Meta-Analysis on the Effect Size of Chatbot Integration in Student Science Performance

Joselito Christian Paulus M. Villanueva*, Gian Del N. Atalia, and John Lorence A. Villamin

School of Arts and Sciences, National University, Manila, Philippines
Email: jcpmvillanueva@nu-moa.edu.ph (J.C.P.M.V.); gdnatalia@nu-moa.edu.ph (G.D.N.A.); jlavillamin@nu-moa.edu.ph (J.L.A.V.)
*Corresponding author
Manuscript received June 25, 2025; revised July 17, 2025; accepted November 24, 2025; published June 24, 2026

***Abstract*—This meta-analysis examines the effectiveness of chatbot integration in science education by synthesizing quantitative findings from selected peer-reviewed studies. Employing a rigorous meta-analytic approach, eight studies published between 2007 and 2024 were analyzed using a random-effects model. The inclusion criteria focused on quasi-experimental research measuring student performance in science subjects. Effect sizes were calculated using Hedges' g, with additional analyses conducted to assess heterogeneity, publication bias, and subgroup differences. Findings indicate that chatbot-assisted instruction significantly improves student performance compared to traditional methods. AI-driven chatbots showed notably higher effectiveness than rule-based systems, highlighting the value of adaptive, interactive technologies in educational contexts. Despite high variability among studies, sensitivity analyses confirmed the stability of the results. The findings suggest that chatbot integration—particularly when designed for dynamic, personalized learning—can meaningfully enhance science education outcomes. These insights offer practical implications for educators and policymakers aiming to implement AI tools in curriculum design.**

***Keywords*—chatbots, science education, meta-analysis, artificial intelligence, student performance, educational technology**

I. Introduction

The landscape of education is rapidly changing, with advancements in Artificial Intelligence (AI) and Internet of Things (IoT) established as cornerstones of the modern pedagogical process. These innovations resulted in a paradigm shift in how teachers implement curricula [1]. Technological interventions significantly impact teaching Science, Technology, Engineering, and Mathematics (STEM) by facilitating the unpacking of concepts across different disciplines [2]. The inclusion of technology in science teaching is grounded on the main tenets of the 21st century skills, learner-centered pedagogy, and constructivism [3], making it essential for curriculum implementation.

The implementation of technology in science education not only complies with the relevant theories of learning but is also proven to enhance student engagement and motivation [4]. Furthermore, Technological interventions in the teaching of science also promote the United Nations Sustainable Development Goals (SDGs) by promoting effective learning experiences (SDG 4) and reducing carbon footprint by going digital (SDG 13), along with other SDGs that align with specific curriculum objectives [5].

Recently, the term "chatbot" has gained prominence in contemporary pedagogy, particularly in science education. Chatbots have evolved in function and feature, from the conception of ELIZA, one of the pioneer language processing computer programs [6]. Widespread utilization of chatbots has been driven by the introduction of contemporary AI programs popularized by Apple's Siri, Microsoft's Cortana and Alexa, and OpenAI's ChatGPT [7]. The integration of chatbots in education called for an exploration of their potential to enhance the pedagogical process through learner-centered activities, collaborative engagements, and personalized learning experiences [8]. Chatbots have a multifaceted effect on education, benefiting both teachers and students by giving immediate access to information, facilitating self-directed learning, and providing feedback on student output [9]. Chatbot use in education has been widely regarded in their potential in aiding the teaching-learning process and knowledge retention. Their potential in enhancing learning outcomes and knowledge retention is well-documented. One study on the ChatGPT chatbot's performance in national licensing examinations [10] showed that the program had significant accuracy in answering questions correctly, encouraging further research into the application of chatbots in exam review and preparation.

While chatbot use is linked to task completion in educational environments, it has also been found to hinder critical and higher order thinking due to the superficial engagement in problem solving [11], warranting precaution in their use. AI chatbots are effective in producing acceptable answers given on lower-order thinking skills but cannot be relied on to answer higher-order thinking questions and understand complex concepts [12]. Despite being ineffective in the advanced cognitive levels, chatbots remain valuable to science education, particularly in self-learning, self-assessment [13], personalized learning [14], and academic support in both science [15] and physical education [16].

Studies highlight the relevant positive impact of chatbots in education, particularly in enhancing metacognitive thinking and student engagement. AI chatbots are used by schools to promote social digital interactions and social engagement, therefore contributing to cognitive development [17]. Pereira [18] found that the @dawebot chatbot, which produces practice materials for exams, improves student preparation and overall academic performance. Another notable chatbot, "Jill Watson, acted as an AI teaching assistant, aiding students at the Georgia Institute of Technology preparing for introductory activities, inquiries, preparations, projects, and assessments [19].

Although chatbots have demonstrated potential in science education, a gap remains between their implementation and measurable improvements in student academic performance. In terms of academic utilization, AI chatbots were perceived to be useful by most students, but only a small proportion of

instructors viewed them as beneficial [20]. A cross-sectional study regarding chatbot use among students revealed that they were primarily used for case-based learning, problem solving, and topic clarification [21]. Chatbots have also successfully answered questions from graduate-level examinations of science courses, performing adequately in multiple-choice, short-answer, and essay questions, implying their potential for exam assistance [22]. These findings collectively imply the usefulness of chatbots among students and exam preparation, but it is yet to be established whether they directly impact academic performance.

Despite their benefits, chatbot utilization has faced scholarly criticisms on its negative effects on student performance. A study found that while chatbots improve productivity among students, dependence on them can impair their comprehension and critical thinking [23]. Additionally, educators are urged to be equipped with proper training to integrate chatbots into academic activities and enhance student academic performance [24]. Although Chatbots can answer short responses and lower-order thinking skills, they struggle with reasoning tasks that require step-by-step solutions, limiting their role in improving academic performance [25]. Furthermore, biases in the training data of chatbots also raise concern for academicians, as generative text patterns may disproportionately resemble certain demographics, hindering diversity and creativity in the outputs of students [26].

This research study employs a meta-analysis approach, utilizing quantitative procedures that synthesize various existing research and assess the overall impact of chatbots on science education performance [27] meta-analysis research aims to evaluate the effectiveness of interventions and arrive at a significant generalization across numerous studies [28]. The validity of meta-analysis studies depends on methodological rigor, statistical precision, and sound data presentation [29]. This meta-analysis systematically examines existing research on chatbot integration in science education, assessing its impact on academic achievement. By aggregating findings from diverse literature, this study aims to illuminate the potential of chatbots in science education through a statistical summary of their results [30].

The integration of chatbots in science education represents a transformative shift in the teaching-learning process, necessitating scientific and objective inquiry into their impact on academic achievement. Therefore, this study contributes to the field of research by:

Bridging the research gap between chatbot use and science education achievement. There is limited research linking chatbot use to academic performance in science. Synthesizing the results from the meta-analysis of multiple studies can provide valuable insights into the significance of chatbots in academic achievement.

Advancing educational technology implementation. By evaluating the size effect and statistical significance of chatbot effects on student performance, this study can contribute to implementation policies and evidence-based decision-making.

Informing educators and policymakers. The findings of this research can generate critical insights for educators and administrators on the effectiveness of chatbots on student achievement. Data-driven insights can guide policy decisions regarding chatbot use in the educational context.

II. Literature Review

This Chatbot utilization in education remains a prominent area of interest. This is mainly due to the accessibility of AI to both educators and students [31]. Chatbots are designed to provide accurate, human-like responses to academic questions and instructions, making them valuable tools for learning [32]. This literature review focuses on the evolution of chatbots towards the role of pedagogy and their impact on science education.

A. Evolution of Chatbots towards Education

Chatbots were programmed as an attempt to answer Alan Turing's question on whether a computer program can mimic human conversation to the point of being undetected [33]. The question, later named the Turing Test, is perceived to mark the beginning of chatbot exploration. The first chatbot, ELIZA, simulated dialog with a psychotherapist [6] despite limited knowledge and a dataset. It was succeeded by PARRY, which voiced dialogues mimicking a person with schizophrenia [34]. Despite being sophisticated at its time of conception, PARRY was usually flagged as a non-person or a chatbot due to limitations in its response system [35]. These two early developments in the 20th century demonstrated limited progress in creating chatbots that can pass the Turing Test.

A breakthrough occurred in the 21st century with the creation of Artificial Linguistic Internet Computer Entity (ALICE), which won the Loebner Prize Turing Test from 2000 to 2001 by simulating natural language patterns and techniques [36]. Copying natural patterns of language and connecting them to the internet caused newer chatbots to become perceptive in answering questions and human interventions. The 2000s was followed by the introduction of SmarterChild, a chat program that provides aid in the teaching-learning process, offering answers to questions asked by both students and teachers. The influx of chatbot use permeated the education platform through the need to address personalized learning.

Notable examples of chatbots in education include:

1) Jill Watson, used for e-learning, personalized learning, and administrative tasks among students at Georgia Institute of Technology [19].
2) Artificial Intelligence chatbot Siri, noted for its use of voice-activated commands that answered questions and provided information [37].
3) IBM's Watson, which won a quiz show against human contestants [38], showcases the range of knowledge it can store as well as the rate of response to questions in a timed contest.

These developments in the early 2010's established the potential of chatbots in three areas of education: in communication, assessment, and [39].

B. Chatbots and the 21st Century Pedagogy

Chatbots play a crucial role in 21st century learning by supporting collaboration, lifelong learning, and teacher-training policies [40]. An exploratory study highlighted personalized learning and asynchronous lesson assistance as the main benefits of chatbot use in learning [41].

A systematic review also stated the benefits of a generative AI chatbot in improving student comprehension [42]. Chatbots facilitate Self-Regulated Learning (SRL) through programmed conversations to address student inquiries and learning styles [43]. Through ease of access, AI chatbots can personalize in addressing student needs, making learning more inclusive and student-centered [44]. AI Chatbots can facilitate homework assistance and self-paced learning [45], flexible personalized learning [46], and skills development through chatbots' instructional guidance and step-by-step problem-solving prompts [47]. Not only did student learning processes benefit, but also the teaching processes as well. Chatbots aid teachers in designing complex tasks to foster higher-order thinking skills, as well as modify assessments to align with learning outcomes [48]. However, educators are cautioned to strategically utilize AI chatbots to promote critical thinking and creativity rather than passive learning [49].

C. Impact of Chatbots in Science Education

AI chatbots such as ChatGPT, Bard, Llama, and DeepSeek have revolutionized education by producing e human-like outputs from advanced algorithms [50]. Chatbot impact in science education can be analyzed in key segments such as pedagogy, research, and student performance.

Chatbots streamline logistics and planning among educators by generating learning plans, rubrics, and quizzes. However, the generic nature of outputs requires teacher oversight and revision [51]. Teachers are encouraged to increase both technical knowledge and pedagogical skills to maximize the potential of chatbots [52]. Chatbots facilitate thought organization, feedback generation, and clerical work [53]. This has prompted academic institutions to set policies in regulating chatbot use in education and research. While chatbots benefit productivity and efficiency among science educators, concerns remain regarding overreliance, impairing comprehension and critical thinking among students [11]. They also challenge traditional methods of assessment, being able to create human-like responses to essay examinations and student tasks [54].

Several studies have explored the benefits of chatbot utilization to learner comprehension and task completion. Comparative studies on student performance stated that generative AI-based chatbots effectively enhance student science knowledge, behavior, and intrinsic motivation [55]. Another comparative study found that chatbots outperformed the average student in most sections of biomedical science examinations, underscoring the prospect of AI chatbots aiding in exam preparation [22]. In allied health sciences, mobile chatbots improved student learning achievement and exam preparation, with participants reporting increased self-efficacy and mental preparation [56]. However, a quasi-experimental study revealed a trade-off of chatbot use among students. Students in the experimental group who regularly used chatbots completed their tasks faster than average but demonstrated less depth of comprehension and critical thinking compared to those in the control group [11].

Despite these promising findings regarding chatbot impact on the pedagogical process, there remains a significant gap in the literature regarding the direct impact of chatbots on student performance in the context of science education. Most existing literature focuses on the chatbots' contributions to general pedagogy, academic organization, and instructional support, yet there is limited quantifiable evidence on the impact of chatbots on science-specific tasks and assessments. By analyzing their direct influence on student performance, this meta-analysis research seeks to contribute valuable insights for educators and academic leaders, informing significant data on chatbot use and its relationship to student performance.

III. Materials and Methods

This study employed a meta-analysis research design to examine the effect of chatbot integration on science performance. Meta-analysis was chosen as it allows for the systematic synthesis of quantitative results from a variety of studies, providing a comprehensive overview of the collective evidence regarding the impact of chatbot integration in science education [57, 58]. The search for relevant studies was conducted across electronic databases, academic journals, and other reputable sources, using predefined criteria for study inclusion. These criteria included studies that investigated the effects of chatbot integration on students' performance in science subjects, with quantitative outcome measures available. Studies were screened based on relevance to the research question, methodological rigor, and availability of data [59, 60]. Data extraction procedures were then implemented to collect pertinent information from each included study, such as sample size, effect sizes, specific science subjects, and methodological characteristics. Additionally, efforts were made to address potential sources of bias, such as publication bias and heterogeneity among studies, through appropriate statistical methods and sensitivity analyses [61].

A. Research Study Procedure

Prior to searching peer-reviewed online journal articles, the researchers established criteria for inclusion and exclusion in the meta-analysis. Various meta-search engines were utilized to gather relevant journal articles, including Google Search, Google Scholar, Education Resources Information Center (ERIC), and Journal Storage (JSTOR). Additionally, the software program Publish or Perish [62] was employed to identify lists of journal articles and analyze academic citations. The search was intentionally confined to articles published from 2007 up to the first quarter of 2024 [63]. The descriptors entered the meta-search engines included terms such as chatbot integration, chatbots in education, students' science performance, and variations of these terms. These descriptors were systematically entered into the search engines, with a persistent focus on the core term "chatbot integration" or "chatbots", until relevant studies were exhaustively identified. This comprehensive search strategy aimed to ensure the inclusion of all relevant studies that investigated the effect of chatbot integration on students' science performance within the specified timeframe.

B. Selection Criteria and Coding Procedures

For this study, research articles relevant to the context were investigated, utilizing a quantitative research design from the period spanning 2007 up to the first quarter of 2024. Inclusion criteria were established to select journal articles meeting specific parameters: Articles must be research

articles from peer-reviewed journals and conference proceedings published between 2007 and the first quarter of 2024 and contain explicit references to chatbot integration in their title or abstract. They must utilize student science performance as a dependent variable, focus on the elementary, secondary, or tertiary level of education, and employ a quasi-experimental design. Additionally, articles should focus on a wide range of scientific disciplines, including Biology, Chemistry, Earth Science, and Physics, while also encompassing related fields such as Health Science, Computer Science, Environmental Science, and Astronomy. They should provide sufficient quantitative data to facilitate effect size computations. The collected journal articles were then filtered using the given inclusion criteria. Fig. 1 illustrates the flow of the search process using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) search strategy diagram [64], providing a visual representation of the systematic search and selection process employed in this meta-analysis.

A total of 1056 studies were initially identified, originating from various countries and conducted across different years. Following a rigorous screening process, only 9 articles met the predefined inclusion criteria and were deemed qualified for the analysis. The countries represented in the qualified studies include the Philippines, Korea, Egypt, Taiwan, Pakistan, and Ghana, spanning years from 2019 to 2024. One study [65] was excluded due to an extreme effect size (g = 19.283), which was identified as a statistical outlier in preliminary analysis. Its inclusion resulted in excessive heterogeneity (I^2 = 99.66%), compromising the stability of meta-analytic estimates. Therefore, it was not included in the final analysis.

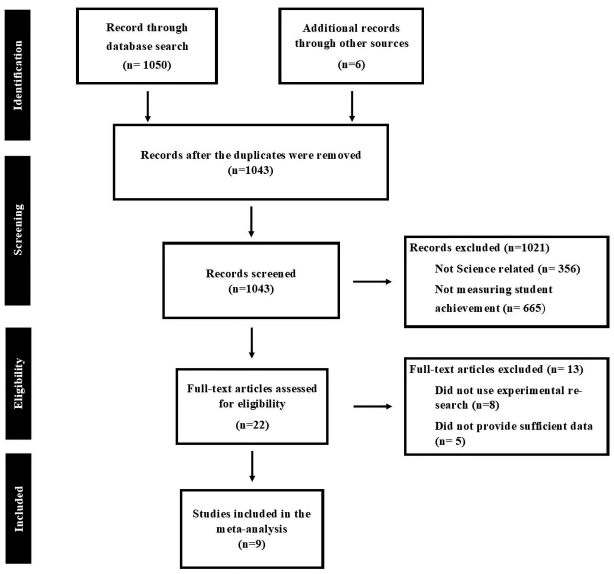

Fig. 1. PRISMA model.

The data collected from these qualified journal articles were systematically coded into several categories to facilitate analysis. These categories encompassed essential aspects of each study, including study identification (author's last name and year of publication), students' grade level, scientific discipline explored, databases utilized, control/comparison condition, instrument used, and outcome measure characteristics (sample size, mean, and standard deviation). This systematic coding allowed for a comprehensive examination of the effect of chatbot integration on students' science performance while ensuring consistency and clarity in data analysis and interpretation.

C. Effect Size Calculation

For Hedges' g was used as the primary effect size measure due to its correction for small sample bias [64, 66]. Effect sizes were computed using STATA18 [67]. A random-effects model was employed to account for heterogeneity among studies. Restricted Maximum Likelihood (REML) estimation was used to estimate the between-study variance [68]. The model was assessed using heterogeneity statistics, including Cochran's Q test, the I^2 statistic to quantify the proportion of variation due to heterogeneity, and T^2 to assess total variability [69].

To detect potential publication bias, several tests were conducted. Egger's Regression Test was used to evaluate the relationship between effect sizes and their standard errors [70]. The Trim-and-Fill Method, a nonparametric approach, was applied to estimate and adjust for missing studies due to publication bias [71]. Additionally, the Galbraith Plot was utilized as a graphical tool to detect outliers and assess heterogeneity [72].

A sensitivity analysis was performed by systematically removing one study at a time to evaluate its influence on the overall effect size. This approach ensured the robustness of the results and helped identify whether a single study disproportionately affected the findings [73].

A meta-regression was conducted using total sample size as a moderator variable to investigate whether study-level sample size influenced effect sizes. Residual heterogeneity was assessed using the Q statistic for residuals [66]. A subgroup analysis was conducted to examine whether chatbot type, specifically AI-driven versus rule-based, influenced the effect size. Studies were grouped accordingly, and separate meta-analyses were performed for each subgroup. The test for subgroup differences was conducted using the Q statistic for between-group heterogeneity [74].

IV. Result and Discussion

The studies reviewed on chatbot use in science education share commonalities in their goals and approaches, while also exhibiting unique features in terms of target audience, platform, and chatbot type (Table 1). A key similarity across the studies is their emphasis on improving student learning outcomes, engagement, and self-regulated learning through chatbot integration. Most studies focused on enhancing conceptual understanding, self-efficacy, and engagement in science subjects by providing interactive and immediate feedback to learners. Additionally, a significant trend observed is the use of AI-driven chatbots employing Natural Language Processing (NLP), knowledge-based systems, or inquiry-based approaches to facilitate student learning (Table 2).

Table 1. Summary of study details

Study	Type of Chatbot	Platform Used	Target Audience	Sample Size	Country of Origin
Chang *et al.* [56]	Mobile chatbot	Mobile-based NLP	University students	36	Taiwan

Study	Type of Chatbot	Platform Used	Target Audience	Sample Size	Country of Origin
Abbasi *et al.* [75]	Android-based chatbot	NLP-based chatbot	University students	110	Pakistan
Lee *et al.* [76]	Rule-based AI chatbot	Network-based instructional system	Grade 6 students	192	South Korea
Lin and Ye [77]	Biology learning chatbot	LINE messaging app	Grade 7 students	34	Taiwan
Prondoza and Panoy [78]	Interactive chatbot	Not specified	Grade 10 students	70	Philippines
Riggs [79]	Inquiry-based chatbot	Not specified (likely NLP-based)	Grade 8 students	60	USA
Chang, Kuo, and Hwang [80]	Knowledge-based chatbot	Mobile learning with NLP	University students	32	Taiwan
Essel *et al.* [81]	Virtual teaching assistant (KNUST-bot)	AI-driven chatbot using zero-coding	University students	68	Ghana

Table 2. Count of AI-driven and rule-based chatbots

Platform Used	Type of Chatbot	Study
AI-Driven (Uses NLP, ML, or Knowledge-Based AI)	6	Abbasi *et al.* [75], Chang *et al.* [56], Chang, Kuo, and Hwang [80], Essel *et al.* [81], Lin and Ye [77], Riggs [79]
Rule-Based	2	Lee *et al.* [76], Prondoza and Panoy [78]

Despite these commonalities, the studies exhibit distinct variations in their implementation. For example, Lee *et al.* [76] developed a rule-based chatbot for sixth-grade students, focusing on a network-based instructional system to improve science conceptual understanding and attitudes toward science, with a particular emphasis on gender differences in engagement and achievement. In contrast, Lin and Ye [77] used a chatbot integrated into the LINE messaging app to support seventh-grade students in biology learning, leveraging mobile technology for personalized interactions beyond the classroom. Similarly, Prondoza and Panoy [78] implemented an interactive chatbot for tenth-grade students, emphasizing self-regulated learning skills alongside science education. These variations highlight differences in chatbot design and technological platforms tailored to specific educational contexts.

Another noteworthy distinction among the studies is the level of chatbot sophistication. AI-driven chatbots, such as those used by Riggs [79] and Chang *et al.* [56, 80] integrated NLP and machine learning to facilitate inquiry-based and knowledge-based learning approaches, respectively. These AI-driven chatbots aimed to create a more dynamic and responsive learning experience. On the other hand, studies like Lee *et al.* [76] and Prondoza and Panoy [78] relied on rule-based systems, which, while effective, had more structured and limited interactions. The impact of these different chatbot types is also reflected in their application, with AI-driven chatbots being more adaptable to students' individual learning paths.

Geographical distribution and sample sizes also vary across the studies, impacting the generalizability of findings. Studies from Taiwan, such as those by Chang *et al.* [56, 80] integrated and Lin and Ye [77] integrated typically had smaller sample sizes (32–36 university or school students), whereas Lee *et al.* [75] in South Korea and Riggs [78] in the USA had larger sample sizes (192 and 60 students, respectively). The study by Essel *et al.* [81] in Ghana focused on higher education and utilized an AI-driven virtual teaching assistant, offering insights into the chatbot's impact on academic performance in a university setting. These differences indicate a range of implementation strategies and learning environments.

A meta-analysis was conducted using a random-effects model to synthesize the effect sizes across eight studies. The pooled effect size, Hedges's g, was 2.41 (95% CI: 1.05–3.76), indicating a statistically significant and large effect of the intervention (Table 3). The test for overall effect was significant (z = 3.48, p < 0.001), suggesting that the intervention had a meaningful impact, with results favoring the treatment group over the control group.

Table 3. Meta-analysis summary

Study	Effect Size	95% CI (Lower, Upper)	Weight (%)
Abbasi *et al.* [75]	1.703	1.269, 2.137	12.92
Chang *et al.* [56]	2.084	1.285, 2.884	12.52
Chang, Kuo, and Hwang [80]	5.935	4.332, 7.539	11.06
Essel *et al.* [81]	4.529	3.635, 5.424	12.39
Lee *et al.* [76]	1.238	0.927, 1.549	13.01
Lin and Ye [77]	0.347	−0.315, 1.008	12.70
Prondoza and Panoy [78]	0.356	−0.111, 0.823	12.90
Riggs [79]	3.644	2.823, 4.465	12.49
Overall (Random-Effects)	2.405	1.052, 3.758	—

Note: Heterogeneity: τ^2 = 3.6378, I^2 = 97.67%, $Q(7)$ = 139.95, p < 0.001, Test for Overall Effect: z = 3.48, p < 0.001

The individual studies reported a wide range of effect sizes, with Chang *et al.* [80] demonstrating the largest effect size (g = 5.94, 95% CI: 4.33–7.54), suggesting a substantial benefit of the intervention in this sample. Similarly, Essel *et al.* [81] also reported a high effect size (g = 4.53, 95% CI: 3.63–5.42), further supporting the intervention's strong positive impact in certain. In contrast, some studies exhibited small or negligible effects, such as Lin and Ye, [77] (g = 0.35, 95% CI: −0.31–1.01) and Prondoza and Panoy, [78] (g = 0.36, 95% CI: −0.11–0.82). These findings suggest that the effectiveness of the intervention may be influenced by study-specific factors, including sample characteristics, implementation fidelity, or contextual variables.

The weight assigned to each study in the meta-analysis ranged from 11.06% [80] to 13.01% [77], indicating that studies with larger sample sizes contributed more to the pooled estimate. Notably, Lee *et al.* [76], which had the highest weight (13.01%), reported a moderate effect size (g = 1.24, 95% CI: 0.93–1.55), reinforcing the robustness of the pooled findings. Interestingly, despite its large effect size, Chang *et al.* [80] had a relatively lower weight (11.06%) due to its small sample size (N = 16 treatment, N = 16 control), highlighting the importance of sample size in determining the influence of individual studies on the overall results (Fig. 2).

A key consideration in this meta-analysis was the substantial heterogeneity observed across studies. The heterogeneity statistics revealed a tau-squared (τ^2) value of 3.64, an I-squared (I^2) value of 97.67%, and an H-squared (H^2) value of 42.99. The high I^2 value suggests that nearly all observed variance was attributable to real differences between studies rather than random sampling error.

Additionally, Cochran's Q test for homogeneity was significant ($Q(7) = 139.95$, $p < 0.001$), confirming substantial heterogeneity among the studies included. Given these findings, a random-effects model was appropriately employed to account for variations between studies [82]. This high level of heterogeneity underscores the need for further investigation into potential moderators, such as differences in study design, intervention protocols, participant characteristics, and outcome measures, which may influence effect sizes.

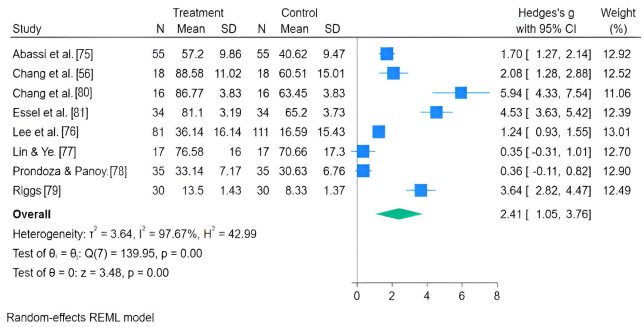

Fig. 2. Forest plot.

To further explore heterogeneity, a Galbraith plot (Fig. 3) was used to assess the distribution of effect sizes and identify potential outliers. The regression line represents the overall trend of effect sizes, while the 95% Confidence Interval (CI) band indicates the range within which most studies are expected to fall. Studies located outside this band suggest significant deviation from the overall effect, potentially contributing to heterogeneity. The observed distribution of data points reveals substantial heterogeneity, as indicated by the spread of studies far from the regression line and outside the 95% CI. This finding aligns with the high $I^2 = 97.67\%$ reported in the forest plot (Fig. 2), confirming considerable variation in effect sizes across studies [83].

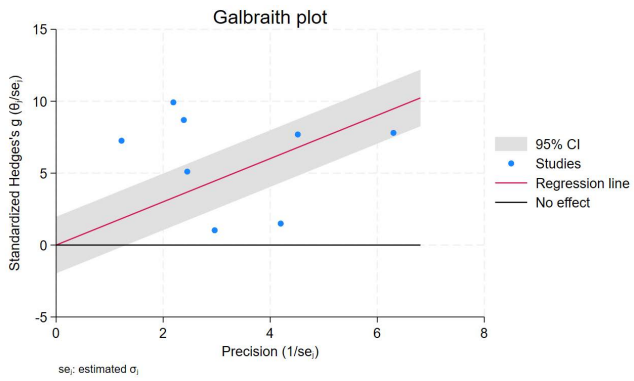

Fig. 3. Galbraith plot.

Several studies are positioned well above the regression line, suggesting stronger intervention effects, while others are closer to the null effect. Notably, a few studies fall outside the 95% CI, indicating potential outliers that may be disproportionately influencing the overall results. For example, studies with extreme effect sizes, such as Chang *et al.* [80] with Hedges's $g = 5.94$, might be driving the observed heterogeneity. Furthermore, the plot illustrates the relationship between study precision and effect size. Studies with higher precision, typically those with larger sample sizes, cluster near the regression line, whereas those with lower precision exhibit greater variability. This pattern is expected, as smaller studies tend to have larger standard errors, leading to wider dispersion in effect estimates [84, 85].

To further assess the stability of the findings, a leave-one-out sensitivity analysis was conducted to evaluate the robustness of the meta-analysis findings by sequentially omitting each study and recalculating the pooled effect size (Hedges's g) with a 95% Confidence Interval (CI). The results, presented in Fig. 4, indicate that the overall effect size remained stable, ranging from 1.96 to 2.71, regardless of which study was omitted. This consistency suggests that no single study had a disproportionate influence on the overall effect size, supporting the robustness of the findings [84].

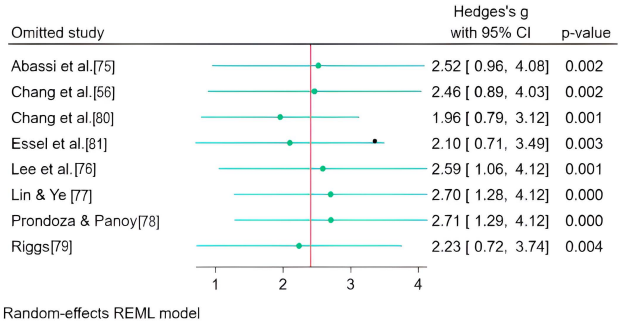

Fig. 4. Leave-one-out sensitivity analysis result.

Statistical significance was maintained across all iterations, with all p-values remaining below 0.05, reinforcing the reliability of the observed effect. The lowest effect size was recorded when Chang *et al.* [80] was omitted ($g = 1.96$, 95% CI [0.79, 3.12], $p = 0.001$), while the highest effect size occurred when Prondoza and Panoy [78] were excluded ($g = 2.71$, 95% CI [1.29, 4.12], $p = 0.000$). These findings indicate slight variability but do not suggest that any single study exerts an undue influence on the overall results.

The results further support the conclusion that the overall heterogeneity observed in the meta-analysis is not driven by a single outlier study. Given this stability, there is no justification for excluding any study from the analysis at this stage. However, the presence of residual heterogeneity suggests that additional subgroup analyses or meta-regression may be useful in identifying potential moderating variables contributing to the observed variance. These additional analyses could help clarify underlying sources of heterogeneity and enhance the interpretability of the findings [86].

To assess potential publication bias, a contour-enhanced funnel plot (Fig. 5) was examined. The plot reveals an asymmetrical distribution of studies around the estimated effect size, with a greater concentration of studies on the right side. This suggests a potential bias favoring larger effect sizes. Additionally, many studies are located within the light gray region ($p < 5\%$), indicating high statistical significance, while few or no studies appear in the non-significant areas, particularly in the lower left quadrant. This pattern suggests the possible absence of unpublished studies with smaller or null effects; a characteristic often associated with publication bias [87].

The presence of asymmetry and the clustering of studies in high-significance regions suggest that studies with non-significant results may be underrepresented in the literature. However, it is also possible that heterogeneity in

study methodologies or sample characteristics contributes to this pattern. To further assess publication bias, additional statistical tests such as Egger's regression test and the Trim-and-Fill method should be conducted to quantify the extent of bias and adjust the effect size accordingly [57, 88]. While the observed asymmetry raises concerns, further analysis is required before drawing definitive conclusions about publication bias.

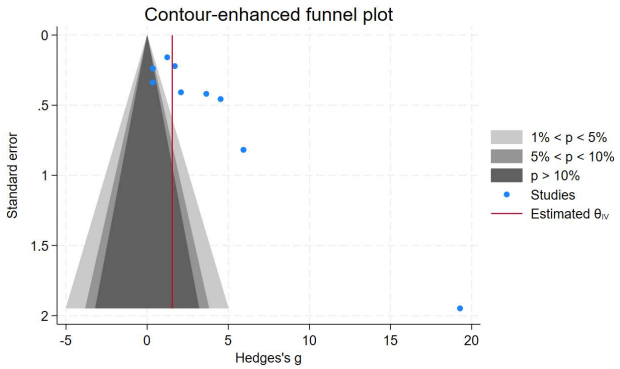

Fig. 5. Contour-enhanced funnel plot.

To formally assess small-study effects, Egger's regression test was performed. The test examines whether there is asymmetry in the funnel plot by evaluating the intercept (β_1) in a regression model. The results showed a significant intercept (β_1 = 8.47, SE = 2.317, z = 3.66, p = 0.0003), providing strong evidence against the null hypothesis (H_0: β_1 = 0) and indicating the presence of small-study effects.

The statistically significant p-value (< 0.05) suggests that smaller studies tend to report larger effect sizes, which align with the asymmetry observed in the funnel plot (Fig. 5). This pattern is commonly associated with publication bias, where studies with larger or significant effects are more likely to be published, while smaller or null results remain unpublished [89]. Given this evidence, further sensitivity analyses are necessary to assess the extent of bias and its potential impact on the overall findings.

To evaluate and adjust for potential publication bias, a Trim-and-Fill analysis was conducted using a nonparametric linear estimator within a random-effects model. The analysis found that no additional studies needed to be removed, as the number of imputed studies was zero. The observed effect size remained Hedges's g = 2.405 (95% CI: [1.052, 3.758]), and this value was unchanged after the Trim-and-Fill adjustment.

The absence of imputed studies suggests that while Egger's test indicated small-study effects, the Trim-and-Fill method did not detect substantial missing studies on the left side of the funnel plot, which would suggest negative bias. This result implies that although some publication bias may be present, its impact on the overall effect size may not be severe. However, caution is still warranted in interpreting the findings, given the significant Egger's test results [90]. Since the observed effect size remained unchanged, no further corrections were necessary. Nonetheless, additional sensitivity analyses, such as meta-regression, could help explore other potential sources of bias.

To examine whether the total sample size influenced the effect size, a meta-regression analysis was conducted. The results, presented in Table 4, indicate that the coefficient for total sample size was −0.011 (SE = 0.014, z = −0.78, p = 0.436, 95% CI [−0.039, 0.017]), indicating a small negative association between sample size and effect size. However, this relationship was not statistically significant. The intercept remained significant (B = 3.252, p = 0.012), suggesting a baseline effect size independent of sample size.

Table 4. Random-effects meta-regression results

Predictor	Coefficient	SE	z	p	95% CI (Lower, Upper)
Total Sample	−0.011	0.014	−0.78	0.436	−0.039, 0.017
Intercept	3.252	1.296	2.51	0.012	0.712, 5.792

Note: Model Fit:
- Residual Heterogeneity: τ^2 = 3.94, I^2 = 97.32%
- Test of Residual Homogeneity: $Q(6)$ = 135.89, p < 0.001
- R^2 = 0.00%, Wald $\chi^2(1)$ = 0.61, p = 0.436

Despite accounting for sample size, residual heterogeneity remained very high (τ^2 = 3.94, I^2 = 97.32%). The test for residual homogeneity ($Q(6)$ = 135.89, p < 0.001) further confirmed that substantial heterogeneity persisted [91]. The proportion of variance in effect sizes explained by total sample size was negligible (R^2 = 0.00%), indicating that sample size did not contribute to explaining the variability in observed effects.

These findings suggest that the total sample size does not significantly influence effect sizes in this meta-analysis. Given persistent heterogeneity, other factors such as study design differences, intervention characteristics, or methodological variations may be contributing to the observed variance. Additionally, the lack of a significant association between sample size and effect size suggests that the publication bias detected by Egger's test is unlikely to be solely attributed to small-study effects, further emphasizing the need to explore alternative explanations for the observed asymmetry.

To further investigate sources of variability, a subgroup analysis examined the differential effects of AI-driven and rule-based chatbots on the outcome measure (Table 5). The results indicated that AI-driven chatbots demonstrated a significantly larger effect size (Hedges' g = 2.969, 95% CI: 1.375, 4.563) compared to rule-based chatbots (Hedges' g = 0.815, 95% CI: −0.048, 1.678). The confidence interval for rule-based chatbots included zero, suggesting a non-significant effect.

The test of group differences (Qb = 5.42, p = 0.020) confirmed a statistically significant moderation effect, indicating that chatbot type influences the observed effect size. AI-driven chatbots consistently produced higher effect sizes across studies, with the largest effects observed in studies by Chang *et al.* [80] and Essel *et al.* [81]. Conversely, the rule-based subgroup showed weaker and more inconsistent effects.

Despite the significant subgroup differences, heterogeneity remained high in both AI-driven (I^2 = 96.30%) and rule-based (I^2 = 89.45%) chatbot studies. The overall heterogeneity (I^2 = 97.67%) suggests that additional factors, such as study design and participant characteristics, may contribute to the observed variability (Table 6). These findings indicate that AI-driven chatbots are more effective than rule-based chatbots, and the consistency of this trend across diverse settings reinforces the validity of the results.

Table 5. Subgroup meta-analysis summary

Group	Study	Hedges's *g*	95% CI (Lower, Upper)	Weight (%)
AI-Driven	Abbasi *et al.* [75]	1.703	1.269, 2.137	12.92
	Chang *et al.* [56]	2.084	1.285, 2.884	12.52
	Chang, Kuo, and Hwang [80]	5.935	4.332, 7.539	11.06
	Essel *et al.* [81]	4.529	3.635, 5.424	12.39
	Lin and Ye [77]	0.347	−0.315, 1.008	12.70
	Riggs [79]	3.644	2.823, 4.465	12.49
	Subtotal (Random-Effects)	2.969	1.375, 4.563	—
Rule-Based	Lee *et al.* [76]	1.238	0.927, 1.549	13.01
	Prondoza and Panoy [78]	0.356	−0.111, 0.823	12.90
	Subtotal (Random-Effects)	0.815	−0.048, 1.678	—
Overall		2.405	1.052, 3.758	—

Table 6. Heterogeneity summary

Group	*Q*(*df*)	*p*	τ^2	I^2 (%)
AI-Driven	93.93(5)	< 0.001	3.748	96.30
Rule-Based	9.48(1)	0.002	0.348	89.45
Overall	139.95(7)	< 0.001	3.638	97.67

Note: Test of Group Differences: $Qb(1) = 5.42$, $p = 0.020$

This meta-analysis aimed to assess the impact of chatbot integration on student performance in science education by synthesizing findings from multiple empirical studies. Given the increasing adoption of artificial intelligence in education, chatbots have emerged as tools that facilitate learning through personalized feedback [92], adaptive questioning [93], and real-time interactions [94]. However, their effectiveness in improving academic performance remains a subject of debate [95], with some studies highlighting significant gains [96] while others indicate negligible to little effects relative to user skill [97]. To systematically evaluate chatbot effectiveness, this study employed a rigorous meta-analytical approach, aggregating quantitative data from peer-reviewed research conducted between 2007 and 2024.

The selection process involved defining strict inclusion criteria, ensuring that only studies with robust methodological designs were considered [91]. The primary focus was on research that explicitly measured student science performance, employed quasi-experimental designs, and provided sufficient quantitative data for effect size computation. Data extraction involved systematically coding studies based on chatbot type (AI-driven or rule-based), educational level (elementary, secondary, or tertiary), scientific discipline (biology, chemistry, physics, etc.), and methodological rigor (sample size, intervention duration, and assessment type). Hedges' g was chosen as the primary effect size measure due to its correction for small-sample bias [98]. A random-effects model was applied to account for variations across studies, ensuring that the meta-analysis captured the diversity in study contexts. Additional statistical analyses, including tests for publication bias and heterogeneity assessment, were conducted to validate the robustness of the findings and minimize the influence of outliers [99].

The results of the meta-analysis indicate that chatbot integration in science education has a statistically significant positive impact on student performance, with an overall effect size of Hedges' $g = 2.41$ ($p < 0.001$). This suggests that students who engaged with chatbot-assisted learning environments outperformed those who relied solely on traditional instruction, like the findings of [100] in the context of language education. The analysis revealed that AI-driven chatbots, such as those employing Natural Language Processing (NLP) and machine learning algorithms, demonstrated superior effectiveness (Hedges' $g = 2.97$) compared to rule-based chatbots (Hedges' $g = 0.82$). These findings highlight the advantage of chatbots capable of generating dynamic and rich responses to quality prompts [101] over those limited to preprogrammed outputs, which may be less flexible in adapting to student needs.

A step-by-step breakdown of how the findings were derived is essential in understanding the study's conclusions. Initially, a comprehensive literature search identified relevant studies, which were then screened based on eligibility criteria. Once selected, data extraction was conducted systematically, ensuring consistency in the coding of key study variables. The said process aligns with the study of Hachfeld and Lazarides [102], which conducted an extensive literature review and data extraction from chatbot research articles. Effect sizes were calculated for each study, followed by an aggregation process to compute the overall effect size using a random-effects model. Heterogeneity was assessed using the I^2 statistic, revealing significant variability among studies ($I^2 = 97.67\%$). This heterogeneity prompted additional subgroup analyses to explore potential moderators, including chatbot type, educational level, and assessment methods. Such factors were explored, like the study of [94], which explored moderating factors of chatbot impact on learning performance. The analysis also included sensitivity checks, such as leave-one-out analysis, to ensure that no single study disproportionately influenced the overall findings. These steps collectively reinforced the credibility of the results and provided a comprehensive understanding of chatbot effectiveness in science education.

In this study, AI-driven chatbots were found to be more effective in enhancing student performance in science education due to their ability to provide real-time feedback [4], facilitate self-regulated learning [79], and adapt to individual student needs [103]. These chatbots leverage machine learning and natural language processing to create interactive learning environments that promote engagement and deeper understanding [104]. The effectiveness of AI-driven chatbots aligns with cognitive load theory, as they help manage the complexity of scientific concepts by breaking them into more accessible components, reducing cognitive overload, and enhancing information retention [105].

Studies reporting higher effect sizes often employed chatbots designed for inquiry-based learning and problem-solving, supporting constructivist learning theories. In these cases, students engaged in dynamic conversations with chatbots, prompting them to ask questions, explore

alternative explanations, and refine their understanding of scientific concepts. This chatbot-student interaction fosters active learning, which has been shown to improve retention and comprehension compared to passive instruction methods [106]. Furthermore, AI-driven chatbots can personalize learning experiences by identifying students' strengths and weaknesses and adjusting instructional content accordingly [102]. This adaptability enhances student motivation and engagement, particularly in self-paced learning environments.

Conversely, rule-based chatbots, which follow a fixed-response and predefined mechanism and offer rigid learning capabilities [107] exhibited lower effectiveness. Their inability to modify flexible responses based on student input limited their capacity to support deep learning and critical thinking skills. While these chatbots can still reinforce factual knowledge, their lack of flexibility prevents them from addressing misconceptions effectively or encouraging students to think critically about complex scientific concepts. This discrepancy underscores the importance of chatbot design in maximizing their educational impact. The success of AI-driven chatbots in science learning suggests that future developments should prioritize enhancing adaptability, responsiveness, and interactivity to optimize learning outcomes [108].

The observed differences in effect sizes across studies can be attributed to several key factors. First, the target audience played a significant role in chatbot effectiveness, with university students benefiting more than younger learners, consistent with the conclusions of [109]. Supporting the insights of [110], this may be due to higher digital literacy, self-regulated learning skills, and the ability to critically engage with AI-generated content among older students. University students are also more likely to utilize chatbots for independent learning, making them particularly well-suited to self-directed educational technologies.

Second, the impact of chatbots varied across scientific disciplines. Studies focusing on conceptual sciences such as biology and chemistry reported greater performance improvements [111, 112], likely because these subjects involve structured knowledge that chatbots can reinforce effectively. In contrast, studies emphasizing computational sciences, where abstract reasoning and stepwise problem-solving are crucial, showed smaller gains [113]. This suggests that chatbot effectiveness is subject to the nature of the scientific content being taught. In highly structured disciplines, chatbots can serve as effective tutors by guiding students through well-defined problems [114]., whereas in more open-ended disciplines, they may be less capable of facilitating deep understanding without additional instructional support [4].

Third, chatbot design and implementation influenced their effectiveness. Studies that integrated chatbots as an integral part of structured instructional frameworks, such as providing formative assessments and personalized explanations, yielded better outcomes than those where chatbots were merely supplementary tools [115]. In well-designed implementations, chatbots guided students through structured learning pathways, adapting responses based on their level of understanding and helping reinforce key concepts through iterative questioning and feedback [116]. This structured integration aligns with best practices in educational technology, which emphasize the importance of aligning digital tools with pedagogical objectives [117].

Lastly, differences in assessment methods contributed to variability in reported effects. Studies using standardized exams as performance measures tended to show higher effect sizes [118], suggesting that chatbot-driven learning is particularly beneficial in preparing students for formal assessments. Standardized tests typically assess factual knowledge and conceptual understanding, areas where chatbots excel in providing reinforcement and immediate feedback. However, studies relying on subjective measures, such as self-reported learning gains, exhibited more modest results, possibly due to variations in student perceptions of chatbot effectiveness [119–121]. These discrepancies indicate that while chatbots can enhance measurable academic performance, their perceived usefulness may depend on students' expectations and prior experiences with AI-assisted learning.

These findings collectively indicate that while chatbots hold significant potential as educational tools, their impact is contingent upon various contextual and implementation factors [122]. The high variability in outcomes suggests that chatbot effectiveness is not universal but rather dependent on how they are designed, integrated into the curriculum, and aligned with student needs [93]. Moving forward, educational institutions and developers should consider these factors when implementing chatbot-assisted learning strategies to maximize their impact on student performance. Additionally, further research should explore the long-term retention effects of chatbot-assisted learning and investigate how these tools can be optimized to support higher-order cognitive skills such as critical thinking, problem-solving, and scientific inquiry.

V. Conclusion

The findings of this meta-analysis indicate that chatbot integration in science education has a statistically significant and positive effect on student performance. The overall effect size (Hedges' $g = 2.41$, $p < 0.001$) suggests that chatbot-assisted learning environments provide notable advantages over traditional instructional methods. AI-driven chatbots demonstrate superior effectiveness in facilitating self-regulated learning, delivering real-time feedback, and promoting deeper engagement with scientific concepts. Despite this, variability in effect sizes across studies highlights the influence of contextual factors such as student level, subject matter, and chatbot design. While chatbot-assisted learning has clear potential, its effectiveness is maximized when implemented within a structured pedagogical framework tailored to student needs.

However, significant heterogeneity in the results underscores the need for cautious interpretation. While chatbots improve factual knowledge retention and learning efficiency, concerns remain about their limitations in fostering critical thinking and higher-order cognitive skills. Rule-based chatbots, which follow rigid response structures, exhibit weaker effects compared to AI-driven counterparts capable of dynamic, context-aware interactions. Given these findings, the role of chatbots in science education should be

carefully integrated, balancing automation with teacher guidance to ensure meaningful learning experiences.

Future meta-analyses on chatbot integration in science education should expand their scope to include a broader range of chatbot types, scientific disciplines, and educational settings to provide a more comprehensive understanding of their effectiveness. Researchers should also examine long-term effects such as knowledge retention and critical thinking development to determine the sustainability of chatbot-assisted learning. Addressing publication bias and ensuring study quality through rigorous inclusion criteria and statistical assessments will enhance the reliability of findings. Investigating moderating factors, such as student demographics, instructional design, and chatbot adaptability, can help identify conditions that maximize chatbot effectiveness. Standardizing outcome measures across studies will improve comparability and strengthen the validity of meta-analytic conclusions. Additionally, promoting open data practices and encouraging replication studies will contribute to the transparency and robustness of chatbot research in science education. By addressing these methodological considerations, future meta-analyses can provide stronger evidence for chatbot implementation, guiding educators and policymakers in optimizing their use for enhanced student learning outcomes.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Mr. Joselito Christian Paulus M. Villanueva contributed to the conceptualization, methodology, study screening and selection, data curation, formal analysis, and visualization of the study. He also participated in the review and editing of the manuscript and provided overall supervision. Mr. Gian Del N. Atalia was responsible for the literature review and took the lead in drafting the original manuscript, as well as contributing to its review and editing. Mr. John Lorence A. Villamin contributed to the original draft preparation, copyediting, formatting, and manuscript review and editing. All authors have read and approved the final version of the manuscript.

References

[1] O. Kandil, R. Rosillo, R. A. E. Aziz, and D. D. L. Fuente, "Investigating the impact of the Internet of Things on higher education: A systematic literature review," *J. Appl. Res. Higher Educ.*, vol. 17, no. 1, pp. 254–273, Jan. 2024. doi: 10.1108/jarhe-05-2023-0223

[2] Journal of Advanced Research in Applied Sciences and Engineering Technology. [Online]. Available: https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index

[3] L. Archambault, H. Leary, and K. Rice, "Pillars of online pedagogy: A framework for teaching in online learning environments," *Educ. Psychol.*, vol. 57, no. 3, pp. 178–191, Jun. 2022. doi: 10.1080/00461520.2022.2051513

[4] D. Lee and S. Yeo, "Developing an AI-based chatbot for practicing responsive teaching in mathematics," *Comput. Educ.*, vol. 191, 104646, Sep. 2022. doi: 10.1016/j.compedu.2022.104646

[5] M. Javaid, A. Haleem, R. P. Singh, and S. Dhall, "Role of virtual reality in advancing education with sustainability and identification of additive manufacturing as its cost-effective enabler," *Sustain. Futures*, vol. 8, 100324, Sep. 2024, doi: 10.1016/j.sftr.2024.100324

[6] J. Weizenbaum, "ELIZA—A computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966. doi: 10.1145/365153.365168

[7] Md. Al-Amin *et al.*, "History of generative Artificial Intelligence (AI) chatbots: Past, present, and future development," arXiv preprint, arXiv:2402.05122, Feb. 2024.

[8] F. Ouyang and P. Jiao, "Artificial intelligence in education: The three paradigms," *Comput. Educ., Artif. Intell.*, vol. 2, 100020, Jan. 2021. doi: 10.1016/j.caeai.2021.100020

[9] L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A review," *IEEE Access*, vol. 8, pp. 75264–75278, Jan. 2020. doi: 10.1109/access.2020.2988510

[10] H. K. Jin, H. E. Lee, and E. Kim, "Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: A systematic review and meta-analysis," *BMC Med. Educ.*, vol. 24, no. 1, 1013, Sep. 2024. doi: 10.1186/s12909-024-05944-8

[11] G. Jošt, V. Taneski, and S. Karakatič, "The impact of large language models on programming education and student learning outcomes," *Applied Sciences*, vol. 14, no. 10, 4115, May 2024. doi: 10.3390/app14104115

[12] R. Murali, D. M. Dhanalakshmy, V. Avudaiappan, and G. Sivakumar, "Towards assessing the credibility of chatbot responses for technical assessments in higher education," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, 2024, pp. 1–9. doi: 10.1109/EDUCON60312.2024.10578934

[13] A. Mahroof, V. Gamage, K. Rajendran, S. Rajkumar, S. Rajapaksha, and D. Wijendra, "An AI-based chatbot to self-learn and self-assess performance in ordinary level chemistry," in *Proc. Int. Conf. Adv. Comput. (ICAC)*, 2020, pp. 216–221. doi: 10.1109/ICAC51239.2020.9357131

[14] G. Molnar and Z. Szuts, "The role of chatbots in formal education," in *Proc. IEEE 16th Int. Symp. Intell. Syst. Informatics (SISY)*, 2018, pp. 197–202. doi: 10.1109/SISY.2018.8524609

[15] E. A. Lambebo and H.-L. Chen, "Chatbots in higher education: A systematic review," *Interactive Learning Environments*, vol. 33, no. 4, pp. 2781–2807, Dec. 2024. doi: 10.1080/10494820.2024.2436931

[16] T. Ha *et al.*, "Editorial," *J. Phys. Educ. Sport*, vol. 24, no. 1, Jan. 2024. doi: 10.7752/jpes.2024.01001

[17] P. Bii, "Chatbot technology: A possible means of unlocking student potential to learn how to learn," *Educ. Res.*, vol. 4, no. 2, pp. 218–221, Jan. 2013.

[18] J. Pereira, "Leveraging chatbots to improve self-guided learning through conversational quizzes," in *Proc. ACM Int. Conf. Learn. Anal. Knowl. (LAK)*, 2016, pp. 911–918. doi: 10.1145/3012430.3012625

[19] Q. Wang, K. Saha, E. Gregori, D. Joyner, and A. Goel, "Towards mutual theory of mind in human–AI interaction: How language reflects what students perceive about a virtual teaching assistant," in *Proc. ACM Conf. Hum. Factors Comput. Syst. (CHI)*, 2021, pp. 1–14. doi: 10.1145/3411764.3445645

[20] P. Tangadulrat, S. Sono, and B. Tangtrakulwanich, "Using ChatGPT for clinical practice and medical education: Cross-sectional survey of medical students' and physicians' perceptions," *JMIR Med. Educ.*, vol. 9, e50658, Dec. 2023. doi: 10.2196/50658

[21] N. K. Qamar, N. S. W. Butt, N. M. Abaid, N. A. Rashid, N. R. Tahira, and N. A. S. Iqbal, "Exploring the adoption and utilisation of ChatGPT in everyday academic practices among medical and dental students of Rawalpindi and Islamabad: A study on the impact and perceived effectiveness," *J. Pakistan Med. Assoc.*, vol. 74, no. 12, pp. 2101–2106, Nov. 2024. doi: 10.47391/jpma.11053

[22] D. Stribling, Y. Xia, M. K. Amer, K. S. Graim, C. J. Mulligan, and R. Renne, "The model student: GPT-4 performance on graduate biomedical science exams," *Sci. Rep.*, vol. 14, no. 1, p. 5670, Mar. 2024. doi: 10.1038/s41598-024-55568-7

[23] N. Baláž, J. Porubän, M. Horváth, and T. Kormaník, "Using ChatGPT during implementation of programs in education," in *Proc. Int. Conf. Perform. Eval. Comput. Telecommun. Syst. (ICPECS)*, Leibniz-Zentrum für Informatik (Schloss Dagstuhl), Jan. 2024, pp. 1–10. doi: 10.4230/OASICS.ICPE.2024.18

[24] M. Montenegro-Rueda, J. Fernández-Cerero, J. M. Fernández-Batanero, and E. López-Meneses, "Impact of the implementation of ChatGPT in education: A systematic review," *Computers*, vol. 12, no. 8, p. 153, Jul. 2023. doi: 10.3390/computers12080153

[25] J. Savelka, A. Agarwal, C. Bogart, Y. Song, and M. Sakr, "Can Generative Pre-Trained Transformers (GPT) pass assessments in higher education programming courses?" in *Proc. ACM Conf. Innov. Technol. Comput. Sci. Educ. (ITiCSE)*, 2023, pp. 117–123. doi: 10.1145/3587102.3588792

[26] D. Baidoo-Anu and L. O. Ansah. (Jan. 2023). Education in the era of generative Artificial Intelligence (AI): Understanding the potential

benefits of ChatGPT in promoting teaching and learning. *SSRN Electronic Journal*. [Online]. Available: https://ssrn.com/abstract=4337484

[27] G. D. Borman and J. A. Grigg, "Visual and narrative interpretation," *The Handbook of Research Synthesis and Meta-Analysis*, New York, NY, USA: Russell Sage Foundation, pp. 497–519, 2009.

[28] J. Gurevitch, J. Koricheva, S. Nakagawa, and G. Stewart, "Meta-analysis and the science of research synthesis," *Nature*, vol. 555, no. 7695, pp. 175–182, Mar. 2018. doi: 10.1038/nature25753

[29] D. Moher *et al.*, "Preferred reporting items for Systematic Reviews and Meta-Analyses: the PRISMA statement," *Ann. Intern. Med.*, vol. 151, no. 4, pp. 264–269, Aug. 2009. doi: 10.7326/0003-4819-151-4-200908180-00135

[30] O. M. Dekkers, J. P. Vandenbroucke, M. Cevallos, A. G. Renehan, D. G. Altman, and M. Egger, "COSMOS-E: Guidance on conducting systematic reviews and meta-analyses of observational studies of etiology," *PLoS Med.*, vol. 16, no. 2, e1002742, Feb. 2019. doi: 10.1371/journal.pmed.1002742

[31] H. Crompton and D. Burke, "Artificial intelligence in higher education: the state of the field," *Int. J. Educ. Technol. Higher Educ.*, vol. 20, no. 1, Apr. 2023. doi: 10.1186/s41239-023-00392-8

[32] T. N. Dinh and M. T. Thai, "AI and Blockchain: A disruptive integration," *Computer*, vol. 51, no. 9, pp. 48–53, Sep. 2018. doi: 10.1109/mc.2018.3620971

[33] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950. doi: 10.1093/mind/lix.236.433

[34] K. M. Colby, "Artificial paranoia," *Artificial Intelligence*, vol. 2, no. 1, pp. 1–25, 1971. doi: 10.1016/0004-3702(71)90002-6

[35] K. M. Colby, F. D. Hilf, S. Weber, and H. C. Kraemer, "Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes," *Artificial Intelligence*, vol. 3, pp. 199–221, Jan. 1972. doi: 10.1016/0004-3702(72)90049-5

[36] R. Wallace. Chatbot A.L.I.C.E. *chatbots.org*. [Online]. Available: https://www.chatbots.org/chatbot/a.l.i.c.e/

[37] J. Aron, "How innovative is Apple's new voice assistant, Siri?" *New Sci.*, vol. 212, no. 2836, 24, Oct. 2011. doi: 10.1016/s0262-4079(11)62647-x

[38] A. Lally and P. Fodor. (Jan. 2011). Natural language processing with Prolog in the IBM Watson system. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.708.9658

[39] M. Naveenkumar and S. Domnic, "Vector quantization-based pairwise joint distance maps (VQ-PJDM) for 3D action recognition," *Procedia Comput. Sci.*, vol. 133, pp. 27–36, Jan. 2018. doi: 10.1016/j.procs.2018.07.005

[40] H. Luan *et al.*, "Challenges and future directions of big data and artificial intelligence in education," *Frontiers in Psychology*, vol. 11, 580820, Oct. 2020. doi: 10.3389/fpsyg.2020.580820

[41] G.-J. Hwang, H. Xie, B. W. Wah, and D. Gašević, "Vision, challenges, roles and research issues of artificial intelligence in education," *Comput. Educ., Artif. Intell.*, vol. 1, 100001, Jan. 2020. doi: 10.1016/j.caeai.2020.100001

[42] M. M. Asad and A. Ajaz, "Impact of ChatGPT and generative AI on lifelong learning and upskilling learners in higher education: Unveiling the challenges and opportunities globally," *Int. J. Inf. Learn. Technol.*, vol. 41, no. 5, pp. 507–523, Oct. 2024. doi: 10.1108/ijilt-06-2024-0103

[43] A. Karoui *et al.*, "Towards an automated adaptive learning web platform through personalization of language learning pathways," *Lect. Notes Comput. Sci.*, vol. 13349, Springer, pp. 448–454, 2022. doi: 10.1007/978-3-031-16290-9_35

[44] B. Elov, M. Samatov, N. Gayibova, N. Samadova, M. Qodirova, and S. Amirovich, "A mantle of chatbots in the place of pedagogy and learning to generate tools to provide assistance using AI," in *Proc. 4th Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE)*, May 14–15, 2024, pp. 1075–1079. doi: 10.1109/ICACITE60783.2024.10617382

[45] J. Crawford, M. Cowling, and K.-A. Allen, "Leadership is needed for ethical ChatGPT: Character, assessment, and learning using Artificial Intelligence (AI)," *J. Univ. Teach. Learn. Pract.*, vol. 20, no. 3, Mar. 2023. doi: 10.53761/1.20.3.02

[46] R. I. Fariani, K. Junus, and H. B. Santoso, "A systematic literature review on personalised learning in the higher education context," *Technol. Knowl. Learn.*, vol. 28, no. 2, pp. 449–476, Nov. 2022. doi: 10.1007/s10758-022-09628-4

[47] M. Benvenuti *et al.*, "Artificial intelligence and human behavioral development: A perspective on new skills and competences acquisition for the educational context," *Comput. Hum. Behav.*, vol. 148, 107903, Aug. 2023. doi: 10.1016/j.chb.2023.107903

[48] T. Karakose and T. Tülübaş, "How can ChatGPT facilitate teaching and learning: Implications for contemporary education," *Educ. Process Int. J.*, vol. 12, no. 4, Jan. 2023. doi: 10.22521/edupij.2023.124.1

[49] A. A. Ka'bi, "Proposed artificial intelligence algorithm and deep learning techniques for development of higher education," *Int. J. Intell. Netw.*, vol. 4, pp. 68–73, Jan. 2023. doi: 10.1016/j.ijin.2023.03.002

[50] (Dec. 8, 2022). The ChatGPT chatbot is blowing people away with its writing skills. *The University of Sydney*. [Online]. Available: https://www.sydney.edu.au/news-opinion/news/2022/12/08/the-chatgpt-chatbot-is-blowing-people-away-with-its-writing-skil.html

[51] G. Cooper, "Examining science education in ChatGPT: An exploratory study of generative artificial intelligence," *J. Sci. Educ. Technol.*, vol. 32, no. 3, pp. 444–452, Mar. 2023. doi: 10.1007/s10956-023-10039-y

[52] I. Celik, "Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate Artificial Intelligence (AI)-based tools into education," *Comput. Hum. Behav.*, vol. 138, 107468, Aug. 2022. doi: 10.1016/j.chb.2022.107468

[53] (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*. *613(7945)*, 612. [Online]. Available: doi: 10.1038/d41586-023-00191-1

[54] C. Stokel-Walke. (Dec. 2022). AI bot ChatGPT writes smart essays—should professors worry? *Nature*. [Online]. Available: https://www.nature.com/articles/d41586-022-04397-7

[55] D. T. K. Ng, C. W. Tan, and J. K. L. Leung, "Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study," *Br. J. Educ. Technol.*, vol. 55, no. 4, pp. 1328–1353, Mar. 2024. doi: 10.1111/bjet.13454

[56] C. Chang, G. Hwang, and M. Gau, "Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training," *Br. J. Educ. Technol.*, vol. 53, no. 1, pp. 171–188, Aug. 2021. doi: 10.1111/bjet.13158

[57] X. Deng and Z. Yu, "A meta-analysis and systematic review of the effect of chatbot technology use in sustainable education," *Sustainability*, vol. 15, no. 4, 2940, Feb. 2023. doi: 10.3390/su15042940

[58] A. C. Gelijns, "Meta-analysis: A quantitative approach to research integration," *Modern Methods of Clinical Investigation*, 1990.

[59] A. Aryankhesal, M. Behzadifar, N. L. Bragazzi, A. Ghashghaee, and M. Behzadifar. (May 1, 2018). A framework for conducting meta-analysis studies: Methodological concerns and recommendations. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6005986/

[60] G. M. Tawfik *et al.*, "A step by step guide for conducting a systematic review and meta-analysis with simulation data," *Trop. Med. Health*, vol. 47, no. 1, p. 46, Aug. 2019. doi: 10.1186/s41182-019-0165-6

[61] G. Frampton *et al.*, "Principles and framework for assessing the risk of bias for studies included in comparative quantitative environmental systematic reviews," *Environ. Evid.*, vol. 11, no. 1, Mar. 2022. doi: 10.1186/s13750-022-00264-0

[62] Publish or perish. (Feb. 6, 2016). Harzing.com. [Online]. Available: https://harzing.com/resources/publish-or-perish

[63] W. Mengist, T. Soromessa, and G. Legese, "Ecosystem services research in mountainous regions: A systematic literature review on current knowledge and research gaps," *Sci. Total Environ.*, vol. 702, 134581, Nov. 2019. doi: 10.1016/j.scitotenv.2019.134581

[64] L. Page, K. Meyer, J. Lee, and H. Gehlbach, "Conditions under which college students can be responsive to text-based nudging," *SSRN Electronic Journal*, Jan. 2024. doi: 10.2139/ssrn.5069370

[65] S. S. Barsoum, M. S. Elnagar, and B. M. Awad, "The effectiveness of using a cognitive style-based chatbot in developing science concepts and critical thinking skills among preparatory school pupils," *Eur. Sci. J.*, vol. 18, no. 22, 52, Jul. 2022. doi: 10.19044/esj.2022.v18n22p52

[66] S. Nakagawa, Y. Yang, E. L. Macartney, R. Spake, and M. Lagisz, "Quantitative evidence synthesis: A practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences," *Environ. Evid.*, vol. 12, no. 1, 8, Apr. 2023. doi: 10.1186/s13750-023-00301-6

[67] Z. Irsova, H. Doucouliagos, T. Havranek, and T. D. Stanley, "Meta-analysis of social science research: A practitioner's guide," *J. Econ. Surv.*, vol. 38, no. 5, pp. 1547–1566, Nov. 2023. doi: 10.1111/joes.12595

[68] StataCorp LLC, *Stata Statistical Software: Release 18*, StataCorp LLC, College Station, TX, USA, 2023.

[69] S. E. Seide, C. Röver, and T. Friede, "Likelihood-based random-effects meta-analysis with few studies: Empirical and simulation studies," *BMC Med. Res. Methodol.*, vol. 19, no. 1, p. 16, Jan. 2019. doi: 10.1186/s12874-018-0618-3

[70] T. B. Huedo-Medina, J. Sánchez-Meca, F. Marín-Martínez, and J. Botella, "Assessing heterogeneity in meta-analysis: Q statistic or I2 index?," *Psychol. Methods*, vol. 11, no. 2, pp. 193–206, Jan. 2006. doi: 10.1037/1082-989x.11.2.193

[71] L. Shi and L. Lin, "The trim-and-fill method for publication bias: Practical guidelines and recommendations based on a large database of

meta-analyses," *Medicine*, vol. 98, no. 23, e15987, Jun. 2019. doi: 10.1097/MD.0000000000015987

[72] M. Kossmeier, U. S. Tran, and M. Voracek, "Charting the landscape of graphical displays for meta-analysis and systematic reviews: A comprehensive review, taxonomy, and feature analysis," *BMC Med. Res. Methodol.*, vol. 20, no. 1, p. 26, Feb. 2020. doi: 10.1186/s12874-020-0911-9

[73] F. I. Mowbray, D. Manlongat, and M. Shukla, "Sensitivity analysis: A method to promote certainty and transparency in nursing and health research," *Can. J. Nurs. Res.*, vol. 54, no. 4, pp. 371–376, Jun. 2022. doi: 10.1177/08445621221107108

[74] L. M. Spineli and N. Pandis, "Problems and pitfalls in subgroup analysis and meta-regression," *Am. J. Orthod. Dentofacial Orthop.*, vol. 158, no. 6, pp. 901–904, Nov. 2020. doi: 10.1016/j.ajodo.2020.09.001

[75] S. Abbasi, H. Kazi, and N. Nawaz Hussaini, "Effect of chatbot systems on students' learning outcomes," *Sylwan*, vol. 163, no. 10, RNNbS, 2019.

[76] J. Lee, T. An, H.-E. Chu, H.-G. Hong, and S. N. Martin, "Improving science conceptual understanding and attitudes in elementary science classes through the development and application of a rule-based AI chatbot," *Asia-Pac. Sci. Educ.*, vol. 9, no. 2, pp. 365–412, Dec. 2023. doi: 10.1163/23641177-bja10070

[77] Y.-T. Lin and J.-H. Yeh, "Development of an educational chatbot system for enhancing students' biology learning performance," *J. Internet Technol.*, vol. 24, no. 2, pp. 275–281, Mar. 2023. doi: 10.53106/160792642023032402006

[78] G. R. M. Prondoza and J. F. D. Panoy, "Development of chatbot supplementary tool in science and the self-regulated learning skills among the grade 10 students," *Asia Pac. J. Adv. Educ. Technol. (APJAET)*, pp. 107–116, Sep. 2022. doi: 10.54476/apjaet/95445

[79] E. C. B. Riggs. (Jan. 25, 2024). Integrating chatbot in an inquiry-based approach in teaching science. *EPRA Journals*. [Online]. Available: https://eprajournals.com/IJMR/article/12192

[80] C.-Y. Chang, S.-Y. Kuo, and G. H. Hwang, "Chatbot-facilitated nursing education: Incorporating a knowledge-based chatbot system into a nursing training program," *Dir. Open Access J. (DOAJ)*, Jan. 2022.

[81] 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," *Int. J. Educ. Technol. Higher Educ.*, vol. 19, no. 1, Nov. 2022. doi: 10.1186/s41239-022-00362-6

[82] J. E. McKenzie and A. A. Veroniki, "A brief note on the random-effects meta-analysis model and its relationship to other models," *J. Clin. Epidemiol.*, vol. 174, 111492, Aug. 2024. doi: 10.1016/j.jclinepi.2024.111492

[83] Z. Meng, J. Wang, L. Lin, and C. Wu, "Sensitivity analysis with iterative outlier detection for systematic reviews and meta-analyses," *Stat. Med.*, vol. 43, no. 8, pp. 1549–1563, Feb. 2024. doi: 10.1002/sim.10008

[84] R. Zejnullahi and L. V. Hedges, "Robust variance estimation in small meta-analysis with the standardized mean difference," *Res. Synth. Methods*, vol. 15, no. 1, pp. 44–60, Sep. 2023. doi: 10.1002/jrsm.1668

[85] J. Sánchez-Meca and J. Botella, "WITHDRAWN: Moderators analysis in meta-analysis: Meta-regression and subgroups analyzes," *Cir. Esp. (Engl. Ed.)*, vol. 102, no. 7, pp. 389–390, Apr. 2024. doi: 10.1016/j.cireng.2024.03.006

[86] J. Afonso, R. Ramirez-Campillo, F. M. Clemente, F. C. Büttner, and R. Andrade, "The perils of misinterpreting and misusing 'publication bias' in meta-analyses: An education review on funnel plot-based methods," *Sports Med.*, vol. 54, no. 2, pp. 257–269, Sep. 2023. doi: 10.1007/s40279-023-01927-9

[87] B. Fernández-Castilla, L. Declercq, L. Jamshidi, S. N. Beretvas, P. Onghena, and W. V. D. Noortgate, "Detecting selection bias in meta-analyses with multiple outcomes: A simulation study," *J. Exp. Educ.*, vol. 89, no. 1, pp. 125–144, Apr. 2019. doi: 10.1080/00220973.2019.1582470

[88] J. P. Souza, C. Pileggi, and J. G. Cecatti, "Assessment of funnel plot asymmetry and publication bias in reproductive health meta-analyses: An analytic survey," *Reprod. Health*, vol. 4, no. 1, p. 3, Apr. 2007. doi: 10.1186/1742-4755-4-3

[89] M. Borenstein, "How to understand and report heterogeneity in a meta-analysis: The difference between I-squared and prediction intervals," *Integr. Med. Res.*, vol. 12, no. 4, 101014, Nov. 2023. doi: 10.1016/j.imr.2023.101014

[90] L. Shi, H. Chu, and L. Lin, "A Bayesian approach to assessing small-study effects in meta-analysis of a binary outcome with controlled false positive rate," *Res. Synth. Methods*, vol. 11, no. 4, pp. 535–552, Jul. 2020. doi: 10.1002/jrsm.1415

[91] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *Syst. Rev.*, vol. 10, no. 1, p. 89, Mar. 2021. doi: 10.1186/s13643-021-01626-4

[92] M. Choque-Diaz, J. Armas-Aguirre, and P. Shiguihara-Juarez, "Cognitive technology model to enhance academic support services with chatbots," in *Proc. IEEE XXV Int. Conf. Electron., Electr. Eng. Comput. (INTERCON)*, Aug. 2018, pp. 1–4. doi: 10.1109/INTERCON.2018.8526411

[93] J.-B. Aujogue and A. Aussem, "Hierarchical recurrent attention networks for context-aware education chatbots," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2019, pp. 1–8. doi: 10.1109/IJCNN.2019.8852445

[94] M. Laun and F. Wolff, "Chatbots in education: Hype or help? A meta-analysis," *Learn. Individ. Differ.*, vol. 119, 102646, Feb. 2025. doi: 10.1016/j.lindif.2025.102646

[95] S. Alneyadi and Y. Wardat, "ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools," *Contemp. Educ. Technol.*, vol. 15, no. 4, ep448, Jun. 2023. doi: 10.30935/cedtech/13417

[96] A. Shoufan, "Can students without prior knowledge use ChatGPT to answer test questions? An empirical study," *ACM Trans. Comput. Educ.*, vol. 23, no. 4, pp. 1–29. Oct. 2023. doi: 10.1145/3628162

[97] G. Schwarzer, G. Rücker, and J. R. Carpenter, *Meta-Analysis with R*, Cham, Switzerland: Springer, 2015. doi: 10.1007/978-3-319-21416-0

[98] W. Viechtbauer, "Conducting meta-analyses in R with the metafor package," *J. Stat. Softw.*, vol. 36, no. 3, Jan. 2010. doi: 10.18637/jss.v036.i03

[99] J. P. T. Higgins and S. G. Thompson, "Quantifying heterogeneity in a meta-analysis," *Stat. Med.*, vol. 21, no. 11, pp. 1539–1558, May 2002. doi: 10.1002/sim.1186

[100] P. Polyzi and L. Moussiades, "An artificial vocabulary learning assistant," *Educ. Inf. Technol.*, vol. 28, no. 12, pp. 16431–16455, May 2023. doi: 10.1007/s10639-023-11810-9

[101] M. Nazari and G. Saadi, "Developing effective prompts to improve communication with ChatGPT: A formula for higher education stakeholders," *Discov. Educ.*, vol. 3, no. 1, Apr. 2024. doi: 10.1007/s44217-024-00122-w

[102] A. Hachfeld and R. Lazarides, "The relation between teacher self-reported individualization and student-perceived teaching quality in linguistically heterogeneous classes: an exploratory study," *Eur. J. Psychol. Educ.*, vol. 36, no. 4, pp. 1159–1179, Oct. 2020. doi: 10.1007/s10212-020-00501-5

[103] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," *Int. J. Educ. Technol. Higher Educ.*, vol. 20, no. 1, Oct. 2023. doi: 10.1186/s41239-023-00426-1

[104] G. P. Ashok, "Virtual bot powered by machine learning and NLP technologies: Emulating human-like conversations through speech-to-text conversions," *Int. J. Sci. Res. Eng. Manag.*, vol. 7, no. 7, Jul. 2023. doi: 10.55041/ijsrem24835

[105] S. Brunswicker, Y. Zhang, C. Rashidian, and D. W. Linna, "Trust through words: The systemize-empathize-effect of language in task-oriented conversational agents," *Comput. Hum. Behav.*, vol. 165, 108516, Dec. 2024. doi: 10.1016/j.chb.2024.108516

[106] V. Robledo-Rella and B.-Y. Toh, "Artificial intelligence in physics courses to support active learning," in *Proc. ACM Conf. Innov. Technol. Comput. Sci. Educ.*, 2024, pp. 68–75. doi: 10.1145/3678610.3678631

[107] R. Winkler and M. Soellner, "Unleashing the potential of chatbots in education: A state-of-the-art analysis," *Acad. Manag. Proc.*, vol. 2018, no. 1, 15903, Jul. 2018. doi: 10.5465/AMBPP.2018.15903abstract

[108] D. E. Gonda, J. Luo, Y.-L. Wong, and C.-U. Lei, "Evaluation of developing educational chatbots based on the seven principles for good teaching," in *Proc. IEEE Int. Conf. Teach., Assess., Learn. Eng. (TALE)*, 2018, pp. 446–453. doi: 10.1109/TALE.2018.8615175

[109] T. Tikhomirova, A. Malykh, and S. Malykh, "Predicting academic achievement with cognitive abilities: Cross-sectional study across school education," *Behavioral Sciences*, vol. 10, no. 10, 158, Oct. 2020. doi: 10.3390/bs10100158

[110] J. Garzón and J. Acevedo, "Meta-analysis of the impact of Augmented Reality on students' learning gains," *Educ. Res. Rev.*, vol. 27, pp. 244–260, Apr. 2019. doi: 10.1016/j.edurev.2019.04.001

[111] A. Mihalache, M. M. Popovic, and R. H. Muni, "Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment," *JAMA Ophthalmology*, vol. 141, no. 6, p. 589, Apr. 2023. doi: 10.1001/jamaophthalmol.2023.1144

[112] C. Munoz-Zuluaga, Z. Zhao, F. Wang, M. B. Greenblatt, and H. S. Yang, "Assessing the accuracy and clinical utility of CHATGPT in laboratory medicine," *Clinical Chemistry*, vol. 69, no. 8, pp. 939–940, May 2023. doi: 10.1093/clinchem/hvad058

[113] D. Dasari *et al.*, "ChatGPT in didactical tetrahedron, does it make an exception? A case study in mathematics teaching and learning,"

Frontiers in Education, vol. 8, Jan. 2024. doi: 10.3389/feduc.2023.1295413

[114] S. Zhang, C. Shan, J. S. Y. Lee, S. Che, and J. H. Kim, "Effect of chatbot-assisted language learning: A meta-analysis," *Education and Information Technologies*, vol. 28, no. 11, pp. 15223–15243, Apr. 2023. doi: 10.1007/s10639-023-11805-6

[115] M. A. M. Trindade, G. S. Edirisinghe, and L. Luo, "Teaching mathematical concepts in management with generative artificial intelligence: The power of human oversight in AI-driven learning," *The International Journal of Management Education*, vol. 23, no. 2, 101104, Dec. 2024. doi: 10.1016/j.ijme.2024.101104

[116] C.-W. Tsai, L. Lee, M. Y.-C. Lin, Y.-P. Cheng, C.-H. Lin, and M.-C. Tsai, "Effects of integrating self-regulation scaffolding supported by chatbot and online collaborative reflection on students' learning in an artificial intelligence course," *Computers & Education*, vol. 232, 105305, Mar. 2025. doi: 10.1016/j.compedu.2025.105305

[117] K. Almeman, F. E. Ayeb, M. Berrima, B. Issaoui, and H. Morsy, "The integration of AI and metaverse in education: A systematic literature review," *Applied Sciences*, vol. 15, no. 2, 863, Jan. 2025. doi: 10.3390/app15020863

[118] A. Gilson *et al*., "How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment," *JMIR Medical Education*, vol. 9, e45312, Feb. 2023. doi: 10.2196/45312

[119] B. Agyare, J. Asare, A. Kraishan, I. Nkrumah, and D. K. Adjekum, "A cross-national assessment of Artificial Intelligence (AI) Chatbot user perceptions in collegiate physics education," *Computers and Education Artificial Intelligence*, vol. 8, 100365, Jan. 2025. doi: 10.1016/j.caeai.2025.100365

[120] Md. R. Awal and Md. E. Haque, "Revisiting university students' intention to accept AI-Powered chatbot with an integration between TAM and SCT: A south Asian perspective," *Journal of Applied Research in Higher Education*, vol. 17, no. 2, pp. 594–608, Mar. 2024. doi: 10.1108/jarhe-11-2023-0514

[121] Y. Song, L. Huang, L. Zheng, M. Fan, and Z. Liu, "Interactions with generative AI chatbots: Unveiling dialogic dynamics, students' perceptions, and practical competencies in creative problem-solving," *International Journal of Educational Technology in Higher Education*, vol. 22, no. 1, Mar. 2025. doi: 10.1186/s41239-025-00508-2

[122] M. A. Kuhail, N. Alturki, S. Alramlawi, and K. Alhejori, "Interacting with educational chatbots: A systematic review," *Education and Information Technologies*, vol. 28, no. 1, pp. 973–1018, Jul. 2022. doi: 10.1007/s10639-022-11177-3

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).