KARYA ILHAM PUBLISHING
CONNECTING MINDS, BRIDGING IDEAS, INSPIRING READERS

Journal of Advanced Research in Computing and Applications

Journal homepage:
https://karyailham.com.my/index.php/arca
2462-1927

The Effectiveness of AI Chatbots as Learning Assistants in Universities: A Systematic Review

Nur Ashraf Mohklis Jahiri[1], Nur Afiqah Mahadhir[1], Nuruljannah Khairul Anuar[1], Nur Nadhirah Husna Kamarul Sham[1], Mohamad Zulhusni Rosmi[1], Che Ghani Che Kob[1,*]

[1] Universiti Pendidikan Sultan Idris, Tanjong Malim, 35900, Malaysia

ARTICLE INFO

Article history:
Received 15 May 2026
Received in revised form 10 June 2026
Accepted 10 June 2026
Available online 6 July 2026

Keywords:
AI chatbot; university learning; systematic review; higher education

ABSTRACT

This Systematic Literature Review (SLR) evaluates the effectiveness of AI chatbots as learning assistants in universities, with a focus on enhancing student engagement, academic performance, and implementation challenges. Using the PRISMA protocol, the study analyzes 20 empirical studies published between 2020 and 2025 from databases such as Scopus, Web of Science, ERIC, IEEE Xplore, and Google Scholar. The findings indicate that AI chatbots generally enhance student engagement and conceptual understanding, with improvements in academic scores particularly evident in interactive courses. However, several challenges are frequently reported, including the accuracy of AI responses, data privacy concerns, and technological accessibility. The study concludes that AI chatbots serve as supportive pedagogical tools rather than full replacements for human instruction. Recommendations include careful evaluation of effectiveness and further research on long-term impacts.

1. Introduction

Quality education is a key driver of sustainable development and is explicitly highlighted in the United Nations Sustainable Development Goal 4 (SDG 4), which aims to ensure inclusive, equitable and quality education and promote lifelong learning opportunities for all. As the lead United Nations agency for education, UNESCO emphasises the strategic and ethical integration of digital technologies, including artificial intelligence to enhance teaching and learning processes in higher education while ensuring inclusivity and educational quality. In line with these global priorities, Malaysia has increasingly focused on innovation-driven and technology-enhanced educational approaches to support the development of future-ready graduates, an orientation that is often reflected in local initiatives associated with sustainable and digitally oriented education, including those framed under the notion of Blue Education. Within this context, AI chatbots have emerged as a promising educational tool with the potential to support personalised learning, provide immediate

* *Corresponding author.*
E-mail address: cheghani@ftv.upsi.edu.my

https://doi.org/10.37934/arca.43.1.235248

feedback and enhance student engagement in university settings. However, despite their growing adoption, empirical evidence regarding the effectiveness of AI chatbots as learning assistants in higher education remains fragmented and inconclusive. Therefore, this systematic review aims to synthesise existing research on the effectiveness of AI chatbots in university education with particular attention to learning outcomes, psychological effects and implementation factors.

AI chatbots demonstrate medium-to-high effectiveness as learning assistants in university education [6], with particularly strong impacts on explicit knowledge transmission tasks. Evidence across 20 systematic reviews and primary studies shows significant improvements in vocabulary acquisition, grammar mastery [18], knowledge retention and learning achievement [6], alongside quantified gains including 14.65% improvement in academic performance and 70.6% increase in learning retention [4]. Psychological benefits include reduced language anxiety [18] and increased learner confidence [14], with student dropout rates decreasing by 8.29% [4]. However, chatbots show no significant effects on critical thinking, learning engagement or motivation [19], with concerns about reduced creativity and collaborative learning capacity [10].

Effectiveness depends strongly on context and the quality of implementation. Blended learning approaches integrating chatbot use with human facilitation emerge as most effective [18], while purely AI-driven implementations show weaker outcomes. Success factors include structured training [2], regular sustained use [14], and integration with teacher support [2], with tech-savvy students benefiting most [14]. Significant challenges include technical limitations such as accuracy issues [2], ethical concerns regarding data privacy and algorithmic bias [18], pedagogical risks of over-reliance [10], and equity barriers for students lacking digital literacy or infrastructure [18]. The evidence indicates AI chatbots are most effective for structured knowledge acquisition and language practice in well-supported implementations, but cannot replace human instruction for higher-order thinking skills, with the technology remaining in early developmental stages [19].

This systematic literature review contributes to the growing body of research on AI chatbots in higher education by providing an integrative, mechanism-oriented synthesis of existing empirical and review-based evidence. Unlike prior studies that predominantly catalogue learning outcomes or focus on technological affordances in isolation, this review simultaneously examines cognitive, psychological and pedagogical dimensions of chatbot effectiveness. Specifically, this study makes three key contributions. First, it synthesises evidence on academic performance, skill development and knowledge retention alongside affective mechanisms such as learner confidence and anxiety reduction, offering a more holistic understanding of how AI chatbots influence learning. Second, it identifies critical implementation factors, including blended learning approaches, structured training and sustained usage, as moderators that shape effectiveness across contexts. Third, the review proposes an implicit conceptual framework linking AI chatbot features, mediating mechanisms and learning outcomes, thereby providing a structured lens to guide future research and pedagogical decision-making in university settings.

2. Methodology

2.1 Paper Search

This systematic literature review followed the PRISMA 2020 guidelines. A comprehensive literature search was conducted across five major academic databases: Scopus, Web of Science, ERIC, IEEE Xplore and Google Scholar. The search was limited to peer-reviewed journal articles published between 2020 and 2025 and written in English. The following search string was used with minor adaptations across databases:

("AI chatbot*" OR "educational chatbot*" OR "ChatGPT") AND ("higher education" OR "university" OR "tertiary education") AND ("learning outcome*" OR "academic performance" OR "student learning").

In addition, the Elicit semantic search engine was used as a supporting tool to cross-check relevant studies and ensure comprehensive coverage of recent publications. Records identified through Elicit were screened and verified manually to avoid duplication and bias. Data extraction was supported by a large language model used solely for organising and summarising information. All extracted data were manually reviewed and verified by the research team through a double-check procedure to minimise bias and ensure accuracy. Methodological quality was assessed using design-appropriate appraisal tools, including the CASP checklist for qualitative studies, the MMAT (2018) for quantitative and mixed-methods studies, and the JBI critical appraisal checklists for systematic reviews and meta-analyses. Based on these assessments, studies were classified as high, moderate or low quality and quality ratings were taken into account during synthesis and interpretation.

2.2 Table of Search Strategy and Databases

As shown in Table 1, five major databases were searched for studies published between 2020 and 2025, yielding a total of 825 records, with Scopus contributing the highest number of retrieved articles.

Table 1
Search strategy and databases

Database	Search Period	Records Retrieved
Scopus	2020–2025	320
Web of Science	2020–2025	210
ERIC	2020–2025	95
IEEE Xplore	2020–2025	80
Google Scholar	2020–2025	120

2.3 Screening and Eligibility

The retrieved studies were screened based on their abstracts using predefined inclusion criteria. Eligible studies focused on undergraduate or postgraduate students in higher education institutions and evaluated AI-powered educational chatbots designed for learning, tutoring or academic support. Studies were required to report learning-related outcomes, such as academic performance, knowledge retention, engagement, satisfaction or skill development and to employ an appropriate research design, including empirical studies, systematic reviews or meta-analyses. Only peer-reviewed academic publications were included. Chatbots limited to rule-based systems or non-educational applications were excluded. Screening decisions were made holistically to ensure relevance and methodological rigor which is the review process followed the PRISMA 2020.

2.4 Included Studies

The PRISMA flow chart illustrates the systematic and transparent process used to identify, screen, assess, and include studies in this review. Following the PRISMA 2020 guidelines, 825 records were identified. After removing duplicates and screening titles and abstracts, 20 studies met the inclusion criteria and were included in this review, including Scopus (320 records), Web of Science (210 records), ERIC (95 records), IEEE Xplore (80 records) and Google Scholar (120 records). No records were identified from study registers. Before the screening process began, 162 duplicate records were

removed while no records were excluded due to automation tools or other reasons. This resulted in 663 unique records being retained for further screening. In the screening phase, the titles and abstracts of the 663 records were carefully examined to determine their relevance to the research objectives. As a result of this initial screening, 560 records were excluded because they did not meet the inclusion criteria or were not directly related to the focus of the study. Consequently, 103 records were considered potentially relevant and were selected for full-text retrieval. All 103 reports were successfully retrieved and none were excluded due to retrieval issues.

The eligibility assessment involved a thorough evaluation of the full-text articles against predefined inclusion and exclusion criteria. Of the 103 reports assessed for eligibility, 83 were excluded for specific reasons. These included studies that were not conducted within a higher education context (32 studies), studies that did not focus on AI chatbot applications (28 studies) and studies that were purely conceptual or review-based without empirical data (23 studies). This rigorous evaluation ensured that only studies closely aligned with the aims of the review were retained. Finally, following eligibility assessment, a total of 20 studies focusing exclusively on higher education contexts met all inclusion criteria and were included in this systematic review. No additional reports of included studies were identified. Overall, the PRISMA flow chart demonstrates a comprehensive and methodical study selection process, ensuring transparency, rigor and reliability in the identification and inclusion of relevant literature for the review.

Overall, the included studies provide a robust and diverse body of evidence demonstrating the potential of AI chatbots as supplementary learning tools in higher education. The findings highlight both the strengths of these systems, particularly in structured knowledge acquisition, language practice and learner confidence, as well as their limitations, including limited effects on critical thinking, higher-order cognitive skills and collaborative learning.

The overall study selection process, including identification, screening, eligibility and inclusion of studies, is illustrated in the PRISMA 2020 flow diagram (Figure 1).

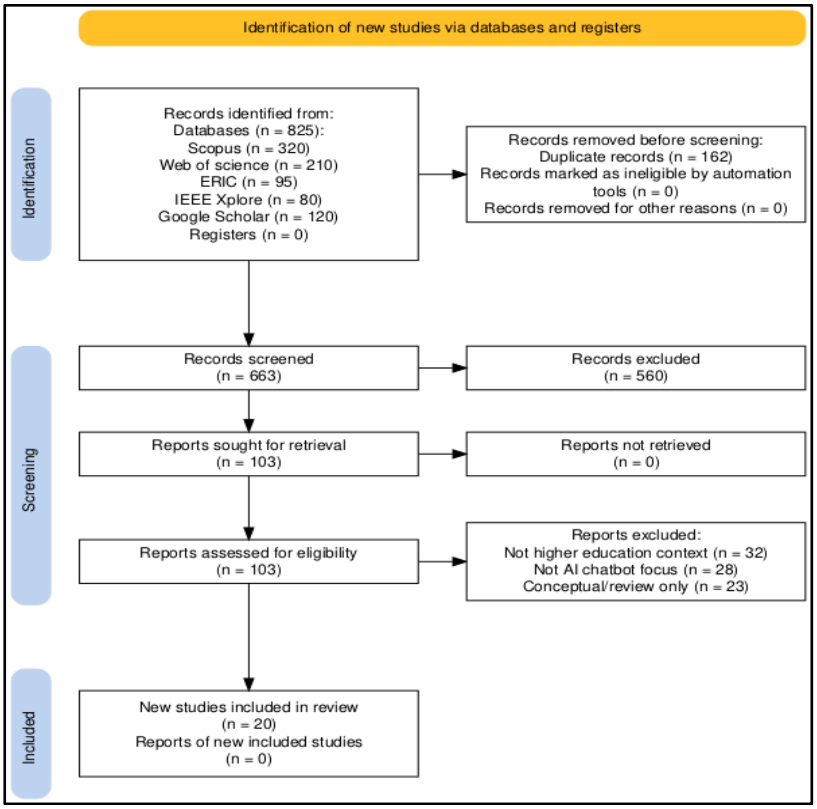

Fig. 1. PRISMA flowchart of literature screening and selection

2.5 Data Extraction

A large language model was used solely to support data organisation and summarisation. All extracted data were manually reviewed and verified by the research team to ensure accuracy and minimise bias from the included studies. The LLM was used as a supporting tool rather than a replacement for researcher judgement, and all extracted data were manually reviewed and verified by the research team to ensure accuracy and consistency. To enhance reliability, a double-check procedure was implemented, whereby extracted information was independently reviewed by at least one researcher and cross-checked against the original sources. Any discrepancies were resolved through discussion and consensus.

2.6 Quality Assessment

The methodological quality of the included studies was assessed to ensure the credibility and trustworthiness of the evidence synthesised in this systematic review. Given the heterogeneity of study designs, a design-appropriate appraisal strategy was adopted. The Critical Appraisal Skills Programme (CASP) checklist was used to evaluate qualitative studies, the Mixed Methods Appraisal Tool (MMAT, 2018) was applied to quantitative and mixed-methods empirical studies and the Joanna Briggs Institute (JBI) critical appraisal checklists were employed for systematic reviews and meta-analyses. Each study was assessed against relevant criteria, including clarity of research aims, appropriateness of research design, adequacy of sampling and data collection procedures, rigour of data analysis and transparency of reporting. Based on the appraisal outcomes, studies were classified into three categories which are high, moderate or low methodological quality.

Of the 20 studies included in this review, 12 studies were rated as high quality, demonstrating robust methodological designs and comprehensive reporting. Six studies were classified as moderate quality, commonly due to limited methodological detail or partial reporting of validity and reliability considerations. The remaining two studies were assessed as low quality, primarily because of insufficient methodological transparency and potential risk of bias. Study quality was taken into account during data synthesis and interpretation. Findings from high-quality studies were given greater interpretive weight when drawing conclusions regarding learning outcomes and implementation effectiveness. Evidence from moderate-quality studies was used to support or contextualise emerging patterns while results from low-quality studies were interpreted with caution and did not substantially influence the main conclusions of this review. A summary of the methodological quality assessment of the included studies is presented in Table 2.

Table 2
Summary of quality assessment of included studies

Study Type	Appraisal Tool Used	Number of Studies	Quality Rating
Systematic reviews & meta-analyses	JBI Critical Appraisal Checklist	10	High (7), Moderate (3)
Quantitative empirical studies	MMAT (2018)	4	High (3), Moderate (1)
Qualitative studies	CASP Qualitative Checklist	3	High (1), Moderate (1), Low (1)

Mixed-methods studies	MMAT (2018)	3	High (1), Moderate (1), Low (1)
Total		**20**	**High (12), Moderate (6), Low (2)**

2.6.1 Table of summary of quality assessment of included studies

The methodological quality of the included studies informed the interpretation of findings throughout the synthesis process. Evidence from high-quality studies was prioritised when identifying key learning outcomes and implementation patterns. Findings from moderate-quality studies were used to triangulate and contextualise results, whereas conclusions from low-quality studies were treated cautiously and were not used as the primary basis for overarching claims.

3. Results

3.1 Characteristics of Included Studies

The systematic review synthesized findings from 20 sources examining AI chatbots in university education. Eight sources provided full text access, while two were abstract-only. The studies varied in their research designs, educational contexts and geographic locations. The characteristics of the included studies are summarised in Table 3.

Table 3
Characteristics of included studies

No	Study	Full text retrieved?	Study Type	Subject Area	Education Level	Geographic Location	Sample Size
1.	Nosheen Malik et al., 2025	Yes	Quantitative	Computer Science	Postgraduate	Pakistan	540 students
2.	O. Chamorro-Atalaya et al., 2023	Yes	Systematic review	General education, second language	Undergraduate and postgraduate	Not specified	Not specified
3.	Xinjie Deng et al., 2023	Yes	Systematic review + Meta-analysis	Nursing, military science, second language	University level	Not specified	2,201 participants across 32 studies
4.	Jinming Du et al., 2024	No	Systematic review	English as a foreign language speaking	Not specified	Not specified	24 studies reviewed
5.	Rizal Bagus Syaifulloh et al., 2025	Yes	Systematic review	English as a foreign language	Not specified	Not specified	43 studies reviewed
6.	Javed Iqbal et al., 2024	Yes	Qualitative study	English language learning	Higher education	Pakistan	100 students

7.	Chunpeng Zhai et al., 2023	No	Systematic review	English as a foreign language	University students	Not specified	28 studies reviewed
8.	MA, W. et al., 2024	Yes	Systematic review	English as a second/foreign language	Undergraduate	Not specified	14 studies reviewed
9.	Chunpeng Zhai et al., 2022	Yes	Systematic review	Second language acquisition	Not specified	Not specified	Studies from 2012-2022
10.	Yazid Albadarin et al., 2024	Yes	Systematic review	English or language education	Higher education	Diverse geographic regions	14 studies reviewed
11.	Okonkwo & Ade-Ibijola (2021)	Yes	Systematic review	Chatbots in education	Primary–Higher education	Global	47 studies
12.	Kuhail et al. (2022)	Yes	Systematic review	Educational chatbots	Higher education	Global	71 studies
13.	Ahmad et al. (2023)	Yes	Systematic review	Educational institutions	Higher education	Malaysia	20 studies
14.	Haristiani, N., 2019	Yes	Qualitative / Inquiry	Language learning	Higher education	Not specified	50 students
15.	Husaeni, D. F. A., Haristiani, N., Wahyudin, W., & Rasim, R. (2024)	Yes	Literature Review & Bibliometric Analysis	Science & Engineering	Higher Education	Not specified	Not specified
16.	Bieletzke (2023)	Yes	Conceptual study	Campus management systems	Higher education	Europe	Not specified
17.	Labadze et al. (2024)	Yes	Systematic literature review	AI chatbots in education	Higher education	Global	62 studies
18.	McGrath et al. (2024)	Yes	Narrative review	Generative AI in higher education	Higher education	Global	Emerging field
19.	Pereira, D. G. et al., 2023	Yes	Scoping review	University systems & learning	Higher education	India / Global	System evaluation
20.	Qiu, W., Su, C. L., Jamil, N. B., Maung Thway, Soo, S., Zhang, L., Lim, F. S., & Lai, J. W. (2025)	Yes	Systematic Review	Education / Technology	Higher Education	Not specified	Not specified

The AI systems examined varied across studies. ChatGPT was the primary focus in several studies while others examined custom chatbots, conversational programs using Dialogflow and speech-recognition chatbots. Key technical features included natural language processing, personalization capabilities, instant feedback mechanisms and emotional support functions.

Implementation approaches varied considerably across studies. Some studies provided structured training and guidelines while others did not specify training protocols. Teacher involvement ranged from monitoring student progress to creating lesson plans and generating quizzes. The duration of chatbot use also varied with some studies conducting interventions for more than one month while others used chatbots for varying periods including less than one week to more than ten weeks.

4. Effects of AI Chatbots on Learning Outcomes

4.1 Academic Performance and Skill Development

Quantitative evidence indicates medium-to-high overall effects of AI chatbots on learning outcomes in higher education. Meta-analytic and systematic review findings demonstrate a substantial positive impact of chatbot use on student learning and academic achievement [11]. Reported gains include improvements in academic performance and enhanced engagement with learning activities, particularly when chatbots are integrated into structured instructional contexts [4].

Learning outcomes varied across cognitive domains. Explicit knowledge acquisition and reasoning-related tasks showed stronger effects compared to higher-order outcomes such as critical thinking [6]. Studies focusing on sustained chatbot use reported improved knowledge retention and recall among university students, suggesting that frequency and duration of use play an important role in learning effectiveness [5].

Language skill development emerged as one of the most consistently improved areas. Prior reviews reported notable gains in vocabulary acquisition, grammatical accuracy and writing fluency among university learners using AI chatbots [13]. Improvements were also observed in receptive skills, including reading and listening comprehension, particularly in tertiary language learning contexts [20]. Speaking-related benefits included reduced communication apprehension and improved pronunciation, especially in chatbots supporting English as a foreign language practice [7].

Despite these positive outcomes, evidence for learning engagement, motivation and critical thinking remains mixed. Several reviews reported non-significant or weak effects for these higher-order outcomes, suggesting that while chatbots are effective for structured knowledge transmission and practice, they may be less capable of fostering deep cognitive engagement without complementary pedagogical support [15].

4.2 Psychological and Affective Outcomes

Overall, the psychological and affective impacts of AI chatbots were predominantly positive. Increased learner confidence was consistently reported, with students expressing greater assurance in their understanding of course content after interacting with chatbots [5]. In language learning contexts, reduced anxiety, particularly speaking related anxiety, was frequently highlighted as chatbots provided a low pressure, non-judgmental environment for practice [7].

Findings related to student engagement were less consistent. While some studies documented higher levels of engagement and interest in learning activities supported by chatbots [14], broader reviews reported no statistically significant overall improvement in engagement across studies [20].

This variation may be attributed to differences in instructional design, chatbot functionality and the extent of teacher involvement.

In addition, positive institutional outcomes were reported. Student satisfaction levels increased, and improved interaction between students and instructors was observed in chatbot-supported learning environments [4]. Evidence also suggests a reduction in dropout rates and improvements in perceived learning efficiency, indicating the potential of AI chatbots to support student retention when appropriately implemented [2].

4.3 Implementation Effectiveness Factors

Several factors were found to moderate the effectiveness of AI chatbots in higher education. Learner characteristics were particularly influential, with students possessing higher levels of digital literacy and technological familiarity reporting greater usability and satisfaction (MA et al., 2024). Regular and sustained chatbot use was also associated with stronger learning outcomes, including improved comprehension and retention [5].

Technical features played a crucial role in successful implementation. Reviews identified usability, responsiveness and contextual relevance as key design elements contributing to positive learning experiences [11]. Personalisation and the provision of immediate, corrective feedback further enhanced learning effectiveness and student interest [6].

Contextual and pedagogical factors were equally important. Blended learning approaches that combined chatbot use with instructor guidance consistently yielded better outcomes than purely AI-driven implementations [15]. Postgraduate students, in particular, were found to benefit from chatbot support for complex academic tasks, while institutional support and a conducive learning environment were identified as essential conditions for effective integration [14].

5. Challenges and Limitations

5.1 Technical and Pedagogical Concerns

Technical limitations posed several challenges in the implementation of AI chatbots in higher education. Issues with accuracy, including “hallucinations” where chatbots generate plausible but incorrect information, raised concerns about reliability [2]. Additionally, programming and system maintenance challenges, as well as performance inconsistencies, affected overall implementation quality [6]. Usability and customization also required careful attention to improve accessibility for diverse learners [14].

Pedagogical concerns centered on potential over-reliance on AI tools. Studies reported risks of reduced creativity and critical thinking skills [10]. The potential for cheating and plagiarism threatened academic integrity [14]. Some studies (21.4%) found negative impacts on innovative capacities and collaborative learning competencies [2]. Concerns about students outsourcing evaluation work to ChatGPT raised questions about authentic learning [10].

AI chatbots’ inability to generate deep insights or analysis [10] and their suitability primarily for beginner-level learners [19] limited their applicability across all learning contexts. The development and implementation of AI dialogue systems remained in their infancy stage, suggesting considerable room for improvement [20].

5.2 Ethical and Access Issues

Ethical concerns emerged as critical challenges. Data privacy issues, algorithmic bias and the need for compliance with educational data protection regulations required careful attention [18]. Privacy and data security concerns necessitated that developers protect personal data [19]. AI chatbots' tendency toward prejudice and lack of cultural sensitivity raised equity concerns [19].

Access and equity issues particularly affected learners in rural or low-income areas who may lack necessary infrastructure or digital literacy [18]. Potential for unequal access due to varying levels of technical proficiency created concerns about widening educational disparities. Resource constraints for IT skills development [14] and lack of training [2] hindered effective implementation.

Implementation barriers included the novelty effect, where improvements in learning outcomes might be due to technology newness rather than inherent effectiveness [6]. The short-term nature of many studies and lack of detailed exploration of ethical issues [18] limited understanding of long-term impacts. Limited literature on considering learners' cultural backgrounds for feedback and the need for better handling of learner frustration and anxiety [20]. highlighted implementation gaps.

5.3 Limitations

Although a large language model was used as a supporting tool during the data extraction process, the possibility of AI-assisted misinterpretation, omission, or bias cannot be entirely eliminated. To mitigate this risk, all extracted information was manually reviewed and verified by the research team through a double-check procedure against the original sources. Any discrepancies identified during this process were resolved through discussion and consensus. Future systematic reviews may benefit from the development of more standardised and transparent human–AI collaboration protocols to further enhance methodological rigor, transparency, and replicability. While a substantial proportion of included studies comprised systematic reviews and meta-analyses, their inclusion enabled cross-validation of findings across contexts. The primary contribution of this review lies not in reproducing existing summaries but in the integrative synthesis of learning outcomes, psychological mechanisms and implementation conditions.

6. Discussion

6.1 Synthesis

Synthesis across the included studies reveals an implicit conceptual relationship underlying AI chatbot effectiveness in higher education. Chatbot features such as instant feedback, natural language interaction, and personalisation operate through mediating psychological and pedagogical mechanisms, including reduced learning anxiety, increased learner confidence, and enhanced opportunities for practice. These mechanisms subsequently influence learning outcomes such as academic performance, retention, and skill development. Importantly, the strength of these relationships is moderated by implementation conditions, including the use of blended learning approaches, structured training, and sustained duration of chatbot use. The proposed conceptual framework illustrating the relationships between AI chatbot features, mediating mechanisms, learning outcomes and implementation factors is presented in Figure 2.

6.2 Conceptual Framework of AI Chatbot Effectiveness in Higher Education

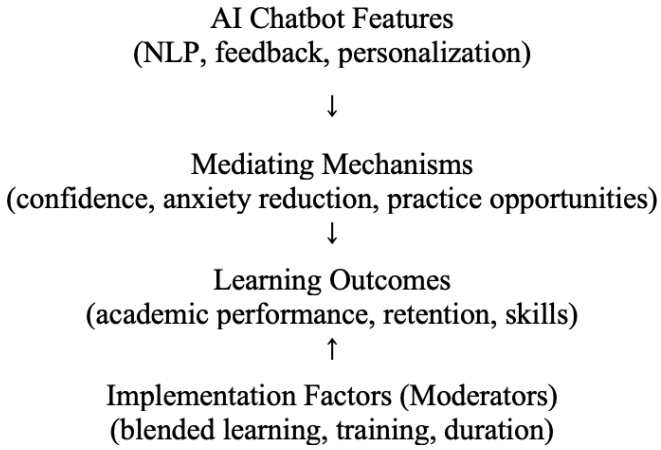

Fig. 2. Conceptual framework of AI chatbot effectiveness in higher education

6.3 Cognitive and Learning Outcomes

AI chatbots demonstrate differential effectiveness across learning dimensions. Studies consistently show strong effects on explicit knowledge transmission tasks, vocabulary acquisition, grammar mastery [13], knowledge retention, and learning achievement [4]. These benefits align with chatbots' core capabilities: providing immediate feedback [16], offering personalized practice opportunities [5] and creating low-pressure learning environments [9].

In contrast, higher-order cognitive skills show weaker or non-significant effects. Critical thinking improvements were not statistically significant [17] and concerns emerged about reduced critical thinking capacity [10]. This pattern suggests that chatbots excel at structured, rule-based learning but struggle with open-ended, creative problem-solving. Their strength in providing immediate, correct answers may inadvertently discourage sustained grappling with ambiguity that develops critical thinking.

Divergent findings on engagement [11] provide insight into implementation effects. While some studies report increased engagement, the meta-analysis found no significant overall effect [6]. This heterogeneity likely reflects implementation quality rather than technology limitations. Studies emphasizing structured training [12], teacher support [13], and blended approaches integrating human facilitation [15] report more positive outcomes than those lacking these features [7].

Tech-savvy students with higher digital literacy found chatbots more usable and satisfactory, while concerns about unequal access due to technical proficiency differences suggest that raw technology deployment without adequate support may widen achievement gaps rather than narrow them. Regular use patterns enhanced effectiveness [14], indicating that sporadic chatbot interactions yield limited benefits compared to sustained, integrated use.

6.4 Psychological and Pedagogical Mechanisms

The consistent finding of reduced language anxiety [18] and increased confidence [14] reveals an important psychological mechanism. Chatbots provide non-judgmental practice environments, allowing students to make mistakes without social consequences. This psychological safety particularly benefits speaking practice [8], where fear of judgment often inhibits participation in traditional settings.

However, this mechanism may explain limited effects on collaborative learning competencies. While chatbots reduce anxiety in individual practice, they cannot replicate social negotiation and interpersonal dynamics essential for collaborative skills development. A proportion of studies noted negative impacts on collaborative learning [2], suggesting over-reliance on chatbots may reduce opportunities for peer-to-peer interaction.

The technology's developmental stage explains some heterogeneity. Multiple studies note that AI dialogue systems remain in their infancy, with gaps in debate and problem-solving skill support and lack of cultural, humor, and empathy functions [20]. Technical limitations including accuracy issues [3], programming and maintenance problems [6], and performance issues [19] reflect immature technology rather than fundamental limitations.

Studies using advanced systems with emotional support capabilities, affective strategies [19] and culturally-informed design [20] report more positive outcomes than those using basic question-answer systems. This suggests that as chatbot technology matures, particularly with large language models, effectiveness across more complex learning dimensions may improve.

6.5 Implementation and Maturity of Technology

Evidence converges on specific conditions maximizing chatbot effectiveness. Blended approaches combining AI assistance with human facilitation [18] outperform purely AI-driven or traditional approaches. This aligns with cognitive load theory: chatbots handle routine knowledge transmission, freeing instructors to focus on higher-order thinking, personalized guidance, and complex problem-solving currently beyond AI capabilities.

Optimal implementation includes structured training for students and educators, clear guidelines to mitigate drawbacks, development of critical evaluation skills, and integration of human interaction [2]. Regular, sustained use [14] rather than sporadic deployment maximizes benefits. Structured implementations support outcomes such as 8.29% reduction in student dropout and 70.6% increase in learning retention [4].

For specific populations, postgraduate students find chatbots particularly adaptable to challenging academic tasks [14], while beginner-level learners benefit most [19], suggesting chatbots serve different functions across educational levels. In language learning, chatbots excel at vocabulary acquisition, grammar practice, and spoken language development but require supplementation for pragmatic language use and interpersonal communication skills [18].

6.6 Implications for Practice

The findings of this systematic review offer several important implications for higher education practice. First, AI chatbots should be positioned as supplementary learning tools within blended learning environments rather than as replacements for human instruction. Lecturers can strategically employ chatbots to support routine academic activities such as vocabulary practice, grammar correction, revision exercises and clarification of course content, thereby freeing classroom time for higher-order thinking, discussion and problem-solving activities.

Second, institutions should provide structured guidance and training for both students and educators on the effective and ethical use of AI chatbots. Clear pedagogical guidelines can help mitigate risks related to over-reliance, superficial learning, plagiarism and academic integrity concerns. Integrating reflective tasks that require students to evaluate and verify chatbot-generated responses may further promote critical engagement with learning materials.

Finally, effective implementation requires ongoing instructor involvement and institutional support. Lecturers should monitor chatbot use, align chatbot activities with learning objectives and ensure that chatbot-supported learning is complemented by meaningful human interaction. These findings underscore that pedagogical integration, rather than mere technological adoption is central to maximising the educational value of AI chatbots in university education.

6.7 Future Research Directions

Despite the growing body of research on AI chatbots in higher education, several directions for future research remain critical. First, future studies should adopt longitudinal research designs to examine the long-term effects of chatbot use on learning outcomes, knowledge retention and learner autonomy. As most existing studies focus on short-term interventions, longitudinal evidence is necessary to determine whether observed benefits persist over time or diminish due to novelty effects.

Second, there is a need for more experimental and quasi-experimental research employing rigorous control conditions. Comparative studies examining chatbot-assisted learning, blended learning and traditional instructional approaches across disciplines and educational levels would provide stronger causal evidence regarding the effectiveness of AI chatbots.

Third, future research should prioritise the development and evaluation of ethical governance frameworks for AI chatbot implementation in higher education. This includes issues related to data privacy, algorithmic transparency, bias mitigation and institutional policies that ensure responsible, equitable and pedagogically sound use of AI technologies. Addressing these dimensions is essential for supporting sustainable and ethical integration of AI chatbots in university learning environments.

References

[1] Ahmad, Che Wan Shamsul Bahri CW, Khirulnizam Abd Rahman Syarbaini Ahmad, Syarbaini Ahmad, and Mokmin Basri. "Kajian Penggunaan Chatbot dalam Institusi Pendidikan: A Study of the Use of Chatbots in Educational Institutions." *Malaysian Journal of Information and Communication Technology (MyJICT)* (2023): 145-157.https://doi.org/10.53840/myjict8-2-103

[2] Albadarin, Yazid, Mohammed Saqr, Nicolas Pope, and Markku Tukiainen. "A systematic literature review of empirical research on ChatGPT in education." *Discover Education* 3, no. 1 (2024): 60. https://doi.org/10.2139/ssrn.4562771

[3] Bieletzke, S. "Ai-chatbot-integration in campus-management-systems." In *EDULEARN23 Proceedings*, pp. 3574-3583. IATED, 2023. https://doi.org/10.21125/edulearn.2023.0971

[4] Chamorro-Atalaya, Omar, Madison Huarcaya-Godoy, Víctor Durán-Herrera, Constantino Nieves-Barreto, Raul Suarez-Bazalar, Yreneo Cruz-Telada, Ronald Alarcón-Anco, Hilda Huayhua-Mamani, Ademar Vargas-Diaz, and Denisse Balarezo-Mares. "Application of the chatbot in university education: A systematic review on the acceptance and impact on learning." *International Journal of Learning, Teaching and Educational Research* 22, no. 9 (2023): 156-178.https://doi.org/10.26803/ijlter.22.9.9

[5] Pereira, Daniela SM, Filipe Falcão, Lilian Costa, Brian S. Lunn, José Miguel Pêgo, and Patrício Costa. "Here's to the future: Conversational agents in higher education-a scoping review." *International Journal of Educational Research* 122 (2023): 102233. https://doi.org/10.1016/j.ijer.2023.102233

[6] Deng, Xinjie, and Zhonggen Yu. "A meta-analysis and systematic review of the effect of chatbot technology use in sustainable education." *Sustainability* 15, no. 4 (2023): 2940. https://doi.org/10.3390/su15042940

[7] Du, Jinming, and Ben Kei Daniel. "Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice." *Computers and Education: Artificial Intelligence* 6 (2024): 100230.https://doi.org/10.1016/j.caeai.2024.100230

[8] Haristiani, Nuria. "Artificial Intelligence (AI) chatbot as language learning medium: An inquiry." In *Journal of physics: conference series*, vol. 1387, no. 1, p. 012020. IOP publishing, 2019. https://doi.org/10.1088/1742-6596/1387/1/012020

[9] Al Husaeni, Dwi Fitria, Nuria Haristiani, W. Wahyudin, and R. Rasim. "Chatbot artificial intelligence as educational

tools in science and engineering education: A literature review and bibliometric mapping analysis with its advantages and disadvantages." *ASEAN Journal of Science and Engineering* 4, no. 1 (2024): 93-118. https://doi.org/10.17509/ajse.v4i1.67429

[10] Iqbal, Javed. "A systematic review on the role of ChatGPT on student learning." *Integration of Mankind and Society* 1, no. 2 (2024): 61-66. https://doi.org/10.56770/ims2024123

[11] Kuhail, Mohammad Amin, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. "Interacting with educational chatbots: A systematic review." *Education and Information Technologies* 28, no. 1 (2023): 973-1018.https://doi.org/10.1007/s10639-022-11177-3

[12] Labadze, Lasha, Maya Grigolia, and Lela Machaidze. "Role of AI chatbots in education: systematic literature review." *International journal of Educational Technology in Higher education* 20, no. 1 (2023): 56. https://doi.org/10.1186/s41239-023-00426-1

[13] Ma, Wenjuan, Wenjing Ma, Yongbin Hu, and Xinyu Bi. "The who, why, and how of ai-based chatbots for learning and teaching in higher education: A systematic review." *Education and information technologies* 30, no. 6 (2025): 7781-7805. https://doi.org/10.1007/s10639-024-13128-6

[14] Malik, Nosheen, Asma Kousar, and Khurram Arshad. "Chat-GPT in Education: Learning Outcomes and Facilitating Knowledge Acquisition." *Inverge Journal of Social Sciences* 4, no. 2 (2025): 97-104. https://doi.org/10.63544/ijss.v4i2.131

[15] McGrath, Cormac, Alexandra Farazouli, and Teresa Cerratto-Pargman. "Generative AI chatbots in higher education: A review of an emerging research area." *Higher Education* 89, no. 6 (2025): 1533-1549.https://doi.org/10.1007/s10734-024-01288-w

[16] Okonkwo, Chinedu Wilfred, and Abejide Ade-Ibijola. "Chatbots applications in education: A systematic review." *Computers and Education: Artificial Intelligence* 2 (2021): 100033. https://doi.org/10.1016/j.caeai.2021.100033

[17] Qiu, W., Su, C. L., Jamil, N. B., Maung Thway, Soo, S., Zhang, L., Lim, F. S., & Lai, J. W. (2025). A systematic approach to evaluate the use of chatbots in educational contexts: Learning gains, engagements and perceptions. Computers, 14(7), 270–270. https://doi.org/10.3390/computers14070270

[18] Syaifulloh, Rizal Bagus, and Shierly Novalita Yappi. "AI CHATBOTS IN EFL: SYNTHESIZING COGNITIVE, PSYCHOLOGICAL, AND ETHICAL IMPLICATIONS." (2025). https://doi.org/10.51878/edutech.v5i1.4701

[19] Zhai, Chunpeng, and Santoso Wibowo. "A systematic review on cross-culture, humor and empathy dimensions in conversational chatbots: the case of second language acquisition." *Heliyon* 8, no. 12 (2022). https://doi.org/10.1016/j.heliyon.2022.e12056

[20] Zhai, Chunpeng, and Santoso Wibowo. "A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university." *Computers and Education: Artificial Intelligence* 4 (2023): 100134. https://doi.org/10.1016/j.caeai.2023.100134