

The Role of AI Chatbots in Student Learning: A Systematic Review

Pema Wangdi*, Sonam Pelden

Paro College of Education, Royal University of Bhutan, Paro, Bhutan

*Corresponding Author: pmwangdi@education.gov.bt

Received: 15 June 2025 / Accepted: 29 August 2025 / Published: 30 August 2025

Abstract: This systematic review evaluates the role of chatbots in student learning, focusing on engagement, motivation, feedback mechanisms, and academic performance. Following PRISMA 2020 guidelines, a structured search was conducted across multiple academic databases using predefined eligibility criteria. Studies published between 2020 and 2025 were analyzed through systematic data extraction and synthesis. Findings suggest chatbots enhance student engagement, provide structured feedback, and facilitate autonomous learning. However, challenges exist in chatbot adaptability, accuracy, and reliance on AI-generated content. Chatbots serve as valuable educational tools, but successful implementation requires structured integration and continuous refinement. AI systems need improved contextual accuracy and adaptability to enhance long-term effectiveness.

Keywords: AI Chatbots, Student Learning, Automated Tutoring, Educational Technology, Motivation

INTRODUCTION

The integration of artificial intelligence (AI) into educational environments is fundamentally transforming pedagogical methods, including how students acquire knowledge and how educators facilitate the learning process (Zawacki-Richter et al., 2019). One of the most prevalent and impactful applications of this technology is the AI-powered chatbot, which is a sophisticated computer program designed to simulate human-like conversation through natural language processing. These chatbots serve as dynamic educational tools that help learners by providing instant, personalized feedback, answering a wide range of questions, and guiding them through complex or difficult topics in a structured manner (Labadze et al., 2023). Furthermore, a significant proposed benefit of these systems is their capacity to support independent and self-directed learning by encouraging students to explore ideas and seek solutions autonomously, thereby fostering a sense of agency over their educational journey (Labadze et al., 2023).

As these AI-driven tools become increasingly common in both physical classrooms and digital learning platforms, the interest of the academic and research communities has correspondingly intensified, focusing on a rigorous empirical understanding of their efficacy, implementation challenges, and overall impact (Gökçearsan et al., 2024). The existing body of literature presents a nuanced picture. Many studies have reported positive findings, indicating that AI chatbots can significantly improve crucial educational metrics such as student motivation, cognitive and behavioral engagement, and the flexibility to learn at one's own pace (Gökçearsan et al., 2024). However, this optimistic view is tempered by considerable concerns and critical perspectives. Specifically, some researchers question the general reliability and contextual appropriateness of chatbots, expressing skepticism about whether their algorithmic responses are consistently accurate across diverse subject matters, and what the potential long-term cognitive and pedagogical effects of their sustained use might be (Labadze et al., 2023). This debate is crystallized in a central tension: while some scholars and practitioners believe AI chatbots are

instrumental in helping students become more independent and resourceful learners, others contend that an over-reliance on these automated tools might inadvertently reduce opportunities for developing critical thinking skills and diminish the frequency and quality of meaningful, interpersonal interactions with human teachers (Hartley et al., 2024).

Because of these conflicting and divergent opinions within the scholarly discourse, this systematic review aims to synthesize and consolidate the findings from a comprehensive range of research studies to construct a more coherent and evidence-based understanding of the effectiveness of AI chatbots in student learning. This review will systematically examine the empirical evidence to illuminate both the documented strengths and the potential risks associated with chatbot integration. The ultimate objective is to provide a balanced and thorough analysis that can inform educators, instructional designers, and policymakers on how these innovative tools can be most effectively and responsibly utilized in future educational contexts.

This systematic review seeks to answer the following research question: “How do AI chatbots help in student learning?”. More precisely, it aims to:

1. Assess the impact of AI chatbots on student engagement, motivation, and learning outcomes.
2. Evaluate AI chatbot-assisted feedback mechanisms and their role in improving student comprehension.
3. Identify limitations in AI chatbot adaptability, accuracy, and integration into formal education.

METHOD

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines (Page et al., 2021). Adhering to this established framework ensured that the processes for study selection, data extraction, and synthesis of findings were transparent, methodical, and replicable. The application of these guidelines serves to minimize selection and reporting biases, enhance the overall transparency of the methodological approach, and thereby support a comprehensive and unbiased evaluation of the existing empirical evidence concerning the impact of AI chatbots on student learning.

Eligibility Criteria

The inclusion and exclusion criteria were meticulously developed a priori to systematically identify studies that provide robust, empirical evidence directly pertaining to chatbot-assisted student learning, while simultaneously filtering out research that is irrelevant, lacks methodological rigor, or falls outside the defined scope. Table 1 provides a detailed outline of the specific inclusion and exclusion criteria applied during the screening process.

Information Sources

A comprehensive, multi-database search strategy was employed to ensure extensive coverage of the relevant literature and to mitigate the risk of publication bias. The primary electronic databases searched included Google Scholar, IEEE Xplore, Scopus, and ERIC (Education Resources Information Center). The ERIC database was prioritized due to its specialization in educational research and its extensive repository of peer-reviewed journal articles and reports (Institute of Education Sciences [IES], 2024). To further identify potentially relevant grey literature and unpublished studies, the search strategy was supplemented by manually examining key conference proceedings in educational technology, scanning systematic review repositories, and conducting direct inquiries to corresponding authors in the field.

Search Strategy

A structured Boolean search protocol was designed and executed to retrieve a comprehensive set of studies relevant to chatbot-assisted learning, following established systematic

review practices (Booth et al., 2016). The specific search queries, which were adapted for syntax across the different databases, included the following primary and secondary strings:

1. Primary search string:
("AI chatbot" OR "automated tutoring system" OR "conversational agent") AND ("student learning" OR "academic improvement" OR "engagement")
2. Secondary search string:
("chatbot feedback" OR "AI-assisted instruction") AND ("writing proficiency" OR "motivation" OR "learning outcomes")

To focus the review on the most current technological iterations, database filters were applied to limit results to peer-reviewed journal articles published between January 2020 and December 2024. The final search across all designated databases was conducted and updated on May 20, 2024.

Table 1. Inclusion and Exclusion Criteria

| Inclusion Criteria | Exclusion Criteria |
|---|---|
| 1. Studies examining chatbot effectiveness in enhancing engagement, motivation, comprehension, or feedback mechanisms. | 1. Studies focusing on AI chatbots in customer service, healthcare, business automation, or other non-educational contexts. |
| 2. Studies reporting measurable learning outcomes via quantitative, quasi-experimental, or mixed-methods research designs. | 2. Conceptual papers, opinion pieces, editorials, and theoretical discussions lacking original empirical research or primary data. |
| 3. Studies published between 2020 and 2024 to reflect recent advancements in AI chatbot technology and its pedagogical applications. | 3. Studies with small sample sizes (e.g., $n < 20$), inadequate control groups, unclear findings, or evident methodological bias that threatened validity. |
| 4. Research set in primary, secondary, or higher education, including contexts such as language learning, STEM, and general pedagogy. | 4. Duplicative studies reporting the same dataset or findings, where only the most comprehensive or original version was retained. |
| 5. Only studies with full-text availability in English to support in-depth analysis and ensure accurate data extraction. | 5. Research focused exclusively on chatbots in corporate or adult professional training contexts without direct relevance or application to formal student education. |

Study Selection Process

The study selection process was rigorously conducted in multiple stages, following the PRISMA 2020 flow diagram and informed by the PICO (Population, Intervention, Comparison, Outcome) framework to refine relevance. The initial database searches identified a total of 300 records. After the removal of 120 duplicate entries, 180 studies remained for the title and abstract screening phase. Two independent reviewers screened these 180 records for relevance based on the predefined inclusion criteria. This initial screening led to the exclusion of 110 studies that either did not focus on chatbot-assisted learning in an educational context or clearly lacked empirical evidence. The remaining 70 full-text articles were retrieved and assessed in detail for eligibility. Studies were excluded at this stage for the following primary reasons: they did not evaluate specific educational outcomes; they focused on non-student populations (e.g., professional trainees); they lacked empirical testing of chatbot efficacy; or they had no control group or valid comparison for assessing effectiveness. This detailed assessment resulted in the exclusion of 56 studies. Ultimately, 14 studies were deemed to meet all eligibility criteria and were included in the final qualitative synthesis. The complete screening and inclusion process is depicted in [Figure 1](#).

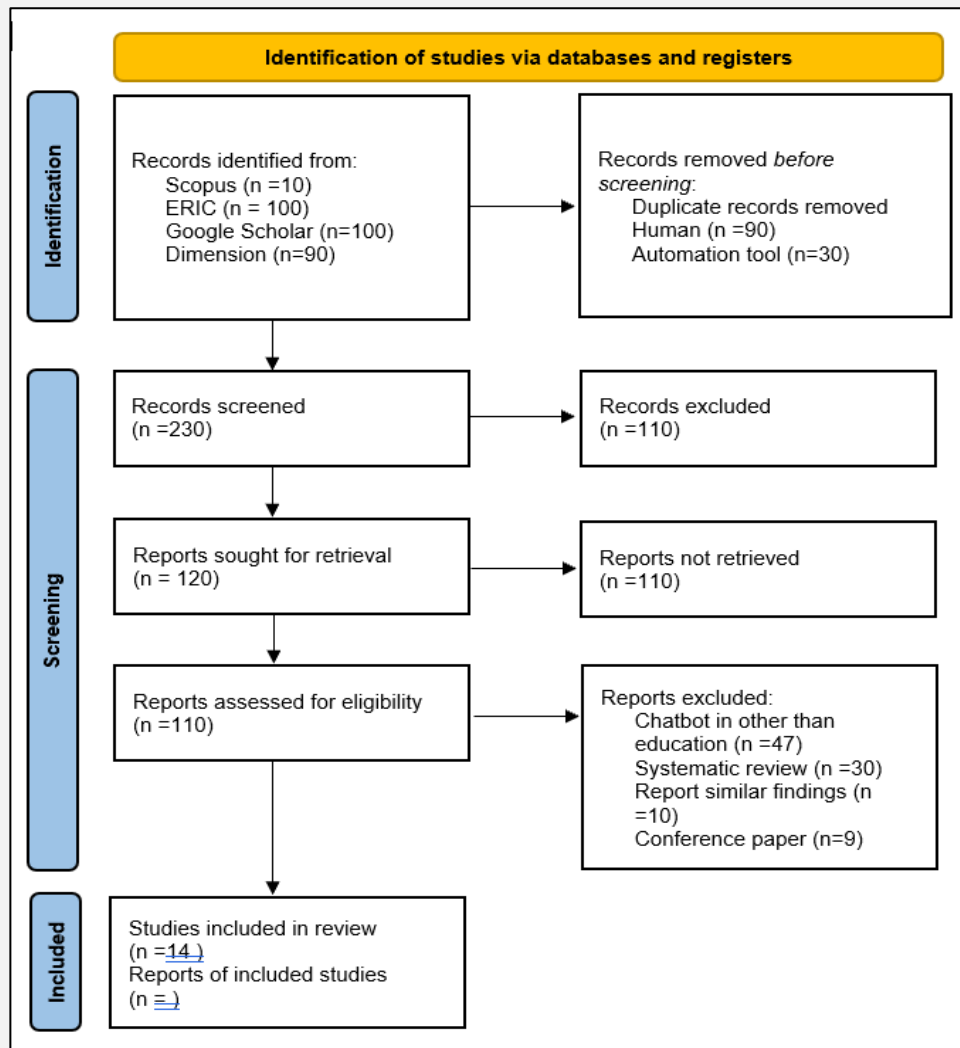


Figure 1. PRISMA 2020 Flow Diagram of Study Selection Process. This diagram illustrates the identification, screening, eligibility, and inclusion stages of the 14 studies analyzed in this review, following PRISMA 2020 guidelines.

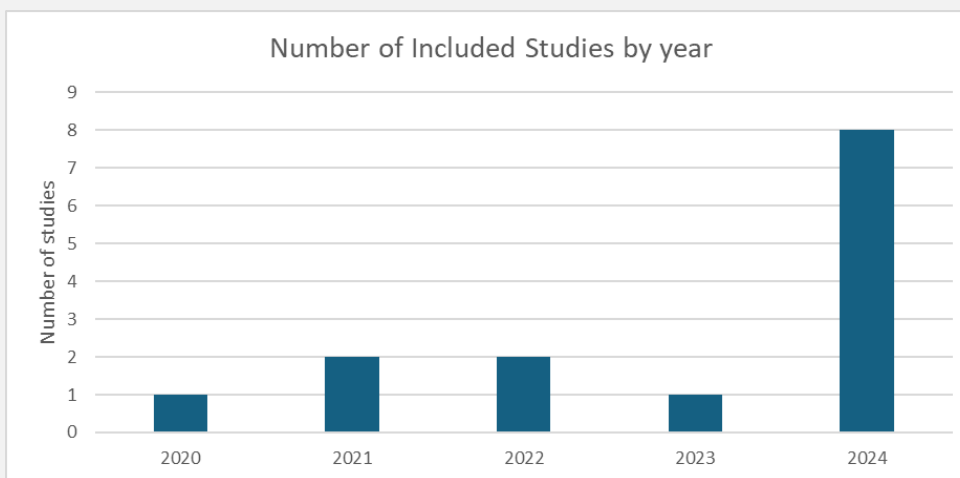


Figure 2. Distribution of Selected Studies by Publication Year.

Data Collection Process

Data were systematically collected from the included studies using a standardized, piloted data extraction form, developed to ensure that key information was recorded consistently and accurately across all studies (Büchter et al., 2020). This structured approach supported a uniform comparison and synthesis of findings. To enhance reliability, two reviewers independently extracted the data into the form. Any discrepancies or disagreements between the reviewers during the extraction process were resolved through discussion until a consensus was reached. The extracted information included: study title, author(s) and publication year, primary research objectives, sample size and characteristics (e.g., age, educational level), research methodology and design, type of chatbot technology implemented, key outcome measures, and main findings.

Study Variables

The data synthesis focused on two main categories of variables—outcome measures and AI chatbot characteristics—to ensure a clear and structured interpretation of the results across the diverse set of included studies.

1. Outcome Measures

- Student engagement levels (e.g., AI chatbot interaction frequency, self-reported motivation changes).
- Writing proficiency improvements (e.g., revision habits, error identification accuracy, sentence structuring ability).
- Learning performance changes (e.g., pre/post-test scores, grades, measures of conceptual understanding).

2. Other Variables

- AI Chatbot technology types (e.g., rule-based vs. generative AI models like GPT).
- Intervention duration (e.g., short-term [< 4 weeks] vs. long-term [≥ 4 weeks] chatbot usage).
- Learning domains (e.g., language education, STEM subjects, general pedagogy).

Study Characteristics and Geographic Distribution

The 14 studies included in this final review represent a diverse range of geographic and educational contexts. The majority of the studies were conducted in the United States and China (4 studies each), followed by Turkey (2 studies). Additional studies contributed data from Ghana, Oman, the United Kingdom, and Saudi Arabia. The research encompassed a variety of student populations, including:

- EFL learners in China, Saudi Arabia, and Oman, where chatbots were primarily used for writing assistance, vocabulary acquisition, and grammar support.
- Higher education students in the US and UK, where studies assessed the effectiveness of AI tutors in facilitating complex problem-solving tasks and providing personalized feedback in subjects like computer science and engineering.
- Primary and secondary school learners in Turkey and Ghana, with a focus on AI chatbot-assisted learning in mathematics and social studies to bolster foundational knowledge and engagement.

The geographic distribution of the selected studies, illustrating the number of studies per country, is presented in [Figure 3](#).

Risk of Bias Assessment

The methodological quality and potential risk of bias for each included study were critically evaluated using the Cochrane Risk of Bias (RoB 2) tool. This assessment was crucial for ensuring the overall methodological rigor and reliability of the data contributing to the review's conclusions. The evaluation focused on several key domains of potential bias:

1. *Selection Bias*: Was the random sequence generation and allocation concealment adequately described and implemented?

2. *Performance Bias*: Were participants and personnel blinded to the intervention, or could knowledge of the assigned intervention have influenced the outcomes?
3. *Measurement Validity*: Were the outcome assessors blinded, and were the tools used to measure learning outcomes objective and validated?
4. *External Influences*: Were there any major confounding factors or external events that could have significantly affected the study's conclusions? (Boutron et al., 2024).

To maintain the highest levels of transparency and credibility, any studies judged to have a high overall risk of bias were excluded from the final synthesis, as per Cochrane recommendations (Boutron et al., 2024). This ensured that only methodologically sound research contributed to the findings and conclusions of this systematic review.

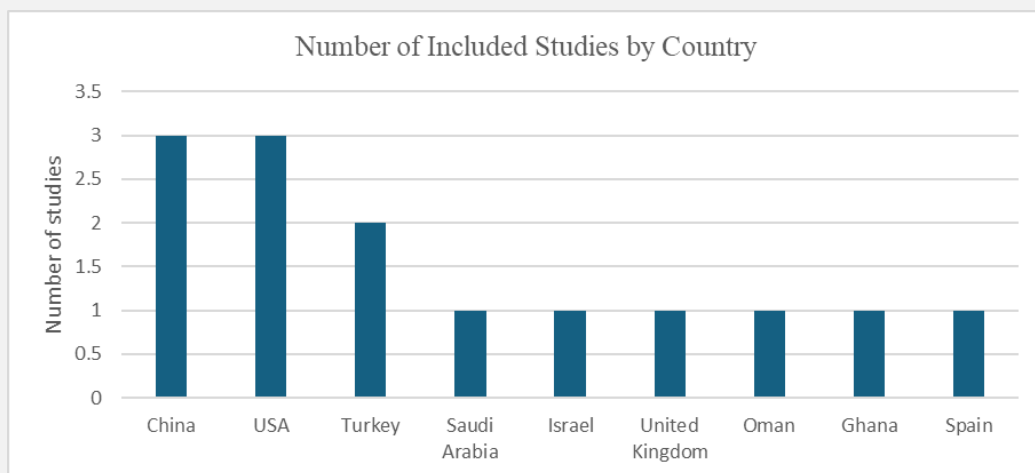


Figure 3. Number of selected studies per country.

RESULTS

This section presents a synthesis of the empirical findings from the 14 studies included in this systematic review, analyzing the role of AI chatbots in student learning. The results are organized around three central themes: (1) student engagement and motivation, (2) feedback mechanisms and proficiency development, and (3) subject-specific learning outcomes. Subsequently, identified challenges and implementation constraints are highlighted. The distribution of these themes across the selected studies is illustrated in Figure 4 below.

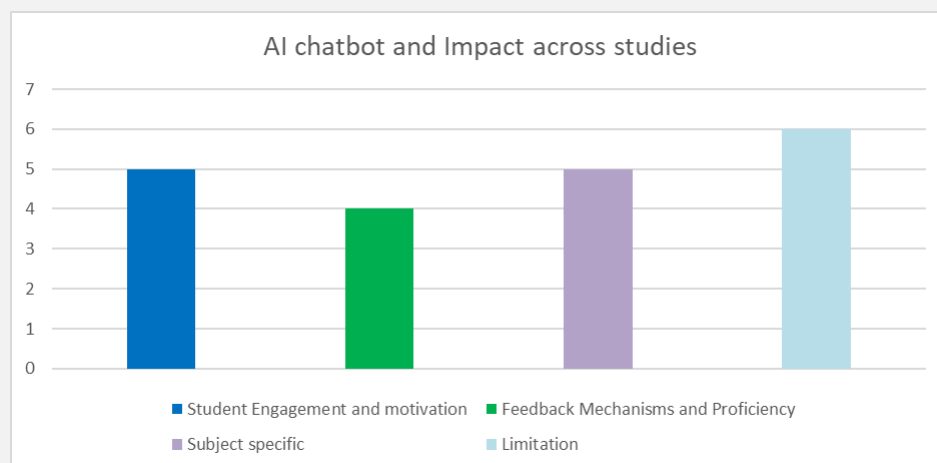


Figure 4. AI Chatbot Impacts and Themes Across Selected Studies.

Chatbots and Student Engagement & Motivation

Across diverse educational contexts, the integration of AI chatbots was found to significantly improve learner engagement and intrinsic motivation. For instance, in a study with Saudi Arabian postgraduate students, those using the Bashayer chatbot demonstrated higher levels of motivation and reported increased use of cognitive and metacognitive strategies compared to a control group (Al-Abdullatif et al., 2024). Similarly, research with undergraduate students in the U.S. indicated an 80% self-reported attentiveness rate when using AI tools, contrasted with 50% in traditional lecture-based classrooms, alongside a measurable 25% improvement in test scores (Gada & Chudasama, 2024). These findings collectively display AI's considerable potential to personalize and invigorate the learning experience through interactive, on-demand support.

However, the outcomes related to engagement were not universally positive and revealed important nuances. While structured chatbot responses were highly beneficial for novice learners, more advanced students often described them as restrictive and lacking depth. This dichotomy was clearly illustrated in the study by Yin & Satar (2020), which found that low-proficiency EFL learners appreciated the scaffolding provided by the "Tutor Mike" chatbot, whereas high-proficiency learners in the same study reported low engagement due to the chatbot's limited conversational and analytical capabilities. This suggests that the effectiveness of chatbot-driven engagement is highly contingent upon the alignment between the tool's functionality and the learner's existing competency level.

Table 2. Engagement and Motivation Gains Across Selected Studies

| Study | Engagement (% Chatbot) | Engagement (% Control) | Motivation Rating (1–5) |
|-----------------------------|------------------------|------------------------|-------------------------|
| Gada & Chudasama (2024) | 80% | 50% | 4.5 (vs. 3.0 control) |
| Al Abdullatif et al. (2024) | High (qualitative) | Moderate | Significant increase |
| Yin & Satar (2020) | High (low-level only) | NA | Mixed by proficiency |

Chatbot-Assisted Feedback and Writing Proficiency

AI chatbots were consistently identified as highly useful in delivering immediate and structured feedback, with a pronounced impact in the domain of writing instruction. A common preference emerged for a hybrid feedback model. Zeevy-Solovey (2024) found that while most EFL students still preferred ultimate feedback from a human teacher, a significant number also valued ChatGPT for its rapid and well-organized preliminary suggestions. Similarly, Wang & Han (2022) observed that although teacher-led feedback resulted in higher-quality immediate revisions, sustained interaction with automated feedback systems contributed significantly to the longitudinal development of overall writing proficiency.

Extending these insights, Song & Song (2023) demonstrated that AI-assisted learning with ChatGPT significantly improved EFL learners' textual coherence, grammatical accuracy, and learning motivation. However, this study also raised pertinent concerns regarding the contextual accuracy of AI feedback and the risk of student over-reliance, noting that learners occasionally accepted the chatbot's flawed or simplistic suggestions uncritically, potentially undermining the development of their own editing skills.

Subject-Specific Learning Support

AI chatbots have demonstrated significant potential to enhance student performance across a variety of discipline-specific domains. In mathematics education, Karaman & Göksu (2024) reported that lesson plans generated by ChatGPT led to substantial improvements in student achievement, evidenced by a large effect size ($d = 1.268$), though they emphasized that teacher interpretation remained crucial for clarifying complex content. Similarly, Yetişensoy & Karaduman (2024) found that chatbot-assisted instruction in social studies improved both immediate academic achievement and long-term knowledge retention, a finding further supported by positive qualitative feedback from both students and teachers regarding the instructional design.

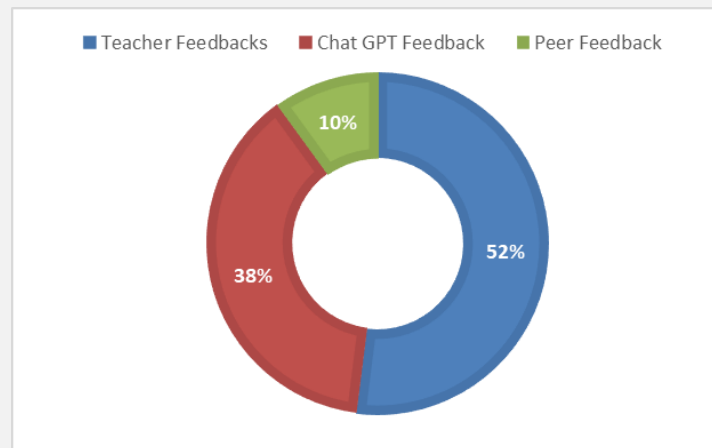


Figure 5. Feedback Quality Comparison (Perceived value of feedback modalities based on Solovey, 2024)

In language education, [Vázquez-Cano et al. \(2021\)](#) observed that Spanish university access students using a chatbot built on the Collect.Chat platform showed greater progress in specific punctuation skills, particularly in the correct use of commas, colons, and periods, compared to peers who relied solely on traditional paper-based materials. Meanwhile, in computer science, [Essel et al. \(2022\)](#) found that Ghanaian students using the "KNUSTbot" outperformed those taught exclusively through human instruction on programming assessments. Although students reported high levels of engagement with the chatbot, they concurrently noted that its explanations for complex programming concepts occasionally lacked the necessary depth for full comprehension.

Table 3. Subject-Specific Gains with Chatbot Intervention

| Study | Domain | Outcome | Note |
|-------------------------------|----------------|--------------------------------------|---|
| Karaman & Göksu (2024) | Math | ↑ Test scores ($d = 1.268$) | No stat. sign. vs. control |
| Yetişensoy & Karaduman (2024) | Social Studies | ↑ Achievement & retention | Teacher praised design |
| Vázquez-Cano et al. (2021) | Punctuation | ↑ Accuracy in comma/colon/period use | Students preferred chatbot to paper tasks |
| Essel et al. (2022) | Programming | ↑ Performance + engagement | Needed deeper explanations |

Identified Limitations and Technical Challenges

Despite the reported benefits, multiple studies highlighted persistent technical and pedagogical challenges that temper the overall enthusiasm for chatbot integration. The primary limitations identified are as follows:

1. ChatGPT and similar tools provided structured help but sometimes critically lacked flexibility and nuanced understanding, which frustrated high-level learners seeking sophisticated dialogue ([Yin & Satar, 2020](#); [Zeevy-Solovey, 2024](#)).
2. [Pardos & Bhandari \(2024\)](#) reported that ChatGPT's generated hints failed quality checks in 32% of mathematics tasks; however, they also noted that employing self-consistency prompting techniques substantially reduced this error rate to 0% for algebra and 13% for statistics.
3. Both [Song & Song \(2023\)](#) and [Hutt et al. \(2024\)](#) caution that students may accept AI suggestions uncritically, which can weaken the development of their independent evaluative and critical thinking skills.

4. In feedback comparison studies, AI-generated explanations were consistently found to lack the empathetic, context-aware quality of human feedback, especially for rhetorical, creative, or metacognitive strategies (Wang & Han, 2022; Hutt et al., 2024).

Summary of Findings

The reviewed studies collectively suggest that AI chatbot interventions can positively impact student learning when they are effectively and thoughtfully implemented. The evidence indicates that chatbots have been shown to improve engagement and motivation, particularly in learning environments that benefit from structured guidance and consistent support. Furthermore, chatbots provide efficient, immediate feedback, which contributes to the gradual development of technical and compositional writing skills. Moreover, their application enhances subject comprehension across both STEM and humanities disciplines. However, the findings uniformly emphasize that successful integration requires thoughtful instructional design and continuous human oversight to mitigate inherent challenges such as response rigidity, factual inaccuracy, or excessive learner dependence.

DISCUSSION

The analysis of the 14 studies affirms that AI-powered chatbots can serve as transformative educational tools when deployed thoughtfully in appropriate contexts. However, their effectiveness is decidedly mixed and is influenced by a complex interplay of factors, including pedagogical design, learner profiles, the method of feedback integration, and the technological maturity of the chatbot systems themselves.

Personalization and Engagement: Chatbots as Cognitive Scaffolds

Many of the reviewed studies underscore the chatbots' ability to function as effective cognitive scaffolds, sustaining learner attention, especially among students who benefit from structured pacing and low-stakes, repetitive interaction. For example, Gada & Chudasama (2024) reported a 30% increase in attentiveness and learning outcomes with AI support, while Al-Abdullatif et al. (2024) observed enhanced motivation and metacognitive strategy use among Saudi postgraduate learners. In these instances, chatbots functioned not merely as instructional tools but as dynamic cognitive prompts—responding instantly, reinforcing key concepts, and creating a psychologically safe space for trial-and-error learning.

However, Yin & Satar's (2020) mixed findings serve as a critical reminder that engagement is not universal. While less proficient learners found the scaffolding invaluable, advanced learners perceived chatbots like Tutor Mike as overly constrained, lacking the semantic range and responsive subtlety needed for more complex, abstract dialogue. This highlights a fundamental takeaway: chatbot effectiveness is relational and contextual, not absolute—it is profoundly shaped by individual learner profiles and specific task complexity.

Feedback Evolution: Efficient but not Exhaustive

The provision of feedback emerged as the second most consistently cited benefit across the literature. AI chatbots like ChatGPT offer immediacy and clarity—attributes that learners explicitly valued for revising written work (Zeevy-Solovey, 2024; Wang & Han, 2022). This automated feedback not only improved surface-level features like grammar and structural precision but also allowed students to iterate more rapidly, particularly in time-bound instructional contexts.

Yet, as the works of Hutt et al. (2024) and Pardos & Bhandari (2024) reveal, the quality and reliability of AI-generated feedback are highly contingent on the disciplinary context and the implementation of advanced prompting techniques. In scenarios involving peer feedback and complex math problem-solving, AI-generated hints were found to be less reliable without technical interventions. Moreover, students do not always differentiate between the chatbot's confident tone and the actual correctness of its suggestions, making ongoing educator supervision crucial in high-stakes or high-cognitive-load tasks. Consequently, AI feedback systems are best understood

as powerful accelerators within the learning process, not as replacements for expert human guidance (Shi & Aryadoust, 2024).

Extending Curriculum Reach: AI Chatbots Across Domains

In disciplines ranging from mathematics and punctuation to social studies and programming, chatbots demonstrated clear utility in helping students visualize step-based learning processes and self-correct routine errors. Karaman & Göksu's (2024) study in math instruction showed large effect sizes when ChatGPT-generated lesson plans were utilized by teachers. Similarly, Essel et al. (2022) and Yetişensoy & Karaduman (2024) highlighted chatbot benefits in higher education and middle school classrooms, respectively, though both noted that deep explanatory support for sophisticated concepts still necessitated a teacher's presence.

Notably, the Vázquez-Cano et al. (2021) study offers a compelling use case for chatbot integration in foundational skill building, where the structure, predictability, and endless patience of chatbot responses were viewed by students as a distinct advantage rather than a limitation.

Ethical and Pedagogical Tensions: Dependency, Accuracy, and Trust

Despite the promising applications, significant limitations persist. AI chatbot adaptability continues to lag behind human instruction, with several studies citing concerns about inaccurate, overly generic, or conceptually shallow responses—especially in open-ended writing and complex problem-solving scenarios (Song & Song, 2023; Pardos & Bhandari, 2024). An overreliance on chatbots may also risk deskilling learners in critical areas such as independent judgment, source evaluation, and nuanced error correction (Zhai et al., 2024).

The variability in student trust reported across studies further complicates adoption. Some learners appreciate AI guidance but struggle with the fundamental ambiguity about when to trust the system and when to doubt its outputs (Bećirović et al., 2025). This suggests an urgent need for parallel digital literacy programs that teach students not just *how* to use chatbots, but *how to critically question and verify* their recommendations.

Implications for Policy, Practice, and Future Research

For educators and policymakers, the evidence suggests that AI chatbots should be integrated deliberately by linking them to specific instructional tasks that leverage their strengths in structure, repetition, and immediate response, such as formative assessment and low-stakes practice (Winkler & Söllner, 2021). To maximize effectiveness, the development of hybrid feedback models is highly recommended. These models would merge the efficiency and scalability of AI-generated responses with the depth, adaptability, and empathy of human insight (Holstein et al., 2019). This combined approach not only improves the timeliness of feedback but also ensures nuanced evaluation where it is most needed.

From a technological perspective, advancing chatbot design through more adaptive natural language processing and context-aware algorithms is essential to better accommodate diverse learner needs and linguistic complexity (Zawacki-Richter et al., 2019). Simultaneously, embedding digital literacy and critical judgment training directly into the curriculum can help students become discerning consumers of AI recommendations, rather than passive, uncritical users (Luckin et al., 2016). Future research should prioritize longitudinal and cross-cultural studies to explore how sustained chatbot interaction influences deeper educational constructs, including learners' cognitive autonomy, academic identity formation, and evolving trust in AI systems (Pimmer et al., 2020).

CONCLUSION

This systematic review, synthesizing evidence from 14 studies published between 2020 and 2024, affirms that AI-powered chatbots are increasingly shaping modern educational landscapes by enhancing student engagement, improving learning outcomes, and delivering scalable instructional support. The evidence demonstrates measurable benefits across varied academic contexts, from improving writing proficiency through structured feedback to elevating motivation and achievement in disciplines such as mathematics, programming, language, and social studies.

However, these benefits are neither uniform nor unconditional. The review reveals significant constraints, including the occasional lack of depth and accuracy in AI-generated feedback, which underscores the indispensable role of educator oversight. Furthermore, advanced learners often find chatbot interactions restrictive, and there are consistent concerns regarding student over-reliance, which can potentially diminish critical thinking and metacognitive skills. These findings necessitate a cautious and deliberate implementation strategy.

To maximize the educational value of this technology, chatbots should be strategically positioned as augmentative tools within a hybrid instructional model, rather than as replacements for human instruction. Effective integration requires a multi-stakeholder approach. Educators should embed chatbots into formative assessments with explicit scaffolding and maintain active roles in validating AI feedback. Policymakers must establish guidelines for ethics and data privacy while supporting digital literacy initiatives. Developers are urged to enhance the contextual accuracy and adaptability of chatbots and to design interfaces that promote learner verification and metacognitive skill development.

Future research should prioritize longitudinal and cross-cultural studies to examine the sustained impact of chatbot interaction on learner autonomy, digital trust, and academic identity. Ultimately, while chatbot-assisted learning offers great promise, its potential to truly empower learners in a digitally enriched future is contingent upon thoughtful integration, continuous pedagogical refinement, and unwavering ethical vigilance.

Acknowledgements

I would like to extend my sincere gratitude to all individuals who contributed, both directly and indirectly, to the completion of this systematic review. Your support and insights have been invaluable.

REFERENCES

- Al-Abdullatif, A. M., Al-Dokhny, A. A., & Drwish, A. M. (2023). Implementing the Bashayer chatbot in Saudi higher education: measuring the influence on students' motivation and learning strategies. *Frontiers in Psychology, 14*, 1129070. <https://doi.org/10.3389/fpsyg.2023.1129070>
- Bećirović, S., Polz, E., & Tinkel, I. (2025). Exploring students' AI literacy and its effects on their AI output quality, self-efficacy, and academic performance. *Smart Learning Environments, 12*(1), 1-25. <https://doi.org/10.1186/s40561-025-00384-3>
- Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic approaches to a successful literature review* (2nd ed.). SAGE Publications.
- Boutron, I., Page, M. J., Higgins, J. P., Altman, D. G., Lundh, A., Hróbjartsson, A., & Cochrane Bias Methods Group. (2019). Considering bias and conflicts of interest among the included studies. *Cochrane handbook for systematic reviews of interventions, 177-204*. <https://doi.org/10.1002/9781119536604.ch7>
- Büchter, R. B., Weise, A., & Pieper, D. (2020). Development, testing and use of data extraction forms in systematic reviews: a review of methodological guidance. *BMC Medical Research Methodology, 20*(1), 259. <https://doi.org/10.1186/s12874-020-01143-3>
- Essel, H. B., Vlachopoulos, D., Tachie-Menson, A., Johnson, E. E., & Baah, P. K. (2022). The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *International Journal of Educational Technology in Higher Education, 19*(1), 1-19. <https://doi.org/10.1186/s41239-022-00362-6>
- Gada, T., & Chudasama, S. (2024). Impact Of Artificial Intelligence on Student Attitudes, Engagement, and Learning. *International Research Journal of Modernization in Engineering Technology and Science, 6*(5), 2695-2698. <https://doi.org/10.56726/irjmets56247>
- Ghailthi, A. A., & Behforouz, B. (2024). The use of an interactive chatbot in grammar learning. *The Journal of Educators Online, 21*(4). <https://doi.org/10.9743/jeo.2024.21.4.12>
- Gökçearslan, Ş., Tosun, C., & Erdemir, Z. G. (2024). Benefits, Challenges, and Methods of Artificial intelligence (AI) chatbots in Education: A Systematic Literature review. *International Journal of Technology in Education, 7*(1), 19–39. <https://doi.org/10.46328/ijte.600>

- Hartley, K., Hayak, M., & Ko, U. H. (2024). Artificial intelligence Supporting Independent Student learning: An evaluative case study of CHATGPT and Learning to Code. *Education Sciences, 14*(2), 120. <https://doi.org/10.3390/educsci14020120>
- Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics, 6*(2), 27–52. <https://doi.org/10.18608/jla.2019.62.3>
- Hutt, S., DePiro, A., Wang, J., Rhodes, S., Baker, R. S., Hieb, G., Sethuraman, S., Ocumpaugh, J., & Mills, C. (2024). Feedback on feedback: comparing classic natural language processing and generative AI to evaluate peer feedback. *Association for Computing Machinery, 55–65*. <https://doi.org/10.1145/3636555.3636850>
- Karaman, M. R., & Göksu, İ. (2024). Are lesson plans created by ChatGPT more effective? an experimental study. *International Journal of Technology in Education, 7*(1), 107–127. <https://doi.org/10.46328/ijte.607>
- Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education, 20*(1), 56. <https://doi.org/10.1186/s41239-023-00426-1>
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson. <https://oro.open.ac.uk/50104/>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ, 372*. <https://doi.org/10.1136/bmj.n71>
- Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS ONE, 19*(5), e0304013. <https://doi.org/10.1371/journal.pone.0304013>
- Pimmer, C., Daniels, F., Ologun-Abiodun, R., & Chipps, J. (2020). A WhatsApp community of practice for nurse graduates in South Africa. *Nurse Education in Practice, 46*, 102826. <https://doi.org/10.1016/j.nepr.2020.102826>
- Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL, 36*(2), 187–209. <https://doi.org/10.1017/s0958344023000265>
- Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*. <https://doi.org/10.3389/fpsyg.2023.1260843>
- Vázquez-Cano, E., Mengual-Andrés, S., & López-Meneses, E. (2021). Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments. *International Journal of Educational Technology in Higher Education, 18*(1), 33. <https://doi.org/10.1186/s41239-021-00269-8>
- Wang, Z., & Han, F. (2022). The effects of teacher feedback and automated feedback on cognitive and psychological aspects of foreign language writing: A Mixed-Methods Research. *Frontiers in Psychology, 13*, 909802. <https://doi.org/10.3389/fpsyg.2022.909802>
- Winkler, R., Söllner, M., & Leimeister, J. M. (2021). Enhancing problem-solving skills with smart personal assistant technology. *Computers & Education, 165*, 104148. <https://doi.org/10.1016/j.compedu.2021.104148>
- Yetişensoy, O., & Karaduman, H. (2024). The effect of AI-powered chatbots in social studies education. *Education and Information Technologies, 29*(13), 17035–17069. <https://doi.org/10.1007/s10639-024-12485-6>
- Yin, Q., & Satar, M. (2020). English As a Foreign Language Learner Interactions with Chatbots: Negotiation for Meaning. *International Online Journal of Education and Teaching, 7*(2), 390-410. <https://iojet.org/index.php/IOJET/article/view/707>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27. <https://doi.org/10.1186/s41239-019-0171-0>
- Zeevy-Solovey, O. (2024). Comparing peer, ChatGPT, and teacher corrective feedback in EFL writing: Students' perceptions and preferences. *Technology in Language Teaching & Learning, 6*(3), 1482. <https://doi.org/10.29140/tltl.v6n3.1482>

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1), 28.
<https://doi.org/10.1186/s40561-024-00316-7>



Copyright (c) 2025 by the authors. This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.