overall	26	0.806	0.215	0.046	0.386	1.227	3.757	0.000
G. User Experience								
Experienced	3	2.120	0.656	0.430	0.835	3.406	3.233	0.001
Novice	20	0.683	0.249	0.062	0.195	1.171	2.745	0.006
Varying	2	0.333	0.767	0.588	-1.169	1.836	0.435	0.664
Not specified	1	0.521	1.09	1.188	-1.615	2.657	0.478	0.633
Overall	26	0.807	0.218	0.048	0.380	1.235	3.701	0.000
H. Assessment Type								
HOTS	10	0.974	0.309	0.096	0.368	1.580	3.150	0.002
LOTS	11	0.225	0.298	0.089	-0.356	0.809	0.757	0.449
MIX	5	1.778	0.457	0.209	0.883	2.674	3.895	0.000
Overall	26	0.942	0.440	0.194	0.079	1.805	2.140	0.032
I. Noncognitive outcomes								
EM	6	0.761	0.407	0.166	-0.036	1.559	1.871	0.061
SP	5	1.803	0.443	0.196	0.934	2.672	4.068	0.000
SL	2	0.609	0.696	0.484	-0.755	1.972	0.875	0.382
ED	1	2.588	0.950	0.902	0.726	4.449	2.724	0.006
SP-EM	2	0.578	0.697	0.486	-0.788	1.945	0.829	0.407
SP-SL	1	0.521	0.956	0.914	-1.352	2.395	0.545	0.586
UN	2	0.361	0.671	0.45	-0.955	1.676	0.537	0.591
Did not measure	7	0.160	0.369	0.136	-0.563	0.883	0.434	0.664
Overall	26	0.801	0.192	0.037	0.424	1.178	4.162	0.000

Legend: HOTS=Higher order thinking skills; LOTS=Lower order thinking skills; MIX= HOTS and LOTS; SP= Satisfaction and Perception; SL= Self-Efficacy and Learning Performance; ED = Equity and Demographics; EM = Engagement and Motivation; UN = Usefulness and Navigation

CONCLUSION

This meta-analysis concludes that AI-powered chatbots have a statistically significant and moderately positive impact on student achievement in science education, confirming the first research objective regarding overall effectiveness. The study also reveals that this effectiveness is not uniform, but is shaped by a combination of moderating factors including pedagogical design, learner experience,

subject area, duration of use, and socio-economic context. Thus, addressing the second research question, chatbots were most effective when used in scaffolded and personalized learning environments, particularly in lower-middle-income countries where they helped bridge gaps in teacher availability and educational resources. The research highlights the importance of using chatbots not as one-size-fits-all tools, but as adaptive learning partners that can scaffold understanding, foster autonomy, and promote educational equity, particularly in under-resourced settings. These findings also suggest that chatbots are not merely supplemental tools, but can function as cognitive and affective mediators of learning when thoughtfully integrated into pedagogy. The study highlights the need for future research to investigate the underlying mechanisms through which chatbots influence learning outcomes, such as their role in fostering metacognition, motivation, and critical thinking. It calls for deeper exploration into chatbot integration at the primary and secondary levels, and within underrepresented subjects like health sciences and mathematics. Longitudinal research is especially needed to assess long-term impacts on learner identity, self-regulation, and scientific reasoning. Furthermore, the study emphasizes that AI adoption in education must be guided by ethical considerations, ensuring inclusivity, data privacy, and equity. Without these safeguards, there is a risk of exacerbating existing educational inequalities. Ultimately, the research reinforces that while AI can enhance efficiency and personalization, the core purpose of education which involves cultivating human understanding, critical reflection, and social responsibility, must remain central in this rapidly evolving digital age.

ACKNOWLEDGMENT

The author would like to thank the Department of Science and Technology – Science Education Institute (DOST-SEI) through the Capacity Building Program in Science and Mathematics Education (CBPSME) for funding the publication of this paper. Their generous support made this research endeavor possible.

REFERENCES

[1] R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, *Intelligence Unleashed: An Argument for AI in Education*. London, UK: Pearson Education, 2016. ISBN: 9780992424886.
[2] L. Chen, F. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.
[3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ: Pearson, 2021. eBook ISBN: 9781292401171.
[4] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promise and Implications for Teaching and Learning*. Boston, MA: Center for Curriculum Redesign, 2019. ISBN: 978-1794293700.
[5] K. VanLehn, "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems," *Educational Psychologist*, vol. 46, no. 4, pp. 197–221, 2011, doi: 10.1080/00461520.2011.611369.
[6] H. F. Li, "Effects of a ChatGPT-based flipped learning guiding approach on learners' courseware project performances and perceptions," *Australasian Journal of Educational Technology*, vol. 39, no. 5, pp. 40–58, 2023, doi: 10.14742/ajet.8923.
[7] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press, 1978. ISBN: 9780674576292.
[8] S. Alneyadi and Y. Wardat, "ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools," *Contemporary Educational Technology*, vol. 15, no. 4, p. ep448, 2023, doi: 10.30935/cedtech/13417.
[9] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100033, 2021, doi: 10.1016/j.caeai.2021.100033.
[10] S. Beltozar-Clemente and E. Díaz-Vega, "Physics XP: Integration of ChatGPT and Gamification to Improve Academic Performance and Motivation in Physics 1 Course," *International Journal of Engineering Pedagogy (iJEP)*, vol. 14, no. 6, pp. 82–92, 2024, doi: 10.3991/ijep.v14i6.47127.
[11] A. D. Topal, C. D. Eren, and A. K. Geçer, "Chatbot Application in a 5th Grade Science Course," *Education and Information Technologies*, vol. 26, pp. 6241–6265, 2021, doi: 10.1007/s10639-021-10627-8.

[12] Y. Xu, J. Zhu, M. Wang, F. Qian, Y. Yang, and J. Zhang, "The Impact of a Digital Game-Based AI Chatbot on Students' Academic Performance, Higher-Order Thinking, and Behavioral Patterns in an Information Technology Curriculum," *Applied Sciences*, vol. 14, no. 6418, 2024, doi: 10.3390/app14156418.
[13] H. Kumar, R. Xiao, B. Lawson, I. Musabirov, J. Shi, X. Wang, H. Luo, J. J. Williams, A. N. Rafferty, J. Stamper, and M. Liut, "Supporting self-reflection at scale with large language models: Insights from randomized field experiments in classrooms," in *Proc. 11th ACM Conf. Learning @ Scale (L@S '24)*, Atlanta, GA, USA, 2024, pp. 86–97. doi: 10.1145/3657604.3662042.
[14] T. Ekukinam, I. N. Udosen, and N. I. Udoh, "The Difference in Academic Performance of Senior Secondary School Biology Students Exposed to Chatbot AI or Expository Method Based on Their Gender," *Universal Academic Journal of Education, Science and Technology*, vol. 6, no. 2, pp. 134–138, Aug. 2024, [Online]. Available: https://www.globalacademicstar.com/article/the-difference-in-academic-performance-of-senior-secondary-school-biology-students-exposed-to-chatbot-ai-or-expository-method-based-on-their-gender-88841.
[15] M. Fırat and S. Kuleli, "GPT vs. Google: A comparative study of self-code learning in ODL students," *Journal of Educational Technology & Online Learning*, vol. 7, no. 3, pp. 308–319, 2024, doi: 10.31681/jetol.1508675.
[16] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and PRISMA Group, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Med.*, vol. 6, no. 7, p. e1000097, Jul. 2009, doi: 10.1371/journal.pmed.1000097.
[17] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, 2009, doi: 10.1002/9780470743386.
[18] J. P. T. Higgins *et al.*, "The Cochrane Collaboration's tool for assessing risk of bias in randomised trials," *BMJ*, vol. 343, p. d5928, Oct. 2011, doi: 10.1136/bmj.d5928.
[19] J. Yin, T. T. Goh, B. Yang, and Y. Xiaobin, "Conversation Technology With Micro-Learning: The Impact of Chatbot-Based Learning on Students' Learning Motivation and Performance," *Journal of Educational Computing Research*, vol. 0, no. 0, pp. 1–23, 2020, doi: 10.1177/0735633120952067.
[20] H. B. Essel, D. Vlachopoulos, A. Tachie-Menson, E. E. Johnson, and P. K. Baah, "The Impact of a Virtual Teaching Assistant (Chatbot) on Students' Learning in Ghanaian Higher Education," *International Journal of Educational Technology in Higher Education*, vol. 19, p. 57, 2022, doi: 10.1186/s41239-022-00362-6.
[21] United Nations, *World Economic Situation and Prospects 2025*. New York: United Nations, 2025. [Online]. Available: https://desapublications.un.org/file/20954/download.
[22] R. Van der Spoel, A. Noroozi, L. E. van Ginkel, and M. E. Mulder, "Teachers' online teaching expectations and experiences during the COVID-19 pandemic in the Netherlands," *European Journal of Teacher Education*, vol. 43, no. 4, pp. 623–638, 2020, doi: 10.1080/02619768.2020.1821185.
[23] H. Y. Sung, G. D. Hwang, and C. Xie, "Artificial intelligence in education: Learning assessment, teacher professional development, and future challenges," *Interactive Learning Environments*, vol. 31, no. 5, pp. 742–761, 2023, doi: 10.1080/10494820.2021.1952615.
[24] J. R. Aguilar-Mejía, S. Tejeda, C. V. Ramirez-Lopez, and C. L. Garay-Rondero, "Design and Use of a Chatbot for Learning Selected Topics of Physics," in *Technology-Enabled Innovations in Education*. Singapore: Springer Nature, 2022, ch. 13, pp. 175–188, doi: 10.1007/978-981-19-3383-7_13.
[25] H. B. Essel, D. Vlachopoulos, H. Nunoo-Mensah, and J. O. Amankwa, "Exploring the Impact of VoiceBots on Multimedia Programming Education Among Ghanaian University Students," *British Journal of Educational Technology*, vol. 00, pp. 1–20, 2024, doi: 10.1111/bjet.13504.
[26] M. Tran, "Generative artificial intelligence as the 'more knowledgeable other': Extending Vygotsky's Zone of Proximal Development," *Frontiers in Education*, vol. 10, p. 12254308, 2025, doi: 10.3389/feduc.2025.12254308.
[27] B. Graefen and N. Fazali, "GPTEACHER: Examining the Efficacy of ChatGPT as a Tool for Public Health Education," *European Journal of Education Studies*, vol. 10, no. 8, pp. 254–259, 2023, doi: 10.46827/ejes.v10i8.4926.
[28] F. E. Çiçek, M. Ülker, M. Özer, and Y. S. Kıyak, "ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: A randomized controlled trial," *Postgraduate Medical Journal*, pp. 1–6, 2024, doi: 10.1093/postmj/qgae170.
[29] H. Liu, "Applicability of ChatGPT in Online Collaborative Learning: Evidence Based on Learning Outcomes," *Proc. Int. Acad. Conf. Educ.*, vol. 1, no. 1, pp. 33–43, 2024, doi: 10.33422/iaceducation.v1i1.656.

[30] S. Al Kahf *et al.*, "Chatbot-based serious games: A useful tool for training medical students? A randomized controlled trial," *PLoS ONE*, vol. 18, no. 3, p. e0278673, Mar. 2023, doi: 10.1371/journal.pone.0278673.
[31] S. Challapalli and J. Leddo, "Comparing the Relative Effectiveness of Chat GPT-generated Content and Human-generated Videos for Teaching Students Calculus," *International Journal of Social Science and Economic Research*, vol. 9, no. 11, pp. 5434–5442, Nov. 2024, doi: 10.46609/IJSSER.2024.v09i11.031.
[32] M. Zawacki-Richter *et al.*, "Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators?" *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, p. 39, 2019, doi: 10.1186/s41239-019-0171-0.
[33] J. Fowlin, "Inventing Distributed Cognition With Artificial Intelligence: Toward a New Theory of Cognition," *Education Sciences*, vol. 15, no. 3, p. 393, 2025, doi: 10.3390/educsci15030393.
[34] A. P. Bhatia, A. Lambat, and T. Jain, "A Comparative Analysis of Conventional and Chat-Generative Pre-trained Transformer-Assisted Teaching Methods in Undergraduate Dental Education," *Cureus*, vol. 16, no. 5, p. e60006, May 2024, doi: 10.7759/cureus.60006.
[35] M. Hakiki, R. Fadli, A. D. Samala, A. Fricticarani, P. Dayurni, K. Rahmadani, A. D. Astiti, and A. Sabir, "Exploring the impact of using Chat-GPT on student learning outcomes in technology learning: The comprehensive experiment," *Adv. Mobile Learn. Educ. Res.*, vol. 3, no. 2, pp. 859–872, 2023, doi: 10.25082/AMLER.2023.02.013.
[36] G. Huesca *et al.*, "Effectiveness of Using ChatGPT as a Tool to Strengthen Benefits of the Flipped Learning Strategy," *Education Sciences*, vol. 14, no. 660, 2024, doi: 10.3390/educsci14060660.
[37] M. Koć-Januchta, K. J. Schöonborn, L. A. E. Tibell, V. K. Chaudhri, and H. C. Heller, "Engaging with Biology by Asking Questions: Investigating Students' Interaction and Learning with an Artificial Intelligence-Enriched Textbook," *Journal of Educational Computing Research*, vol. 58, no. 6, pp. 1190–1224, 2020, doi: 10.1177/0735633120921581.
[38] A. A. K. Kusuma *et al.*, "Effectiveness of Artificial Intelligent Independent Learning (AIIL) with Physics Chatbot of Global Warming Concept," *Momentum: Physics Education Journal*, vol. 8, no. 1, pp. 42–54, 2024, doi: 10.21067/mpej.v8i1.8942.
[39] Y. T. Lin and J.-H. Ye, "Development of an Educational Chatbot System for Enhancing Students' Biology Learning Performance," *Journal of Internet Technology*, vol. 24, no. 2, pp. 275–278, Mar. 2023, doi: 10.53106/160792642023032402006.
[40] R. Mellado-Silva, A. Faúndez-Ugalde, and M. Blanco-Lobos, "Effective Learning of Tax Regulations using Different Chatbot Techniques," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 6, pp. 439–446, Nov. 2020, doi: 10.25046/aj050652.
[41] Z. A. Pardos and S. Bhandari, "ChatGPT-Generated Help Produces Learning Gains Equivalent to Human Tutor-Authored Help on Mathematics Skills," *PLoS ONE*, vol. 19, no. 5, p. e0304013, May 2024, doi: 10.1371/journal.pone.0304013.
[42] T. T. Wu, H. Y. Lee, P. H. Chen, C. J. Lin, and Y. M. Huang, "Integrating peer assessment cycle into ChatGPT for STEM education: A randomised controlled trial on knowledge, skills, and attitudes enhancement," *J. Comput. Assist. Learn.*, vol. 40, no. 1, Oct. 2024. [Online]. Available: https://doi.org/10.1111/jcal.13085.
[43] Y. Xue, H. Chen, G. R. Bai, R. Tairas, and Y. Huang, "Does ChatGPT Help With Introductory Programming? An Experiment of Students Using ChatGPT in CS1," in *Proc. 46th Int. Conf. on Software Engineering: Software Engineering Education and Training (ICSE-SEET '24)*, Apr. 2024, Lisbon, Portugal, doi: 10.1145/3639474.3640076.
[44] J. A. Kulik and J. D. Fletcher, "Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review," *Review of Educational Research*, vol. 86, no. 1, pp. 42–78, 2016, doi: 10.3102/0034654315581420.
[45] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: A systematic literature review," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 56, 2023, doi: 10.1186/s41239-023-00426-1.
[46] C. Y. Chang, G. J. Hwang, and M. L. Gau, "Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training," *British Journal of Educational Technology*, vol. 53, no. 1, pp. 171–188, 2022, doi: 10.1111/bjet.13158.
[47] D. H. Schunk and J. A. Greene, *Handbook of Self-Regulation of Learning and Performance*, 2nd ed. New York: Routledge, 2021, doi: 10.4324/9781315697048.
[48] B. Kim and T. C. Reeves, "Reframing research on learning with technology: In search of the meaning of cognitive tools," *Instructional Science*, vol. 35, no. 3, pp. 207–256, 2007, doi: 10.1007/S11251-006-9005-2.