Computers & Education 234 (2025) 105323

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts

Tim Debets [a,*], Seyyed Kazem Banihashem [a], Desirée Joosten-Ten Brinke [b], Tanja E.J. Vos [c,d], Gideon Maillette de Buy Wenniger [d], Gino Camp [a]

[a] *Faculty of Educational Science, Open Universiteit, Heerlen, Netherlands, Valkenburgerweg 177, Heerlen, 6419, AT, the Netherlands*
[b] *School of Health Professions Education, Maastricht University, Maastricht, Netherlands, Universiteitssingel 60, Maastricht, 6229, ER, the Netherlands*
[c] *Escuela Técnica Superior de Ingeniería Informática, Universidad Politecnica de Valencia, Valencia, Spain, Camino de Vera, s/n, Valencia, 46022, Spain*
[d] *Faculty of Science, Open Universiteit, Heerlen, Netherlands, Valkenburgerweg 177, Heerlen, 6419, AT, the Netherlands*

ARTICLE INFO

Keywords:
Artificial intelligence
Chatbots in education
Educational bot
Literature review
Conversational agent

ABSTRACT

There is a growing body of literature on the development and use of chatbots in education. However, insights into the objectives, underlying technologies, theories, and criteria for developing and evaluating chatbots in education are lacking. This study presents a systematic review of 71 papers, identified through a thorough search in Web of Science and Scopus in October 2023, to understand, compare, and create a comprehensive overview of these elements. Papers were selected through a step-by-step procedure based on the guidelines of the PRISMA framework. The main results indicate that most chatbots in education are teaching-oriented and have increasingly been developed using chatbot builder platforms. However, many of these chatbots are integrated into educational settings without a solid theoretical foundation. Despite this, they are primarily evaluated based on perceptual factors and, in most cases, prove effective for their intended objectives. This review also identified a few limitations of the included studies, such as heterogeneity, publication bias, and focus on short-term impact. Overall, this review enhances our understanding of chatbots in education and provides valuable insight and guidance for educators, practitioners, and researchers.

1. Introduction

A chatbot is a software program designed to simulate conversations with users via text or voice (Adamopoulou & Moussiades, 2020). For more than fifty years, chatbots have been used in educational contexts. Laurillard (2002) traces their use as conversational agents back to the 1970s. One early example is the Socratic System (Suppes, 1971), which allowed students to interact with a computer by answering questions and receiving customised feedback. In the 1990s, Geometry Tutor (Anderson et al., 1995) provided tailored

[*] Corresponding author. Open Universiteit, Valkenburgerweg 177, Heerlen, 6419, AT, the Netherlands.
E-mail addresses: tim.debets@ou.nl (T. Debets), kazem.banihashem@ou.nl (S.K. Banihashem), desiree.joosten-tenbrinke@maastrichtuniversity.nl (D. Joosten-Ten Brinke), tanja.vos@ou.nl, tvos@dsic.upv.es (T.E.J. Vos), gideon.maillettedebuywenniger@ou.nl (G. Maillette de Buy Wenniger), gino.camp@ou.nl (G. Camp).

https://doi.org/10.1016/j.compedu.2025.105323
Received 23 July 2024; Received in revised form 9 April 2025; Accepted 10 April 2025
Available online 11 April 2025

feedback to students using a rule-based system. Moving into the 2000s, a well-known example is Autotutor (Graesser et al., 2005), which could answer students' questions, receive explanations, and engage in interactive problem-solving across various subjects. More recently, the bot Jill Watson (Goel & Polepeddi, 2018), based on IBM Watson, was developed as a virtual teaching assistant for a graduate course. Recently, the evolution and application of chatbots in education, also known as conversational agents or dialogue systems, have grown explosively due to advances in Artificial Intelligence (AI), supporting both students and teachers (Chiu et al., 2023; Kuhail et al., 2023).

Research on chatbots in education has grown exponentially since 2019. According to Scopus (accessed February 2024), publications on this topic increased from 53 in 2019 to 255 in 2023, underlining their rising importance in education. Potential explanations for this growth include the rise in online education due to the COVID-19 pandemic (e.g., Han et al., 2022), the rapid advancement and adoption of educational technologies to provide personalised learning experiences (Mohamad Noor, 2023), the increasing teacher-student ratio leading to higher teacher workloads (e.g., Chen et al., 2023), and the round-the-clock availability of chatbots enabling students to engage in learning at their convenience.

The advantages of using chatbots in education have become clearer. For example, chatbots can answer frequently asked questions (Wijayawardena et al., 2022) and support students during learning (Okonkwo & Ade-Ibijola, 2021b). As a consequence, research interest in this field has noticeably grown, leading to numerous literature reviews that summarise the rapid expansion of research on educational chatbots. Building on Kuhail et al.'s (2023) study, Table 1 summarises the key areas of focus in existing literature reviews, revealing that research on chatbots in education primarily explores their applications, objectives, design, technology, evaluation methods, and associated challenges.

Several studies investigate the educational contexts and disciplines in which chatbots are employed. Hwang and Chang (2021), Kuhail et al. (2023), Wollny et al. (2021), and Zhang et al. (2023) reviewed chatbot applications across various learning domains, including computer science, language learning, and general education. Huang et al. (2022) specifically examined chatbots used in language learning, while Smutny and Schreiberova (2020) focused on educational chatbots deployed via Facebook Messenger, classifying them by subject area. Winkler and Söllner (2018) analysed the settings in which chatbots are implemented, placing particular emphasis on the field of application. Hwang and Chang (2021), Wollny et al. (2021), and Zhang et al. (2023) found that language learning is the most common area of chatbot use, whereas Kuhail et al. (2023) reported computer science as the most prevalent domain. Smutny and Schreiberova (2020) reported that Facebook Messenger chatbots are mainly used for delivering information, while Winkler and Söllner (2018) did not provide specific quantitative data on chatbot applications.

Another major area of interest concerns the educational objectives of chatbots and their impact on student experiences. Hobert and von Wolff (2019), Kuhail et al. (2023), Pérez et al. (2020), and Wollny et al. (2021) explored the roles of chatbots in different learning settings and formats, such as teaching assistants or peer collaborators, and how these roles shape learning experiences. Winkler and Söllner (2018) investigated their influence on motivation, self-efficacy, and engagement, and their function in supporting metacognitive thinking and feedback provision, noting that their effectiveness is highly context-dependent. Hwang and Chang (2021) examined chatbot-supported learning strategies and their effects on engagement, motivation, and learning outcomes. Lo et al. (2024) focused on the influence of ChatGPT on students' behavioural, emotional, and cognitive engagement. Jeon, Lee, and Choi (2023) assessed the impact of chatbots on language proficiency, motivation, and learner confidence. Zhang et al. (2023) specifically investigated the affective outcomes of chatbot-assisted learning and reported largely positive effects. In addition, Zhang et al. (2023) and Huang et al. (2022) analysed the educational objectives of chatbots, including support for speaking and writing skills in language learning contexts.

Several studies categorise chatbots according to their design and functionality. Martha and Santoso (2019) classified pedagogical agents into text-based, voice-based, 2D character, 3D character, and human-like avatars. Huang et al. (2022) identified different functions of language learning chatbots, describing them as interlocutor, simulation, transmissive, helpline, or recommendation-based. Jeon, Lee, and Choi (2023) categorised chatbots by function, including capacities such as conversational partner, feedback provider, evaluator, resource provider, and interviewer. Kuhail et al. (2023) further explored chatbot types of engagement, such as teaching agent, peer, teachable agent, and motivational agent. They also investigated different interaction styles (chatbot-driven vs. user-driven) alongside design principles like personalisation and social dialogue. Huang et al. (2022), Hobert and

Table 1
The focus of different relevant literature reviews.

Focus	Relevant studies
Field of application	(Huang et al., 2022; Hwang & Chang, 2021; Kuhail et al., 2023; Smutny & Schreiberova, 2020; Winkler & Söllner, 2018; Wollny et al., 2021; Zhang et al., 2023)
Objectives and learning experiences	(Hobert & von Wolff, 2019; Huang et al., 2022; Hwang & Chang, 2021; Jeon, Lee, & Choi, 2023; Lo et al., 2024; Kuhail et al., 2023; Pérez et al., 2020; Winkler & Söllner, 2018; Wollny et al., 2021; Zhang et al., 2023)
Design approaches	(Hobert & von Wolff, 2019; Huang et al., 2022; Jeon, Lee, & Choe, 2023; Jeon, Lee, & Choi, 2023; Kuhail et al., 2023; Lai & Lee, 2024; Martha & Santoso, 2019; Winkler & Söllner, 2018)
Technology used	(Hobert & von Wolff, 2019; Pérez et al., 2020; Smutny & Schreiberova, 2020)
Evaluation methods	(Hobert, 2019; Hobert & von Wolff, 2019; Hwang & Chang, 2021; Jeon, Lee, & Choi, 2023; Kuhail et al., 2023; Lai & Lee, 2024; Pérez et al., 2020)
Challenges	(Du & Daniel, 2024; Huang et al., 2022; Kuhail et al., 2023; Okonkwo & Ade-Ibijola, 2021a; Zhang et al., 2023)

Note. This is an updated table from Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*, 973–1018. https://doi.org/10.1007/s10639-022-11177-3.

von Wolff (2019), Jeon, Lee, and Choi (2023), and Kuhail et al. (2023) also categorised chatbot design according to platform (e.g., web-based or smartphone). Jeon, Lee, and Choe (2023) proposed design categories based on goal orientation, embodiment, and multimodality. Winkler and Söllner (2018) classified chatbots by technological design, distinguishing among flow-based systems, AI-driven chatbots, those with speech recognition, and systems that incorporate contextual data. Lastly, Lai and Lee (2024)analysed research designs concerning conversational AI in English language teaching.

The technological infrastructure behind educational chatbots is another area of attention. Hobert and von Wolff (2019) examined the system architectures of chatbots, while Pérez et al. (2020) discussed the frameworks and natural language processing tools used in chatbot development. Smutny and Schreiberova (2020) focused on chatbot development specifically for the Facebook Messenger platform.

Evaluation methods for educational chatbots also represent a research focus. Hobert (2019) distinguished among evaluation procedures, measurement instruments, and objectives. Hobert and von Wolff (2019) identified two main evaluation approaches, Wizard-of-Oz experiments and field studies, and highlighted the absence of a standardised evaluation framework. Hwang and Chang (2021) reviewed various research designs and found that quantitative methods dominate. Jeon, Lee, and Choi (2023) analysed evaluation approaches and noted a strong preference for mixed-methods research using a single-chatbot group design. Kuhail et al. (2023) reported that experiments and evaluation studies are the most frequently employed methodologies. Similarly, Pérez et al. (2020) observed that chatbot evaluations often rely on learner perception surveys. Lai and Lee (2024) also reviewed data collection strategies, identifying mixed methods as the most common, and examined the types of research design, including quasi-experimental approaches.

Finally, researchers address the challenges and limitations related to chatbot implementation in education. Kuhail et al. (2023) identified both technical and pedagogical challenges. Okonkwo and Ade-Ibijola (2021a), and Zhang et al. (2023) outlined issues such as technical constraints, limited user acceptance, and concerns around scalability. Huang et al. (2022) highlighted not only technical challenges but also the novelty effect and cognitive load as barriers to effective chatbot use. Du and Daniel (2024) additionally discussed the difficulties of implementing AI chatbots, specifically in the context of English-speaking skills development.

While we acknowledge the efforts of prior review studies in addressing different aspects of educational chatbots, several gaps remain unaddressed. First, an in-depth analysis of the underlying technology of chatbots in education is missing. 'Underlying technology' refers to the technology, algorithm, or model on which a chatbot is built. For example, with the underlying technology, we aim to answer questions such as whether deep learning is used or whether the chatbot has been developed using rule-based methods. While Pérez et al. (2020) briefly discuss frameworks and language processors, they do not delve into specific models, algorithms, or approaches. Huang et al. (2022), Hobert and von Wolff (2019), Jeon, Lee, and Choi (2023), and Kuhail et al. (2023) only discuss the platform on which a chatbot was developed (e.g., smartphone, desktop PC, or smart speaker). Understanding the underlying technology can inform educational chatbots' design, implementation, and optimisation, potentially enhancing their effectiveness, addressing scalability challenges, and ensuring meaningful learning experiences. For these reasons, investigating the underlying technology of the chatbots in the literature is included in the current study.

Second, another gap that emerges is the lack of a comprehensive analysis of theories underpinning the integration of chatbots in educational settings. While Zhang et al. (2023) suggest which theoretical frameworks chatbots could be built on, they do not conduct a literature review to see whether chatbots are based on these theoretical frameworks. Furthermore, Hwang and Chang (2021) primarily analysed the learning strategies used by educational chatbots but did not systematically examine the broader educational theories that support these strategies. Our study addresses this gap by not only analysing theories concerning learning strategies but also investigating other relevant educational theories, including Social Learning Theories, Personal Development, and Motivational Theories. Since the use of chatbots should be built on a well-thought-out pedagogical and theoretical foundation, the absence of underlying theories may lead to their ineffective or misguided implementation, limiting their potential benefits in educational settings.

Third, it is unclear based on which criteria chatbots are being evaluated. While several literature reviews (Hobert & von Wolff, 2019; Hwang & Chang, 2021; Jeon, Lee, & Choi, 2023; Kuhail et al., 2023; Lai & Lee, 2024; Pérez et al., 2020) examine methods and analyses used to evaluate chatbots, such as questionnaires, Wizard-of-Oz experiments, and ANCOVA, only Hobert (2019) briefly discusses the specific evaluation objectives measured. A comprehensive overview of the criteria applied across the 25 included research studies in Hobert's research is still missing. For example, if a researcher seeks to evaluate purely using technical methods, such as accuracy and BLEU, which methods are most appropriate, and how can they be integrated? Alternatively, for user-based evaluations, such as psychological factors, what criteria can be measured, how are these assessed, and how can these be combined with technical evaluations? This gap highlights the need for a more comprehensive analysis to show which criteria are measured and how these criteria are evaluated.

Furthermore, even when a study provides a detailed analysis of which criteria are measured and how they are assessed, the effectiveness of chatbots remains unclear without examining the actual results. Do the findings demonstrate that chatbots improve learning outcomes? Which chatbot designs are most effective? Addressing this gap is crucial for providing researchers and practitioners with actionable insights into chatbot effectiveness. Therefore, to develop a thorough analysis, the effectiveness of chatbots should not be overlooked. A holistic overview can help identify effective chatbots and the foundational elements that contribute to their success.

Lastly, no literature review thus far has analysed the interactions among the objectives, technologies, theories, evaluation criteria, and effectiveness of chatbot applications in education. Such an overview could help future research by creating an understanding of the outcome these factors have on the success of chatbot implementations in education. This can lead to more informed pedagogical decision-making, more effective chatbot designs, and improved user satisfaction. Moreover, this overview can also highlight which interactions have been underexplored or overlooked, indicating which areas need further research.

Therefore, this paper aims to address the above-mentioned gaps in the existing literature. By systematically analysing and incorporating the latest research findings, this study contributes to an in-depth analysis of (1) the underlying technologies and methods for developing chatbots in education, (2) the role of theories in the application of chatbots in education, (3) the criteria used to evaluate chatbots in education and how these criteria are evaluated, and (4) the interactions among objective, technology, theory, criteria, and effectiveness. Such a review can potentially add value to guide future research in the promising research area of chatbot use in education. It can also assist educators and practitioners in building more trust in the results of chatbots in education by showing educational underpinnings. Taking this into account, the following research questions have been formulated to guide this study.

RQ 1. What are the objectives of chatbots in education?

RQ 2. Which technologies are utilised when developing chatbots in education?

RQ 3. How do various educational theories inform the integration of chatbots in educational settings?

RQ 4. Which criteria are used to evaluate chatbots in education?

RQ 5. How effective are chatbots in education?

RQ 6. How do the interactions between objectives, technologies, theories, and evaluation criteria influence the overall outcome and success of effective chatbot implementation?

2. Method

Research questions 1–5 will be answered by a systematic literature review. Research question 6 will be answered by using a social network analysis (Wasserman & Faust, 1994) on the findings of RQ1 – RQ5. The chosen method to systematically review the literature was the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) method (Page et al., 2021). Then, a critical quality appraisal strategy was used to refine the remaining studies (Gerritsen-van Leeuwenkamp et al., 2017; Theelen et al., 2019). This strategy helped to include only publications with solid methodology, results, and discussion sections. Finally, the information from each remaining publication in the final selection was systematically and explicitly aggregated, and analysed in a framework based on the research questions (Gerritsen-van Leeuwenkamp et al., 2017). Fig. 1 shows the process of selecting publications for this systematic review and the resulting number of publications.

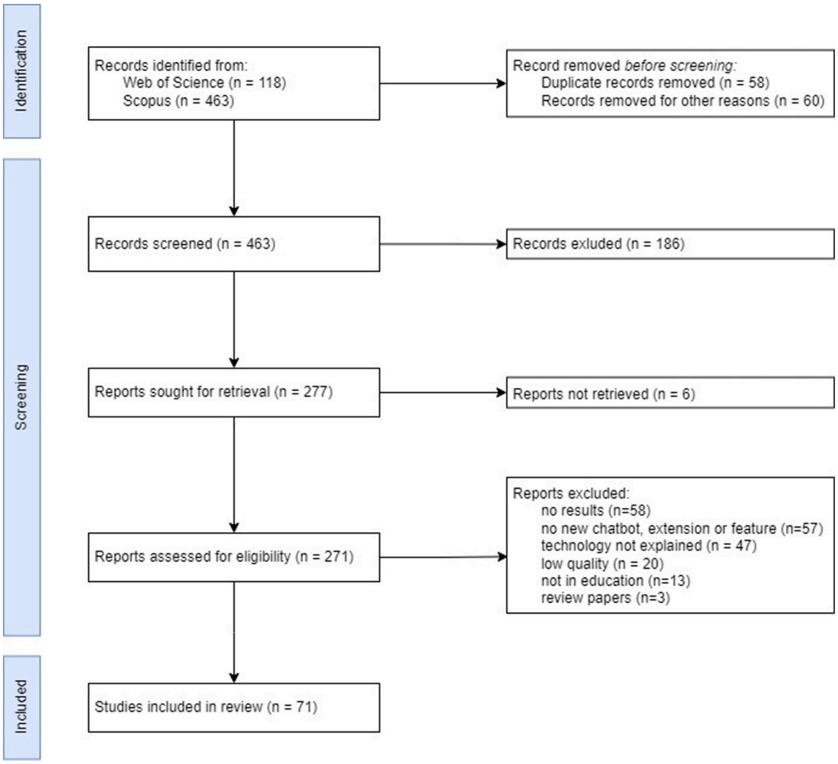
Identification
Records identified from:
Web of Science (n = 118)
Scopus (n = 463)
Record removed before screening:
Duplicate records removed (n = 58)
Records removed for other reasons (n = 60)
Screening
Records screened (n = 463)
Records exluded (n = 186)
Reports sought for retrieval (n = 277)
Reports not retrieved (n = 6)
Reports assessed for eligibility (n = 271)
Reports excluded:
no results (n=58)
no new chatbot, extension or feature (n=57)
technology not explained (n = 47)
low quality (n = 20)
not in education (n=13)
review papers (n=3)
Included
Studies included in review (n = 71)

Fig. 1. Flowchart of the selection process based on PRISMA (Page et al., 2021).

2.1. Search strategy

To search for relevant publications, the two main databases, Web of Science and Scopus, were selected and accessed on October 9, 2023. The relevant terms for the search query were defined based on the key concept of the study, namely, chatbots in education. Relevant synonyms for chatbots and education have been searched for and used to get a more complete search, including suggestions made by the thesaurus of EBSCO. The following keywords were used to search in both databases: ("chatbot" OR "chatbot virtual assistants" OR "chat bots" OR "conversational agents" OR "pedagogical agents" OR "dialogue system" OR "dialogue system" OR "smart personal assistants" OR "smart assistants") AND ("education" or "academic education" OR "higher education" OR "secondary education" OR "primary education").

2.2. Inclusion and exclusion criteria

This study used a two-phase process to apply the inclusion and exclusion criteria. The first phase was implemented during the initial screening process. The primary inclusion and exclusion criteria were: (a) Only publications in the English language were included since the majority of scientific papers are typically published in English. (b) Only publications from 2019 to 2023 were included. This timeframe was selected because GPT-2, a milestone in large language model research, was introduced in 2019 (Radford et al., 2019), likely stimulating advancements in chatbot-related research. (c) Only peer-reviewed empirical studies, articles or conference papers were included, due to their level of originality, reliability, and validity. (d) Review papers were excluded to focus on studies presenting new findings or methodologies. This resulted in 463 papers.

In the second screening phase, additional screening criteria were applied to focus on studies directly relevant to the research objectives. The inclusion and exclusion criteria were: (a) Only publications in the field of education were included, with a focus on chatbots used in tertiary, secondary, and primary educational settings. (b) Studies were required to develop or introduce an educational chatbot and should not be exact duplicates of previous studies. This excludes papers that did not implement or introduce a chatbot, such as those conducting qualitative analyses of user expectations without chatbot deployment. Additionally, studies that were exact duplicates of prior research were excluded to prevent skewing the results. (c) Papers that did not provide details about the underlying technology of the chatbot, such as the development platform, AI model (e.g., deep learning model), or rule-based method, were excluded to ensure technical insights were available. For example, papers that develop a chatbot using a base AI model, such as BERT, were included since they provide the necessary technical insights. (d) Papers that lacked methodological details or empirical research results were excluded to avoid including conceptual papers. This left 91 papers for quality appraisal.

2.3. Quality appraisal

The quality appraisal framework utilised for this study combined suggestions of both Gerritsen-van Leeuwenkamp et al. (2017) and Theelen et al. (2019). The criteria for the quality appraisal originated in Gerritsen-van Leeuwenkamp et al. (2017)(Table 2), while, as suggested by Theelen et al. (2019), each criterion in the quality appraisal framework was assigned a score ranging from missing (0) to

Table 2
Quality and critical appraisal criteria.

Code	Description of code
Problem formulation	The problem formulation gives a clear statement of the objectives and the scope of the study (AERA, 2006; Spencer et al., 2003).
Contribution to knowledge	Research goals, questions or hypotheses are set in the context of existing knowledge and identify new areas for investigation (AERA, 2006; Spencer et al., 2003).
Context of research	Specifics of the context in which the research was conducted are described (AERA, 2006; Spencer et al., 2003).
Appropriateness of research method for goal(s)	The appropriateness of the research methods to meet the research goal(s) is adequately described (AERA, 2006; Spencer et al., 2003).
Data collection/method	The process of data collection is precisely and transparently described and argued (AERA, 2006; Spencer et al., 2003).
Data analysis	The process of data analysis is precisely and transparently described and argued (AERA, 2006; Spencer et al., 2003).
Substantiated result(s)	Results are supported by the required research evidence (AERA, 2006; Spencer et al., 2003).
Logic of result(s)	Results 'make sense'/have a coherent logic (Spencer et al., 2003).
Coherence of result(s) with other knowledge	Results are coherent with other knowledge and experience (Spencer et al., 2003).
Corroborating evidence	Corroborating evidence is used to support or refine results (Spencer et al., 2003).
Presentation of result(s)	Results are presented or conceptualised in a way that offers new insights (Spencer et al., 2003).
Link between result(s) and goal(s)	Results are linked to the research goals (Spencer et al., 2003).
Link between conclusion(s) and goal (s)	The conclusion is linked to the research goals (AERA, 2006; Spencer et al., 2003).
Implications of study	The theoretical, practical, or methodological implications are credible and discussed (AERA, 2006; Spencer et al., 2003).
Limitations of evidence	Limitations of evidence and what remains unknown/unclear or is needed are discussed (Spencer et al., 2003).
Error or bias	Error(s) or (potential) biases in the research are discussed (AERA, 2006; Spencer et al., 2003).

Note. Taken from Gerritsen-van Leeuwenkamp, K. J., Joosten-ten Brinke, D., & Kester, L. (2017). Assessment quality in tertiary education: An integrative literature review. *Studies in Educational Evaluation, 55*, 94–116. https://doi.org/10.1016/j.stueduc.2017.08.001.

extensively mentioned (3), indicating the extent a criterion was mentioned in the paper. As per the recommendations of Theelen et al. (2019), studies that obtained an average score of less than 2 were excluded from our analysis. This led us to exclude 20 papers. The remaining 71 papers were included in the final analysis.

2.4. *Characteristics of included publications*

Appendix A summarises the characteristics of the included publications by author and year, publication source, context (education level and number of participants), country, domain, and research methodology. The reviewed articles were published between 2019 and 2023, with a significant increase in the number of publications every year. This indicates the gaining attention of chatbots in education. The number of unique publication sources is 56, and the three most common sources are the HCI International Conference (n = 4, 6%), Computers and Education (n = 3, 4%), and International Journal of Educational Technology in Higher Education (n = 3, 4%). The majority of research was conducted in higher education (n = 53, 75%), followed by secondary education (n = 13, 18%). The number of participants involved in the reviewed publications varied from 0 to 1063. One article (Gonçalves et al., 2022) only described the number of conversations investigated (n = 8309). The first authors of the included publications came from 36 different countries, with the majority coming from the United States (n = 9, 13%), followed by Malaysia, Saudi Arabia, China, Germany, and Spain (n = 4, 6%). Chatbots in education have been utilised in 31 unique domains, with 49 per cent developed for STEM subjects (n = 35). The majority of these are for Computer Science (n = 11, 15%), followed by Healthcare (n = 7, 10%). After STEM subjects, the most common chatbots are developed for multi-domain applications (n = 5, 7%) and language teaching chatbots (n = 4, 6%). The majority of included publications used a quantitative methodology to conduct their research (n = 39, 55%), followed by a mixed-method approach (n = 30, 42%). Only 2 papers (Cai et al., 2021; Draxler et al., 2022) used a qualitative methodology approach.

2.5. *Analysis strategy*

A coding scheme was developed based on the framework of Gerritsen-van Leeuwenkamp et al. (2017) to systematically and explicitly aggregate all the relevant information based on the research questions (Table 3). Using this coding scheme, all the relevant information was first coded in text and later analysed in Microsoft Excel. Finally, the inter-rater reliability between two coders was examined for the screening and coding process, similar to the approach used in Gao et al. (2024). First, during the screening process, 35 articles were randomly selected to be screened by the two coders. Cohen's kappa coefficient analysis was used to measure this inter-rater reliability between the coders, and the results showed that there is a high degree of consistency between the coders (Kappa = 0.76, $p < 0.001$) during screening. Finally, the inter-rater reliability between two coders was also examined during coding by randomly selecting ten per cent (n = 7) of the papers. Cohen's kappa coefficient analysis showed that there was a high level of agreement between the two coders (Kappa = 0.78, $p < 0.001$). When the coders did not agree, a short discussion took place, the framework was adjusted accordingly, and a consensus was reached. After that, the first coder screened and coded the remaining publications. To perform a social network analysis, the results of RQ1- RQ5 were binary coded in Excel, and the analysis was conducted on this binary dataset using Python. The Excel file was imported using the pandas library for analysis. The co-occurrence of categories across research questions was analysed by calculating frequencies using standard Python code. Then, a network graph was constructed with nodes representing categories and edges indicating co-occurrence frequencies. Finally, a visual representation of this network was created using draw.io to highlight these relationships.

3. Results

Table 4 includes a full list of all reviewed publications, showing how each research question was addressed through categories and sub-categories; a more in-depth analysis of these results is provided below.

Table 3
Information related to the research questions.

Criterion	Description of criterion
Objective of the chatbot	The objective of the chatbot
Primary Function	The primary function of the chatbot: teaching-oriented, service-oriented or both
Technique of the chatbot in general	Rule-based, AI, hybrid, development/chatbot builder platform (e.g. Dialogflow or Rasa) or other. In the case of others, it is explained what is used
Which rule-based approach is used	An overview of which rule-based approach is used for this chatbot
Which AI algorithm/model is used	An overview of which AI algorithms or models have been used
Which development package is used	The software package/development platform/chatbot builder platform is used to develop the chatbot
Theory	Which theory is mentioned to inform the introduction of the chatbot
Evaluation criteria	Which evaluation criteria is used to measure the effectiveness of the chatbot
Evaluation tool	Which tool is used to measure the evaluation criteria, i.e. questionnaire, automatic evaluation, etc.
Effectiveness of the chatbot	The effectiveness of the chatbot based on the found criteria

Table 4
Analysis of the reviewed publications.

Reference	Objective		Chatbot Technology & Category		Theory	Evaluation criteria	Tool	Effectiveness
	Chatbot objective	Primary Function	Technology	Category				
Abdelhamid and Katz (2020)	Support student's learning processes and engagement	Teaching	Dialogflow	Platform	–	Usability	Questionnaire	Positive Usability (77 out of 100 SUS score)
Adair et al. (2023)	Assess and support 8th-graders in conducting investigations and developing mathematical models to describe scientific phenomena	Teaching	–	Rule-based	–	Competencies	Pre- & Post-test	Significant increase in Competencies
Al-Abdullatif et al. (2023)	Support students' motivation and learning strategies, and help set learning goals	Teaching	Freshdesk	Platform	–	Motivation & Learning strategies use	Questionnaire	Significant difference in Motivation & Learning Strategies
Ali et al. (2022)	Provide information support on the admission process	Service	Dialogflow	Platform	–	Usability, Accuracy, Response time	Questionnaire, Automatic	Positive Usability (72 out of 100 SUS score), 76.8% Accuracy, 216 ms Average response time
Aljameel et al. (2019)	Enhance learning for children with Autism Spectrum Disorder by adapting to the Visual, Auditory and Kinaesthetic learning styles	Teaching	String similarity, Pattern matching	Rule-based	–	Experience, Knowledge, Accuracy, Chatbot behaviour	Questionnaire, Pre- & Post-test, Chatlogs	Positive Experience, Significant difference in Knowledge, 89% Accuracy
Aloqayli and Abdelhafez (2023)	Answer frequently asked questions about university admission	Service	Bostify	Platform	–	Experience, Accuracy, Precision, Recall, F1-score	Questionnaire, Automatic	Positive Experience (76.6 out of 100 CUQ score), 92% Accuracy, 0.97 Precision, 0.95 Recall, 0.95 F1-score
Alqahtani and Alrwais (2023)	Answer frequently asked questions about the Blackboard system	Service	Rasa	Platform	–	Accuracy, F1-score, Precision	Automatic	93% Accuracy, 0.93 F1-score, 0.93 Precision
Aujogue and Aussem (2019)	Inform prospective students about a Data Science master's program	Service	Hierarchical recurrent neural network	AI	–	Accuracy	Automatic	99% Accuracy
Cai et al. (2021)	Reduce the learning time by supporting students in learning math by explaining concepts, providing practice questions and offering tailored feedback by using personalised pedagogical policies	Teaching	Finite state machine, Contextual bandit algorithm, Uniform random data, Rule-based	Hybrid	Educational Theories, Learning Theories	Chatbot behaviour, Learning gain, Experience	Questionnaire, Pre- & Post-test, Chatbot use	Positive personalisation of chatbot, No significant difference in Learning gain, Indifference between chatbot and another platform
Chae et al. (2023)	Help students to learn English	Teaching	BART-large model	AI	–	Accuracy	Automatic	88% Accuracy
Chen et al. (2021)	Inquire about different aspects of students' information and automatically infer their need deficiencies	Service	LSTM, DQN, Template	Hybrid	Educational Theories	Identifying need deficiencies & number of turns	Automatic	44% Correctly identifying Need Deficiencies in 12 turns
Chen et al. (2023)	1) Interview students to gather their perceptions of their learning needs and explore	Both	Juji inc	Platform	–	Experience	Chatbot use, Questionnaire	Positive Experience

(continued on next page)

Table 4 (*continued*)

Reference	Objective		Chatbot Technology & Category		Theory	Evaluation criteria	Tool	Effectiveness
	Chatbot objective	Primary Function	Technology	Category				
	ideas on how to utilise chatbots in student success 2) Teach AI concepts							
Deveci Topal et al. (2021)	Help students in understanding the "Matter and the changing state of matter" unit	Teaching	Dialogflow	Platform	–	Experience, Knowledge	Focus group, Test	Positive Experience, No significant difference in Knowledge
Draxler et al. (2022)	Evaluate students' well-being and experiences in online education	Service	Telegram	Platform	–	Usability	Chatlogs	Using the chatbot for experience sampling had a positive impact on the experience sampling procedure
Durall and Kapros (2020)	Assess and enhance users' core competencies by facilitating self-regulated learning through assessments	Teaching	Conversational Form Library	Rule-based	–	Experience, Confidence in ability	Questionnaire	Positive feedback, but online forms are preferred over chatbot, Increase in confidence in Ability
El Hefny et al. (2021)	Provide students with information about the content, dates, timings, locations, and deadlines for their course	Service	Dialogflow	Platform	–	Usability	Questionnaire	Positive usability (84 out of 100 SUS score)
Essel et al. (2022)	Answer multimedia programming questions	Teaching	Flow.xo	Platform	–	Experience, Learning gain	Focus group, Pre- & Post-test	Positive Experience, Significant difference in Learning gain
Fidan and Gencel (2022)	Give elaborate and personalised feedback when a student answers a question	Teaching	Microsoft power virtual agents	Platform	–	Knowledge, Intrinsic motivation, Perception	Pre- & Post-test, Questionnaire	Positive Perceptions, Significant difference in Learning performance & Motivation
Giler et al. (2023)	Support the student by providing information related to the Computer Science faculty	Service	Telegram	Platform	–	Acceptance & Satisfaction	Questionnaire	High level of Acceptance and Satisfaction
Gonçalves et al. (2022)	Answer questions about the higher education institution	Service	IBM Watson	Platform	–	Chatbot use	Chatbot analytics	Demands for student services were met at a lower cost
González et al. (2022)	Recommend which courses a student should follow based on real-life examples and the current status of projects	Service	Amazon services, Dialogflow	Platform	–	Experience	Questionnaire, Interviews	Improved recommendations, Improved Learning experience
Han et al. (2022)	Promote nursing skills related to electronic fetal monitoring	Teaching	Landbot.io	Platform	–	Experience, Knowledge, Confidence, Interest in education, Self-directed learning	Questionnaire, Test	No significant difference in Knowledge, Confidence and Experience, Significant difference in Interest in education and Self-directed learning
Hew et al. (2023)	Help students set personal learning goals based on the SMART framework and help students learn a foreign language by acting as a learning buddy	Teaching	Dialogflow	Platform	Personal Development and Motivational Theories, Social Learning Theories	Perception, Engagement	Questionnaire, Interviews, Chatlogs	Positive Perception, On average, 8 utterance turns

(*continued on next page*)

Table 4 (*continued*)

Reference	Objective		Chatbot Technology & Category		Theory	Evaluation criteria	Tool	Effectiveness
	Chatbot objective	Primary Function	Technology	Category				
Hirose et al. (2021)	Advice students on their learning strategies	Teaching	ID3 decision tree algorithm	AI	Personal Development and Motivational Theories	Experience, Self-efficacy, Learning strategy use	Questionnaire	Positive Experience, Significant difference in Self-efficacy & Learning strategy use
Hsu et al. (2023)	Facilitate individual students' knowledge construction by guiding them to focus on their learning goals throughout the learning process	Teaching	IBM Watson	Platform	Educational Theories	Knowledge, Learning anxiety and enjoyment	Test, Questionnaire	Significant improvement in Knowledge and Learning enjoyment, Significant decrease in Learning Anxiety
Iku-Silan et al. (2023)	Guide students in a decision-making manner to conduct learning, and provide feedback and suggestions after solving learners' questions	Teaching	IBM Watson	Platform	Learning Theories	Experience, Learning gain, Self-efficacy, Engagement, Extrinsic motivation, Mental efforts	Questionnaire, Pre- & Post-test	Significant difference in Learning gain, Extrinsic motivation, Self-efficacy, Engagement and satisfaction, and Mental efforts
Jariwala et al. (2021)	Help visually impaired students learn mathematical concepts	Teaching	Neural network, Expression tree	AI	–	Accuracy	Automatic	99% Accuracy
Khalil and Rambech (2022)	Provide functionality for acquiring lecture notes and course schedules, completing of course related quizzes and contacting course professors through a conversational messaging interface	Service	Telegram	Platform	–	Usability	Questionnaire, Interviews	Positive Usability
Khin and Soe (2020)	Answer frequently asked questions about university-related information	Service	Recurrent neural network encoder + decoder	AI	–	BLEU	Automatic	0.41 BLEU
Kohnke (2023)	Assist second language learners by chatting with them and guiding learners to appropriate learning resources	Both	Dialogflow	Platform	–	Experience	Questionnaire, Interviews	Positive Experience
Kumar (2021)	Facilitate group work collaboration by registering group members, sharing information, monitoring progress and sharing peer-to-peer feedback	Service	Textit	Platform	–	Learning gain, Self-efficacy, Need for cognition, Perceived motivation, Perception of learning, Teamwork	Pre- & Post-test, Questionnaire	Significant difference in Learning gain and Teamwork, No significant difference in the Need for cognition, Perceived motivation, Self-efficacy, and Perception of Learning
Lam et al. (2023)	Answer factual questions about the university and engage students in general conversation	Service	Transformer, kNN algorithm	AI	–	BLEU	Automatic	0.88 BLEU
Lee and Yeo (2022)	Support preservice teachers' development of responsive teaching practices in mathematics	Teaching	IBM Watson	Platform	Educational Theories	Experience, Accuracy, Chatbot behaviour	Questionnaire, Chatlogs, Chatbot analytics	Positive Experience, 84% Accuracy
Leonardi and Torchiano (2023)	Answer student questions on object-oriented programming	Teaching	Rasa, Matching algorithm	Hybrid	–	Accuracy, F1-score, Precision, Recall	Automatic	93% Accuracy, 0.92 F1-score,

(*continued on next page*)

Table 4 (*continued*)

Reference	Objective		Chatbot Technology & Category		Theory	Evaluation criteria	Tool	Effectiveness
	Chatbot objective	Primary Function	Technology	Category				
								0.91 Precision, 0.93 Recall
Liaw et al. (2023a)	Train nursing students in sepsis care and interprofessional communication	Teaching	Dialogflow	Platform	–	Knowledge, Self-efficacy, Inter-professional communication	Pre- & Post-test, Questionnaire	Significant difference in Sepsis care knowledge, No significant difference in Interprofessional communication knowledge and Self-efficacy
Liaw et al. (2023b)	Train nursing students in interprofessional communication	Teaching	Dialogflow	Platform	–	Acceptance, Knowledge, Self-efficacy	Questionnaire, Focus group, Pre- & Post-test	Positive Acceptance, Significant improvement in Knowledge and Self-efficacy
Lin and Ye (2023)	Support students in conducting biology learning activities outside the classroom	Teaching	–	Rule-based	Educational Theories	Learning gain	Pre- & Post-test	Significant difference in Learning gain
Mageira et al. (2022)	Teach students cultural content in a foreign language	Teaching	Snatchbot	Platform	–	Experience, Learning gain	Questionnaire, Interviews, Pre- & Post-test	Positive Experience for Culture Learning, but not Language Learning No significant difference in Learning gain
Mai et al. (2022)	Engage students in conversation about exam anxiety to answer solution- and resource-oriented questions	Service	Rasa	Platform	Personal Development and Motivational Theories, Learning Theories	Acceptance, Working alliance, Technical functionality	Questionnaire	Moderate Acceptance & Working alliance scores, Positive Technical functionality (7.6 out of 10)
Mamani et al. (2019)	Support students when searching for academic information	Service	IBM Watson	Platform	–	Satisfaction	Questionnaire	63% of users are satisfied
Martha et al. (2023)	Improve self-regulation and co-regulation skills by providing scaffolding	Teaching	Dialogflow	Platform	Educational Theories	Learning gain, Experience	Pre- & Post-test, Chatlogs, Survey	Positive Experience, Significant difference in Learning gain
Memon et al. (2021)	Facilitate student's learning towards an outcome-based education domain	Teaching	AIML 2.0, Template	Rule-based	–	Accuracy, F1-score, Recall, False discovery rate, False negative rate, False positive rate, Matthews correlation coefficient, Negative predictive value	Automatic	0.94 Sensitivity, 0.53 Specificity, 0.92 Precision, 0.62 Negative predictive value, 0.47 False positive rate, 0.08 False discovery rate, 0.06 False negative rate, 88% Accuracy, 0.93 F1 score, 0.50 Matthews correlation coefficient
Mendoza et al. (2020)	Serve as an extra-school tool and as an intermediary between students and teachers	Service	Dialogflow	Platform	–	Perceived workload	Questionnaire	Positive effect on Perceived workload
Michos et al. (2020)	Assist students in collaborative learning by using intervention strategies	Teaching	Colmooc	Platform	Social Learning Theories	Experience, Chatbot behaviour	Questionnaire, Chatlogs	Knowledge-based prompts are more effective than Social based prompts,

(*continued on next page*)

Table 4 (*continued*)

Reference	Objective		Chatbot Technology & Category		Theory	Evaluation criteria	Tool	Effectiveness
	Chatbot objective	Primary Function	Technology	Category				
								Knowledge-based prompts are preferred over Social based prompts
Mohamed et al. (2021)	Answer student questions in different domains of education	Teaching	Transformer	AI	–	Accuracy, BLEU	Automatic	35% Accuracy, 0.41 BLEU
Mokmin and Ibrahim (2021)	Advise on health issues and fitness-related questions	Service	Dialogflow	Platform	–	Perception	Questionnaire, Interviews	74% had a good Perception, 37% exited the application early
Mzwri and Turcsányi-Szabo (2023)	Answer students' questions in an open-domain	Teaching	BERT. Wit.ai, internet wizard	Hybrid	–	Experience	Questionnaire	Positive Experience
Nasharuddin et al. (2021)	Teach AI concepts to students	Teaching	Dialogflow	Platform	Learning Theories	Experience, Chatbot behaviour	Questionnaire, Experts	Positive chatbot behaviour, Mixed Experience
Nguyen et al. (2021)	Give explanations to students' questions and instruct students to enhance their programming skills	Teaching	Ridge classifier, conditional Random field, Probabilistic model, Template, Rela-scripts-model	Hybrid	–	Accuracy, F1-score	Automatic	82% Accuracy, 0.91 F1-score
Nguyen et al. (2022)	Offer intelligent and personalised learning services by providing information and recommendations for a learning path	Service	Rasa	Platform	–	F1-score, Precision, Recall, Response time, Satisfaction	Automatic, Survey	0.88 Precision, 0.92 Recall, 0.90 F1-score, Positive Feedback
Okonkwo and Ade-Ibijola (2021b)	Help students understand Python's basic syntactic structure and semantics	Teaching	Snatchbot	Platform	–	Experience	Questionnaire	Positive Experience
Oralbayeva et al. (2022)	Assist students in learning the Kazakh Latin alphabet	Teaching	Q-learning	AI	–	Perception, Knowledge	Questionnaire, Pre- & Post-test	No significant difference in Learning gain & Perception
Palasundram et al. (2019)	Answer computer science-related questions	Teaching	seq2seq bidirectional GRU encoder and GRU decoder	AI	–	BLEU	Automatic	0.95 BLEU on word embedding, 0.78 BLEU on Character embedding
Paschoal et al. (2023)	Assist students in expressing their doubts and opinions by providing feedback on uncertainties and convictions	Teaching	AIML	Rule-based	–	Usability, knowledge	Questionnaire, Test	No significant difference in Knowledge, Moderate Usability
Pereira et al. (2023)	Train medical students in the clinical interview approach	Teaching	SBERT, Cosine similarity, Decision tree with confidence value	AI	–	Accuracy, F1-score, Precision, Recall, Confidence Value	Chatlogs, Automatic	86% Accuracy, 0.85 Precision, 0.66 Recall, 0.72 F1-score, 0.87 Confidence Value
Raiche et al. (2023)	Support students in assessing the risks and needs of juvenile offenders	Teaching	Rasa	Platform	–	Perception	Questionnaire	Positive Perception
Ruan et al. (2019)	Help students learn factual information by quizzing them	Teaching	Smooth inverse frequency model, DASH algorithm, Rule-based	Hybrid	–	Experience, Knowledge, Engagement, Accuracy	Questionnaire, Test, Chatlogs	Positive Experience, Significant difference in Recognising and Recalling correct answers,

(*continued on next page*)

Table 4 (*continued*)

Reference	Objective		Chatbot Technology & Category		Theory	Evaluation criteria	Tool	Effectiveness
	Chatbot objective	Primary Function	Technology	Category				
								Higher Engagement 97% Accuracy
Sáiz-Manzanares et al. (2023)	Aid university students by providing feedback and enhancing their learning through metacognitive strategies	Teaching	Dialogflow	Platform	–	Perception, Knowledge	Questionnaire, Test	This study compared whether the current level of degree being studied influences Learning outcomes and Perceived Satisfaction. It does not evaluate the general performance of the chatbot.
Salas-Velasco (2023)	Train students on the convenience of pursuing a master's degree and the suitability of taking out a graduate student loan	Teaching	–	Rule-based	–	Knowledge	Test	Significant difference in Knowledge
Salazar (2023)	Support nursing students in acquiring knowledge in a nursing pharmacology course	Teaching	Chatfuel	Platform	–	Experience	Questionnaire	Positive Experience
Schmitt et al. (2022)	Support students in retrieving course-relevant information and answering course-related questions	Service	Naïve Bayes classifier, Chatterbot, Semantic similarity matching	Hybrid	–	Experience, Knowledge	Questionnaire, Test	Positive Experience, Higher Trust, Significant higher level of Knowledge
Sharma et al. (2023)	Assist maritime trainees in learning Collision Avoidance Regulations	Teaching	IBM Watson	Platform	Educational Theories	Usability	Questionnaire	Positive Usability (73.72 out of 100 SUS score)
Tanana et al. (2019)	Train basic interview and counselling skills of psychotherapy students by providing real-time feedback	Teaching	LSTM encoder + decoder	AI	Learning Theories	Experience, Knowledge	Questionnaire, Test, Automatic	No significant difference in Experience, No significant difference in the Use of open questions, Significant difference in the Use of reflections
Vázquez-Cano et al. (2021)	Teach students about punctuation in the Spanish language	Teaching	Collect chat	Platform	–	Experience, Learning gain	Forum discussion, Pre- & Post-test	Positive Experience, Significant difference in Learning gain
Wambsganss et al. (2020)	Evaluate a course through conversation	Service	Snatchbot	Platform	Social Learning Theories	Experience	Questionnaire	Positive Experience
Wambsganss et al. (2022)	Evaluate a course through conversation	Service	Naïve Bayes classifier, Chatterbot, Google Cloud Services	Hybrid	Social Learning Theories	Experience, Response quality & quantity	Questionnaire, Automatic	Positive Experience, Higher Response quality, Lower Response quantity
Wijayawardena et al. (2022)	Answer questions on a particular subject, give students personalised feedback to improve their overall learning experience and suggest relevant online learning content on a specific subject	Both	Feedforward neural network	AI	–	Accuracy	Automatic	99% Accuracy

(*continued on next page*)

Table 4 (*continued*)

Reference	Objective		Chatbot Technology & Category		Theory	Evaluation criteria	Tool	Effectiveness
	Chatbot objective	Primary Function	Technology	Category				
Winkler et al. (2020)	Teach students about Python by scaffolding	Teaching	NLP.js, Rule-based	Hybrid	Learning Theories, Educational Theories	Experience, Knowledge	Questionnaire, Pre- & Post-test	Positive Experience, Significant improvement in Learning gain
Winkler et al. (2021)	Instruct students to use their knowledge to solve problems by providing scaffolding	Teaching	Amazon services	Platform	Learning Theories, Educational Theories	Experience, Learning gain	Questionnaire, Focus group, Pre- & Post-test	Positive Experience, Significant increase in Knowledge
Xie et al. (2021)	Enhance learners' engagement in CSCL by using motivational interviewing	Teaching	AIML + Pattern Matching	Rule-based	Social Learning Theories, Personal Development and Motivational Theories	Collaborative Learning Engagement	Questionnaire	Significant difference in Engagement
Xu et al. (2022)	Help children to find solutions to problems they encounter while watching a TV show, the CA represents the show's main character	Teaching	Dialogflow	Platform	Learning Theories	Experience, Knowledge, Accuracy, Engagement	Questionnaire, Interviews, Test, Automatic, Chatlogs	Positive Experience and Engagement, Significant increase in Knowledge 89% Intent accuracy, 81% Speech-to-text accuracy

3.1. RQ1: what are the objectives of chatbots in education?

As shown in Table 4, the objectives of the included chatbots were almost always unique. Therefore, the objectives were further analysed by categorising the chatbots based on the primary function of the objective: teaching-oriented and service-oriented. This distinction has previously been introduced by Pérez et al. (2020). The primary argument for this categorisation is the distinct functional goals these chatbots have, which are crucial for developing chatbots effectively for their intended purpose. Teaching-oriented chatbots direct educational outcomes. These outcomes include enhancing user knowledge, improving comprehension, and fostering skill development through interactive and personalised learning experiences. For instance, teaching-oriented chatbots might answer educational questions (e.g., Mzwri & Turcsányi-Szabo, 2023) or provide feedback (e.g., Ruan et al., 2019).

In contrast, service-oriented chatbots are developed to assist with tasks that indirectly support the educational experience. These tasks often involve administrative tasks, logical support, or access to educational resources. For example, service-oriented chatbots can help students manage their course schedules, retrieve information about assignments or deadlines (e.g., Khalil & Rambech, 2022) or recommend personalised learning materials based on a student's preference or progress (e.g., Nguyen et al., 2022). While they contribute to an enhanced learning environment, service-oriented chatbots focus on facilitation and support rather than direct teaching, knowledge delivery, and learning motivation.

Some overlap may occur, for example, when a chatbot assists with goal-setting. In these cases, classification depends on the chatbot's approach and purpose. A teaching-oriented chatbot focuses on setting educational goals by providing guidance or explaining how these goals contribute to improved learning outcomes. On the other hand, a service-oriented chatbot might focus on logistical aspects, such as asking users to input deadlines for tasks. Therefore, understanding the chatbot's intent and the nature of its interaction with users is critical for classification.

This analysis of the objectives showed that the majority of chatbots are teaching-oriented chatbots (n = 46, 65%), followed by service-oriented chatbots (n = 22, 31%), and chatbots that are both teaching- and service-oriented (n = 3, 4%). Teaching-oriented chatbots are primarily aimed at answering questions on specific topics, providing feedback on assignments, and assisting students in their studies. For example, Inq-ITS (Adair et al., 2023) was a chatbot designed to assess and support 8th graders in conducting investigations and developing mathematical models to describe scientific phenomena. Liaw et al. (2023a) developed an AI-powered doctor to train nursing students in sepsis care and interprofessional communication. CikguAIBot (Nasharuddin et al., 2021) was developed to teach university students AI concepts. Students can use CikguAIBot to learn about specific AI concepts, ask questions, or take quizzes with feedback on their answers. These examples demonstrate the versatility of subjects and objectives of teaching-oriented chatbots.

Service-oriented chatbots primarily focus on answering university-related questions (n = 12), e.g., Aujogue and Aussem (2019), El Hefny et al. (2021), Giler et al. (2023), and Gonçalves et al. (2022). Service-oriented chatbots provided additional functionalities beyond answering questions. For example, QMT212 (Kumar, 2021) was designed to use a diversified number of interactions to aid teamwork activities by registering group members, helping with information sharing, monitoring progress and helping with peer-to-peer feedback. Nguyen et al. (2022) developed a chatbot that gives personalised learning advice to job seekers in the IT domain by providing information and recommendations for a learning path according to market trends and the learner's profile. Just like teaching-oriented chatbots, these examples show that service-oriented chatbots were developed to offer a wide variety of services.

Three articles discussed chatbots that could be classified as both teaching-oriented and service-oriented. Kucko Version 2 (Kohnke, 2023) assisted second-language learners by providing a possibility for learners to chat in a foreign language. Furthermore, Kucko Version 2 also offered a service by guiding learners to appropriate learning resources. Chen et al. (2023) developed Sammy, which was first used as a service-oriented chatbot by conducting chatbot-led interviews to gather undergraduate students' perceptions of their learning needs and how a chatbot might support them. After these interviews, Sammy was repurposed to teach AI concepts to business students, thus being teaching-oriented. Wijayawardena et al. (2022) developed a chatbot for 6th-graders, which answers general questions about a particular subject, provides personal feedback to improve the student's learning experience (teaching-oriented functionalities), and can suggest relevant learning content on a specific topic (service-oriented). These examples show that it is possible to develop chatbots that can both offer a service and teach students.

Finally, Fig. 2 shows the trend of the objective functionality. Teaching-oriented chatbots are the majority of chatbots being developed every year, except for 2022, where service-oriented chatbots are the majority. This also coincides with a decrease in teaching-oriented chatbot research. Chatbots that have both objective functions were introduced in 2022 and have been increasing since.

3.2. RQ2: which technologies are utilised when developing chatbots in education?

As shown in Table 4, many different techniques, methods, and algorithms were used to develop chatbots. Therefore, these underlying technologies and approaches were categorised. Agarwal and Wadhwa (2020) distinguished between AI or neural network-based chatbots, and rule-based chatbots, while it is also possible to combine these two techniques in a *hybrid* chatbot. Finally, during the analysis and coding of the papers, several studies mentioned only the development platform or which chatbot builder tool was used without providing any technical details (e.g., Deveci Topal et al., 2021; Kohnke, 2023). These platforms offer prebuilt functionalities that allow developers to create chatbots with minimal custom coding and support a range of approaches depending on implementation choices. In other words, chatbots developed on these platforms may employ AI, rule-based techniques, or a combination of both. However, since many papers that utilised a platform to build their chatbot did not disclose the underlying technologies, and to avoid making assumptions about their implementation, a fourth category, platform chatbot, was introduced. The final four

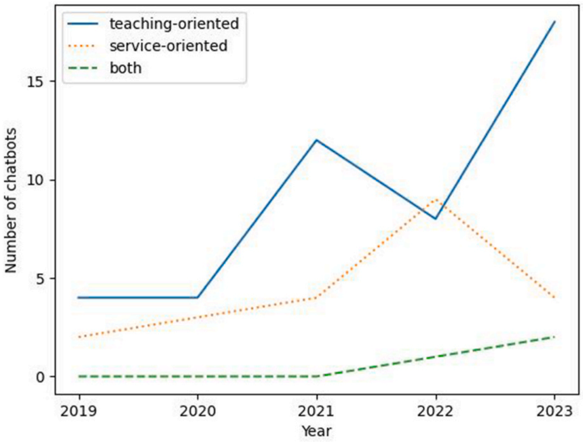

Fig. 2. The trend of the objective function of chatbots in education.

categories are.

(1) AI chatbot: Chatbots that leverage machine learning or artificial intelligence techniques to learn from data, identify patterns, and generate responses. This category includes models utilising neural networks, reinforcement learning, natural language processing (NLP), and other AI-driven methodologies.
(2) Rule-based chatbot: Chatbots using predefined templates and responses.
(3) Platform chatbot: Chatbots developed using a development or chatbot builder platform such as Dialogflow or Rasa, where the underlying technology is not always explicitly specified.
(4) Hybrid chatbot: Chatbots that combine two or more of the above categories.

The most common method to build a chatbot was to use a development or chatbot builder platform (n = 42, 59%). The second most common method was utilising AI as the underlying technology to develop chatbots (n = 12, 17 %), followed by a hybrid approach (n = 9, 13%). Finally, the rule-based method was the least common (n = 8, 11 %).

A diverse range of development and chatbot builder platforms were used to create chatbots in the reviewed studies. In total, sixteen different platforms were identified. Five of these platforms were used in multiple studies, while the remaining eleven platforms were each used in only one study. The most commonly used platform was Google's Dialogflow (n = 14; e.g., Hew et al., 2023). Other frequently used platforms included IBM Watson (n = 6; e.g., Hsu et al., 2023), Rasa (n = 4; e.g., Raiche et al., 2023), Snatchbot (n = 3; e.g., Wambsganss et al., 2020), and Telegram (n = 3; e.g., Giler et al., 2023). Additionally, some studies reported using multiple platforms. For example, González et al. (2022) combined the Dialogflow platform with Amazon services.

A similar trend was observed with AI chatbots, where the majority of AI models and algorithms were exclusively used in the identified chatbots. Only the (feedforward) neural network (Jariwala et al., 2021; Wijayawardena et al., 2022) and the Transformer architecture (Lam et al., 2023; Mohamed et al., 2021) were the two AI models that were used twice. Other techniques that were used include LSTM encoder and decoder (Tanana et al., 2019), hierarchical recurrent attention network (Aujogue & Aussem, 2019), and seq2seq bidirectional GRU encoder and decoder (Palasundram et al., 2019). This highlights the wide range of AI possibilities for chatbots.

When developing a chatbot with a hybrid approach, most methods were used only once. Only the Naïve Bayes classifier and the Chatterbot chatbot builder platform combination were found more than once (Schmitt et al., 2022; Wambsganss et al., 2022). Even several of the utilised development platforms, Wit.ai (Mzwri & Turcsányi-Szabo, 2023) and NLP.js (Winkler et al., 2020), were only introduced in the hybrid development method. Furthermore, certain AI models or algorithms were exclusively found in hybrid chatbots and were not present in AI chatbots, e.g., BERT (Mzwri & Turcsányi-Szabo, 2023) and DQN (Chen et al., 2021). This shows the variety of algorithms, techniques, and platforms available to develop a hybrid chatbot.

The most common approach for the rule-based method was to use AIML or AIML 2.0 in combination with another rule-based method, such as Pattern Matching (Xie et al., 2021) or using a template (Memon et al., 2021). Three articles (Adair et al., 2023; Lin & Ye, 2023; Salas-Velasco, 2023) only mention that they are a rule-based chatbot, but it has not been specified exactly which method they utilised.

Finally, Fig. 3 illustrates the trend of the utilised technologies per year. This shows that using a platform to develop a chatbot has been the most used method from 2020 onwards, while using AI, building a rule-based chatbot, or using a hybrid approach never differentiated a lot from each other.

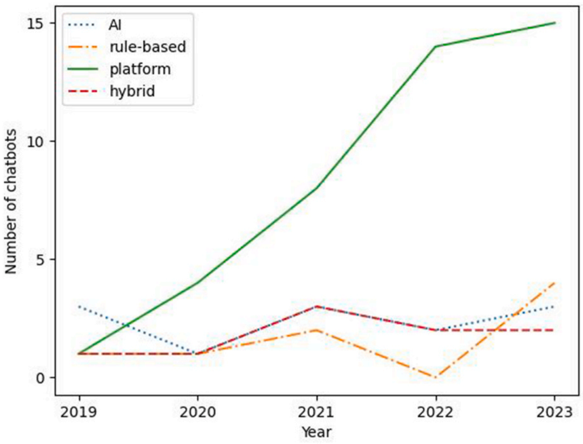

Fig. 3. The trend of the underlying technologies of chatbots in education.

3.3. *RQ3: how do various theories inform the integration of chatbots in educational settings?*

As shown in Table 4, only twenty papers (18%) informed the integration of chatbots in education with a theory or theories. To effectively analyse the identified theories, they have been categorised into four categories.

(1) Learning Theories: This group includes theories that fundamentally address how learning occurs and the frameworks supporting these learning processes, e.g., Media Equation Theory (Mai et al., 2022).
(2) Educational Theories: This group encompasses specific methods, strategies, or practices designed to enhance the learning experience and address the educational needs of learners, e.g., Practice-based Teacher Education (Lee & Yeo, 2022) and scaffolding (Winkler et al., 2021).
(3) Social Learning Theories: This group focuses on the social aspects of learning environments, e.g., Computer-Supported Collaborative Learning (CSCL) (Xie et al., 2021).
(4) Personal Development and Motivational Theories: This group includes theories and techniques that are oriented towards personal growth and the psychological aspects of learning and development, e.g., self-regulated learning (Hew et al., 2023) and intervention techniques (Mai et al., 2022).

Educational Theories was the most found category (n = 9), followed by Learning Theories (n = 8), Social Learning Theories (n = 5), and Personal Development and Motivational Theories (n = 4). Some papers used multiple theories across different categories, so the total count of theories exceeds twenty, which is the number of papers that use at least one theory to inform their integration into education.

Fig. 4 illustrates the trend of the theory used over the last few years. This figure shows that there is no visible trend in the type of theories that are used to inform the introduction of chatbots in education. The line that shows the chatbots that use a theory to inform the introduction of chatbots in education, Chatbots with Theory, has remained constant at around five since 2021. In contrast, there

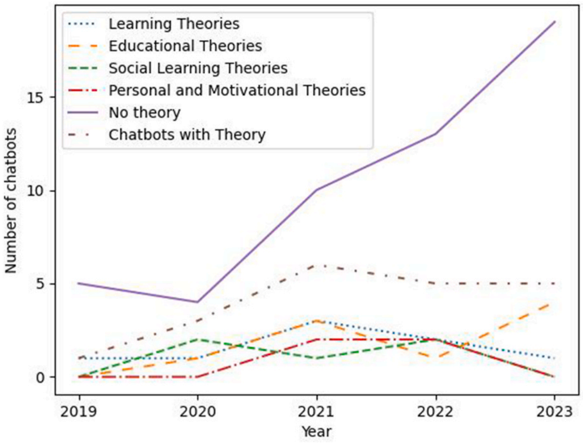

Fig. 4. The trend of the theories to inform the introduction of chatbots in education.

has been a significant increase in the number of chatbots introduced without any theoretical foundation.

3.4. RQ4: which criteria are used to evaluate chatbots in education?

The analysis of the reviewed articles (see Table 4) identified 37 unique criteria for evaluating chatbot effectiveness. While most articles used multiple criteria, 25 of the 37 criteria were measured in only one of the 71 reviewed articles. To be able to analyse these evaluation criteria, they have been categorised into four categories.

(1) Perceptual factors: These relate to users' perceptions and experiences with the chatbot, including how useful, useable, and motivating they find the chatbot.
(2) Behavioural factors: These involve how students behave in response to using the chatbot, including their engagement, learning strategies, collaboration with peers, demonstration of competencies, and learning gain.
(3) Technical and Analytical factors: These focus on the technical performance and analytical measures of the chatbot, including its functionality, such as accuracy and BLEU, and the quality of student responses.
(4) Psychological factors: These address the psychological impacts on users, such as motivation, self-efficacy, and learning-related attitudes.

These categories were inspired by well-established frameworks in the fields of information systems and technology acceptance. The DeLone and McLean Model of Information Systems (DeLone & McLean, 2003) provides a comprehensive structure for evaluating system success, aligning closely with perceptual, behavioural, and technical factors. Meanwhile, the Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2003) further informed this categorisation by highlighting performance expectancy, behavioural intention and self-efficacy, which align with perceptual, behavioural, and psychological factors. The analysis of these categories showed that Perceptual factors were the most used criteria (n = 49, 69%), followed by Behavioural (n = 32, 45%), Technical and Analytical (n = 26, 37%), and Psychological (n = 9, 13%) factors.

Several tools were identified for measuring Perceptual factors. One questionnaire that was repeatedly used, is the system usability score (SUS). El Hefny et al. (2021) used the SUS questionnaire to evaluate the usability of their service-oriented chatbot, which provided students with information regarding the content, dates, timings, locations and deadlines for their course. Another questionnaire that looked at the usability or acceptance of a chatbot is the Technology Acceptance Model (TAM). Liaw et al. (2023b) developed an AI medical doctor to help nursing students with interprofessional communication and used the TAM questionnaire to measure the chatbot's acceptance by students. Besides questionnaires, other methods such as interviews, focus groups, forum discussions, or a combination of several tools have also been used to measure Perceptual factors. Fenbotum (Deveci Topal et al., 2021) was a chatbot that helped secondary school students understand a science unit. A focus group was used to gain a better understanding of how students experienced the chatbot. Vázquez-Cano et al. (2021) analysed the use of a forum to see how students experienced the chatbot that helps them to learn Spanish. Elinor (Xu et al., 2022) was a chatbot developed for 4-6-year-olds. The authors not only measured the experience of the children, but they also interviewed parents about their experiences. This demonstrates a multitude of ways to measure Perceptual factors.

Behavioural factors were frequently measured using a pre- and post-test design, a post-test, or a questionnaire. For example, LANA-I (Aljameel et al., 2019) is a chatbot specifically developed for children with Autism Spectrum Disorder. They measured the learning gain using a pre-and post-test of their chatbot when using the Visual, Auditory and Kinaesthetic learning styles model (VAK). Vázquez-Cano et al. (2021) developed a chatbot to teach students about punctuation in the Spanish language, and they measured the learning gain using a pre- and post-test. Xie et al. (2021) used a questionnaire to measure the increase in engagement of students in CSCL while using their developed chatbot.

Technical and Analytical factors were often automatically measured. For example, Leonardi & Torchiano (2023) developed a chatbot to answer questions from students about object-oriented programming, and they used accuracy to measure if the chatbot recognised the correct intent. They also used other automatic criteria to evaluate the effectiveness of their chatbot. In the research of Chen et al. (2021), the measured accuracy is whether the chatbot can correctly recognise students' need deficiencies. Psychological factors were most frequently measured using questionnaires. For example, Durall and Kapros (2020) measured the confidence a user had in their ability to use core competencies using a questionnaire.

Fig. 5 illustrates the trend of the four categories. Perceptual factors are the most common to evaluate chatbots, followed by Behavioural, Technical and Analytical, and Psychological factors. A steady increase has been observed in the measurement of Psychological factors and Behavioural factors, particularly since 2020. In contrast, the measurement of Perceptual and Technical factors declined in 2023.

3.5. RQ5: how effective are chatbots in education?

The results of RQ4 indicate that there is no standardised approach to evaluating chatbots. Chatbot effectiveness was categorised based on how well chatbots achieved their intended objectives, considering performance metrics, learning outcomes, user experience, and findings from the reviewed studies. Chatbots were categorised into four categories.

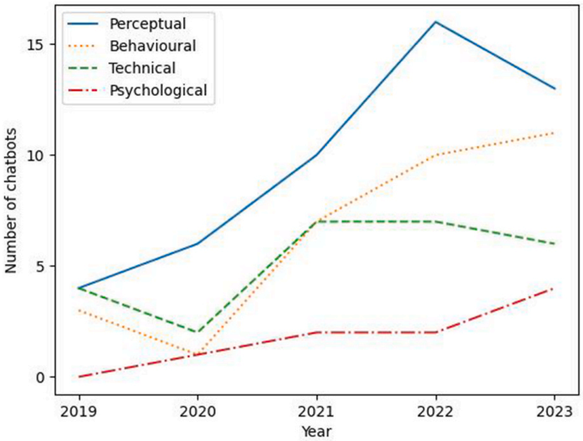

Fig. 5. The trends of evaluation criteria of chatbots in education.

(1) Effective chatbots: They successfully achieved all their intended objectives, demonstrated a significant impact (e.g., measurable learning gains, enhanced user engagement, or strong usability scores), and/or received consistently positive evaluations from users or experts.
(2) Partially effective chatbots: They demonstrated some positive outcomes but did not fully meet all intended objectives. For example, a chatbot might improve user engagement but show no measurable effect on learning gain or knowledge retention.
(3) Ineffective chatbots: They failed to achieve their primary objectives, showing no significant impact on learning or performance. Chatbots were also placed in this category if they delivered a predominantly negative user experience, even if minor positive aspects were reported.
(4) Unclear effectiveness: The included studies lacked sufficient evidence to determine effectiveness. This includes cases where two chatbot versions were compared without evaluating overall impact or where performance metrics (e.g., accuracy or BLEU score) were presented without clear benchmarks for interpretation.

According to the evaluated criteria and claims in the effectiveness column of Table 4, 54 chatbots (76%) were classified as effective. Effectiveness in this context refers to the extent to which chatbots achieve their intended objectives, the performance based on automatic metrics, or the user experience of interacting with the chatbot. Some chatbots were only partially effective (n = 9, 12%), others were ineffective (n = 4, 6%), and for some, effectiveness was unclear (n = 4, 6%). The chatbots that fall under these latter three categories are discussed further below.

First, we discuss the partly effective chatbots. Han et al. (2022) developed a chatbot which did have a significant effect on several of the students' interests in education and self-directed learning, but there was no significant difference in knowledge increase, confidence increase, and experience. Mathbot (Cai et al., 2021) did learn a personalised policy for each user, however, there was no significant difference in learning gain between using the chatbot and using videos from Khan Academy. The virtual agent developed by Liaw et al. (2023a) trained nursing students in sepsis care and interprofessional communication. The virtual agent did not significantly impact interprofessional communication knowledge or self-efficacy but significantly helped nursing students learn about sepsis care. The users who communicated with Fenbo (Deveci Topal et al., 2021) had a positive experience, however, Fenbo had no significant influence over the gained knowledge of the participant. The objective of the chatbot developed by Durall and Kapros (2020) was to raise awareness, spark curiosity and understand core competencies. This chatbot received positive feedback from participants and participants had an increase in confidence about their core competencies, but when compared with online forms, online forms were preferred. An expert positively evaluated CikguAIBot's behaviour (Nasharuddin et al., 2021), but user feedback on the chatbot experience was mixed. The chatbot QMT212 (Kumar, 2021), who facilitated group work, had a significant influence on the learning gain and teamwork criteria, but there was no significant difference in the need for cognition, perceived motivation, self-efficacy and perception of learning. NEdBOT (Ali et al., 2022) received a positive usability score, demonstrating measurable success. However, its accuracy (77%) and response time (216ms) were reported without clear benchmarks, making their impact on effectiveness uncertain. Due to its confirmed usability success but unclear performance in other areas, it is classified as partially effective. In students who interacted with Clientbot (Tanana et al., 2019), there was no significant difference in the use of open questions or user experience, but a significant difference was observed in the use of reflections.

Second, we discuss the chatbots that are ineffective. The developed chatbot by Mohamed et al. (2021) had an accuracy score of 35 per cent, meaning that the chatbot only recognises the intent of the user in 35 per cent of the cases, which is low. Therefore, this chatbot has been categorised as ineffective. The chatbot TOB-STT (Paschoal et al., 2023) had no significant impact on the gained knowledge and only had a moderate usability score. The objective of AsasaraBot (Mageira et al., 2022) was to teach students cultural content in a foreign language. Since AsasaraBot had no significant impact on learning gain in both culture learning and language learning, and user experiences were mixed, it was classified as ineffective. Oralbayeva et al. (2022) developed a chatbot to assist students in learning the Kazakh Latin alphabet, but there was no significant difference in learning gain and perception of using the chatbot.

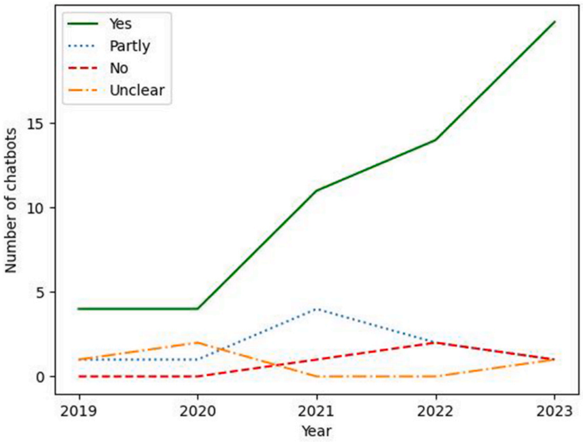

Fig. 6. The trend of effectiveness of chatbots in education.

For several chatbots, it was not clear whether the developed chatbots were effective. For example, Michos et al. (2020) compared two CSCL strategies with each other, but it was not clear if using a conversational agent was effective at all. The same holds for the chatbot developed by Aljameel et al. (2019). They compared two chatbots with each other; one uses VAK learning strategies, and the other does not, not explaining if using a chatbot is effective at all. The chatbot developed by Sáiz-Manzanares et al. (2023) was not being evaluated on performance. This study compared whether the current level of degree being studied influences learning outcomes and perceived satisfaction when using a conversation agent. Khin & Soe (2020) evaluated the chatbot automatically by looking at the BLEU score. The chatbot received a BLEU score of 0.41, but it is not clear whether this is a good score or a bad score.

Fig. 6 illustrates the trend of the effectiveness of chatbots in education. The major visible trend is that effective chatbots have always been in the majority. Furthermore, the ratio of effective chatbots has been increasing since 2021. What the figure also shows is that in the first two years, there were no ineffective chatbots, while every year, there has been at least 1 chatbot that was partly effective or it was not clear whether the chatbot was effective. This figure also shows that there is never a big difference between the no, partly and not clear categories. There is, at maximum, a difference of four.

3.6. RQ6: how do the interactions between objectives, technology, theories, and evaluation criteria influence the overall outcome and success of effective chatbot implementation?

Fig. 7 illustrates the constructed social network, highlighting the interactions between objectives (RQ1), technology (RQ2), theories (RQ3), and evaluation criteria for effective chatbots (RQ4/RQ5). An interaction represents the simultaneous presence of two elements in an effective chatbot. For example, the analysis examined whether teaching-oriented chatbots were rule-based or whether chatbots grounded in learning theories were evaluated using perceptual factors. The nodes have been coloured based on the element they belong to. The size of the nodes indicates how often a specific category was found in the effective chatbots. The thickness of edges between nodes indicates how often this relation was found in effective chatbots, hence, these can be read in two ways. Interactions with an insignificant amount of representation, less than ten per cent, were not included in the figure.

When examining the nodes in Fig. 7, the results of RQ1 – RQ 4 are represented accordingly. The most common category for each specific element (RQ) is represented by the largest node, highlighting their prevalence in effective chatbots. For the objective element, teaching-oriented chatbots were the most common. Among the technology elements, the most effective chatbots were developed using a development or chatbot builder platform. For the theory element, the majority of chatbots did not justify their use in education with a theory; among those that did, educational theories were the most common. Lastly, perceptual factors were most frequently measured in the evaluation criteria element. The interactions between the largest nodes of each element were also most commonly found. Thus indicating that these interactions influence the development and implementation of effective chatbots.

However, several interactions might be underexplored or overlooked since they are not found in this literature review (see Fig. 7). Five categories are not represented by interactions in Fig. 7: rule-based chatbots, psychological evaluation factors, chatbots that use a social or personal theory to inform the integration in education, and chatbots that are both teaching- and service-oriented, indicating these categories are overlooked or underexplored. Furthermore, Fig. 7 shows that both teaching- and service-oriented chatbots are only represented when developed by a platform. It is also noteworthy that only educational theories are significantly represented by teaching-oriented chatbots, while the other theories are not represented by any objective element. It also shows that there is not a big difference in using perceptual, behavioural, and technical factors to measure the effectiveness of teaching-oriented chatbots. Service-oriented chatbots focus more on perceptual factors, followed by behavioural and technical factors.

When looking at the interactions between the technology and evaluation criteria for platform and hybrid chatbots, perceptual factors are the most common. Meanwhile, these are not represented for AI and rule-based chatbots, AI chatbots focus more on technological factors. Furthermore, only chatbots developed while using a platform have an interaction with the theory category educational theories. On the other hand, interactions between AI, rule-based, and hybrid chatbots with a theory element are not

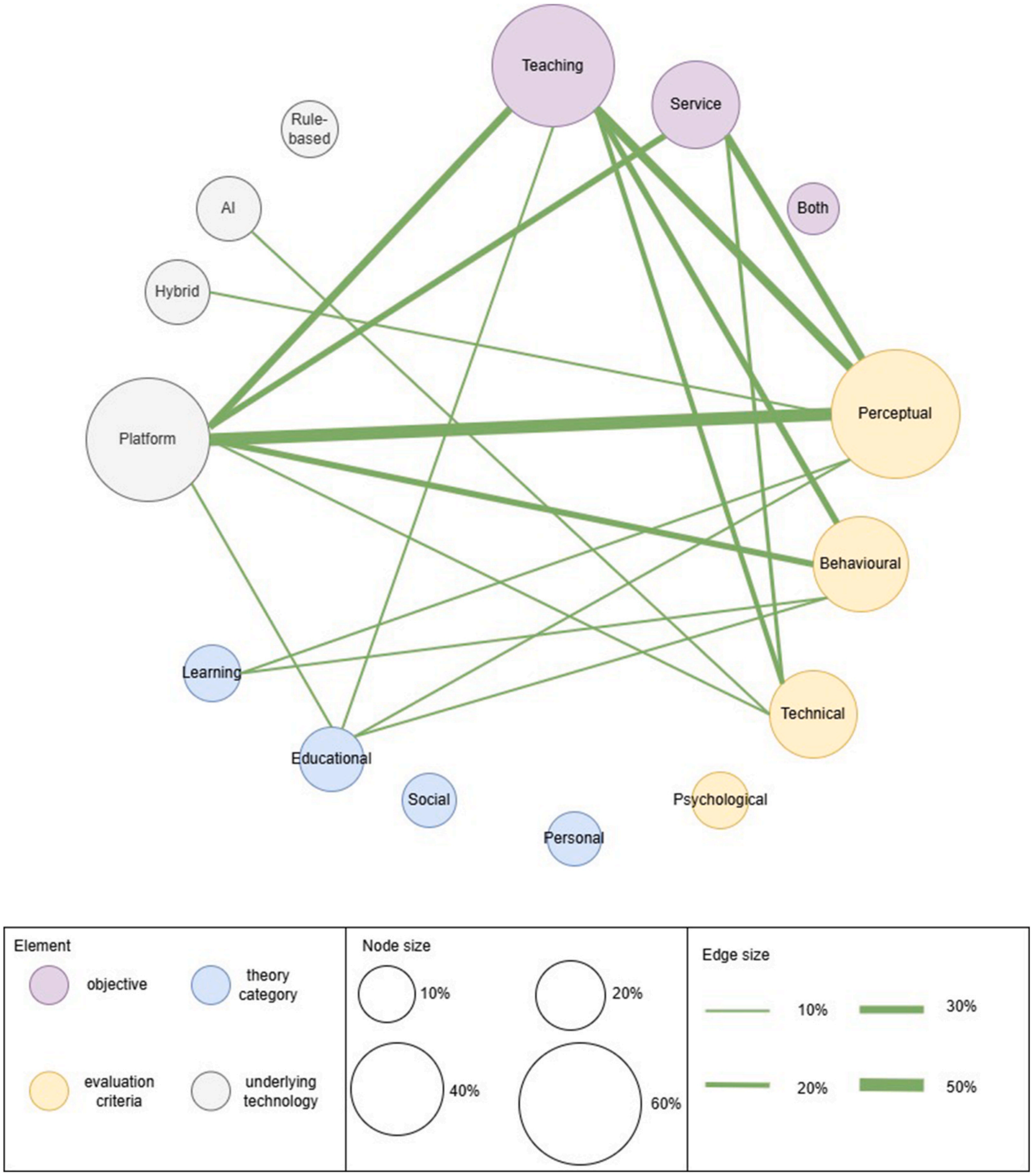

Fig. 7. The interactions of different elements of effective chatbots.

represented. It is also noteworthy that chatbots that use a learning or educational theory focus mainly on perceptual and behavioural factors and less on technical factors.

4. Discussion

This systematic literature review presents a comprehensive overview of the current research on chatbots in education with a focus on objectives, underlying technology, theory, evaluation criteria, effectiveness, and the interactions of these elements. When looking at all the trend analyses, only articles published before October 2023 were taken into account. These analyses might change slightly when looking at the whole year of 2023. In the following sections, an exploration and interpretation of the five components is undertaken followed by an in-depth analysis of the interactions between these five components.

4.1. Objectives of chatbots in education

Classifying chatbots in our review proved challenging due to the lack of clear definitions in existing literature. Pérez et al. (2020)

categorised chatbots as teaching- and service-oriented but did not provide precise definitions, leading to ambiguity in applying their framework. None of the reviewed studies explicitly adopted this classification, making it difficult to distinguish between chatbot types. Additionally, inconsistent reporting on chatbot objectives and functionalities across studies further complicated the categorisation process and highlighted the need for a standardised framework in future research.

Despite these classification challenges, our findings still reveal clear trends in chatbot development. In particular, there is a growing emphasis on STEM-focused teaching chatbots, followed by language-teaching chatbots. This aligns with the findings of Kuhail et al. (2023), however, it misaligns with Hwang and Chang (2021) and Wollny et al. (2021). Both reviews identified language learning as the most educational topic. One possible explanation for this difference is the variation in search criteria. This research focused on the term education, whereas Hwang and Chang (2021) and Wollny et al. (2021) employed broader search strategies. Hwang and Chang (2021) used general chatbot-related terms to search within the Web of Science database, restricting their analysis to categories such as "Education Educational Research," "Education Scientific Disciplines", and "Psychology Educational". While their broad search terms may have led to the inclusion of more language-learning chatbots, their restrictive filtering resulted in a dataset of only 30 publications, potentially limiting the generalisability of their findings. Similarly, Wollny et al. (2021) included general education-related keywords (e.g., learner and teaching), which likely captured a wider dataset and contributed to the greater occurrence of language learning chatbots in their findings. By contrast, the more targeted use of the term education in this study may have amplified the visibility of STEM-related research.

Another factor is the temporal context. The data of this research shows an increase in chatbots developed for STEM subjects starting in 2021, suggesting a shift in research focus. This aligns with Kuhail et al. (2023), who also employed general education-related keywords yet observed similar trends. Furthermore, the inclusion criteria in both this study and Kuhail et al. (2023) prioritised technological depth, which might favour STEM-focused research. Since our findings highlight a growing emphasis on STEM applications in chatbot development, they point towards a need for broader subject coverage in future research. This shift also underscores the evolving landscape of educational technology and the importance of adapting research methodologies to capture these trends.

Beyond the focus on STEM education, trends in chatbot functionalities also show interesting shifts over time. The trend analysis indicated a decline in the development of service-oriented chatbots in 2023 compared to 2022. One possible explanation for this trend is that due to the increase in online education (Mohamad Noor, 2023), there is a shift in the focus of researchers towards chatbots that directly support the learning process. However, service-oriented chatbots remain important. For instance, online education presents challenges in group work, where chatbots could provide support (e.g., Kumar, 2021). Additional services include course recommendation systems (e.g., Nguyen et al., 2022) and library assistance (Mckie & Narayan, 2019).

While the focus on service-oriented chatbots has declined, our findings suggest an emerging trend towards integrating both teaching and service functionalities into a single chatbot, contrary to Pérez et al. (2020). This indicates the expanding capabilities of chatbots, likely driven by advancements in AI. As technology improves, integrating multiple functionalities into a single chatbot becomes more feasible. However, the number of chatbots that combine teaching and service-oriented functionalities is still relatively small. This is surprising since the development of multifunctional chatbots could also contribute to a more inclusive educational experience by addressing both academic and administrative needs in a single tool, making them particularly valuable in resource-constrained settings. One possible explanation for this finding is that developing a multifunctional chatbot is more challenging. It may require advanced AI models, in-depth knowledge of how to develop or properly utilise these models, and significant resource investment, which may not be feasible for many developers.

4.2. *The underlying technology of chatbots in education*

This review, based on the categories suggested by Agarwal and Wadhwa (2020) , categorised the underlying technology of chatbots into four categories: rule-based, AI, hybrid, and platform. It is important to note that these four categories are not entirely conceptually distinct, as some inherently overlap. Hybrid chatbots explicitly combine rule-based and AI-driven approaches, while platform chatbots may employ AI, rule-based, or hybrid methods but were categorised separately due to the lack of reported implementation details in many studies (e.g., Deveci Topal et al., 2021; Kohnke, 2023). Maintaining a distinct category for platform chatbots was, therefore, necessary to prevent misclassification and to ensure an accurate representation of chatbot technologies without making assumptions about missing details. While this classification system does not offer fully discrete conceptual boundaries, it ensures a rigorous and transparent approach to analysing chatbot development trends.

The categorisation used in this study revealed significant insights. The most noticeable finding was the predominance of development or chatbot builder platforms, suggesting the ease of use and availability of pre-built tools. This trend towards platform-based development highlights the accessibility and convenience these platforms offer. However, many reviewed studies that utilised a platform to develop their chatbot were ambiguous and will not easily be reproducible. They often only mention which platform was used, lacking other technical information or they did not share the data or knowledge base used to develop their chatbot. As a result, reproducibility is not possible without extensive knowledge of these systems. This ambiguity further complicates the ability to evaluate and compare the efficacy and costs of different platforms, even when designed for the same objective.

While this review provides a structured categorisation of chatbot technologies, earlier studies approached classification differently. For example, Pérez et al. (2020) did not explicitly categorise chatbots based on their underlying technology but focused on the language processor used to interpret user messages. Thus, it remains unclear which technology was used to develop these chatbots. The current review includes more recent development approaches and methods that might not have existed in 2020. Even so, several techniques, such as Dialogflow and Chatfuel, were found in both reviews.

One common denominator across all categories (rule-based, AI, hybrid, and platform) was that most technologies were only utilised once, indicating the wide range of possibilities and the rapid technological advancements in chatbot development. This aligns with findings from Ouyang et al. (2022), who reviewed educational AI applications and noted the variety of AI algorithms and models used. However, by focusing solely on AI, the review of Ouyang et al. (2022) misses out on development platforms that incorporate state-of-the-art models, which can still provide significant benefits in practical applications. In contrast to AI chatbots, there is a more standardised method for developing a rule-based chatbot, namely, using AIML in combination with another rule-based approach. This standardisation might indicate reliability and maturity in this rule-based approach.

Interestingly, the trend analysis did not show an expected increase in utilising AI as the underlying technology of educational chatbots despite the rapid advancements in AI. There was no significant difference between AI, rule-based, or hybrid chatbots. One possible explanation for the preference for using platforms to develop chatbots is that these platforms lower the entry barrier, allowing developers to create chatbots without extensive coding knowledge. Additionally, these development or chatbot builder platforms incorporate state-of-the-art AI models. For example, Dialogflow incorporates foundational models developed by Google, such as BERT, enabling advanced conversational capabilities (DialogFlow, n.d.). However, the lack of detailed implementation descriptions in many papers complicates the classification process. Because platform-based chatbots may use AI, rule-based, or hybrid techniques, categorising them separately helps avoid unverified assumptions. Including platform chatbots in the hybrid category without knowing their architecture could skew the statistical analysis. This reinforces the rationale, introduced earlier, for maintaining platform chatbots as a distinct category, supporting classification accuracy and interpretative transparency.

Despite their growing prominence in chatbot development, foundational AI models like Large Language Models (LLMs) are underrepresented in the reviewed literature. Foundational models like ChatGPT, Google Bard, and other LLM-based systems have recently transformed the landscape of chatbot applications by enabling advanced conversational capabilities, personalised learning support, and scalability. These models lower the barrier to creating sophisticated chatbots by providing pre-trained architectures that can be fine-tuned for specific educational contexts (Albadarin et al., 2024; Yigci, Eryilmaz, Yetisen, Tasoglu, & Ozcan, 2024).

The limited use of foundational models in the studies analysed in this research may stem from the timeframe of the publications or challenges inherent in integrating these technologies, such as the computational cost and expertise required. Their absence in the reviewed literature highlights a critical gap. As foundational models continue to gain traction, their incorporation into educational chatbot development could address existing limitations, such as providing multilingual support and dynamically adapting to individual learners' needs by leveraging the advanced capabilities of generative AI (Albadarin et al., 2024; Yigci, Eryilmaz, Yetisen, Tasoglu, & Ozcan, 2024). The shift towards foundational models presents a significant opportunity to deepen theoretical understanding and enhance practical applications of educational chatbots across diverse learning environments. Nevertheless, our research suggests that an alternative exists alongside chatbots directly based on foundational models. Development platforms, which often integrate well-tested foundational models (Google Cloud, n.d.), offer a more accessible option, enabling chatbot creation with greater ease and lower technical barriers. This distinction between platforms and foundational models provides practical guidance for educators: While platforms are ideal for creating entry-level chatbots, foundational models offer advanced functionalities for complex, personalised interactions. For instance, an educator could use a platform like Dialogflow to build a chatbot for basic FAQs while working with foundational models like GPT-4 for complex, domain-specific tutoring.

4.3. Theories that inform the integration of chatbots in education

The analysis of the reviewed papers showed that all studies justified introducing chatbots in education by referencing prior successful implementations or practical reasons, such as the COVID-19 pandemic (e.g., Deveci Topal et al., 2021) and increased teacher-student ratios (e.g., Chen et al., 2023). It is concerning that many papers lack theoretical explanations and foundations for their chatbot. The absence of a theoretical foundation in many chatbot implementations aligns with Chen et al. (2020), who highlighted the lack of theoretical grounding in highly cited educational AI applications.

In educational contexts, theories explain how learners absorb, process and retain knowledge (Ertmer & Newby, 2013). Without a theoretical foundation, chatbots may miss opportunities to align with cognitive and learning processes, reducing their effectiveness in delivering meaningful educational experiences. Theories provide a structured approach, ensuring chatbots address specific cognitive and motivational needs rather than functioning as generic tools. Banihashem and Macfadyen (2021) highlight three essential roles of theories in the effective implementation of educational tools: (1) Theories underpin the design and selection of tools. For example, the cognitive theory of multimedia learning (Mayer, 2014) guides the design of chatbots that effectively integrate text, audio, and visuals to enhance learning. (2) Theories inform the use of tools by clarifying assumptions about learning processes. For instance, the constructivist learning theory (Jonassen, 1991) supports the development of chatbots that engage students in active learning through problem-solving and social interaction. (3) Theories aid in interpreting data by explaining learning behaviours and outcomes. For example, cognitive load theory (Sweller et al., 1998) explains how cognitive overload occurs when too much information is presented, helping to interpret data on student struggles. In other words, integrating learning theories can help educators select and utilise educational chatbots more effectively. This includes determining the best contexts for implementation, choosing the appropriate technology, identifying the data that needs to be collected from the chatbot, and determining how that data should be visualised and reported to optimise educational outcomes. Therefore, a solid theoretical foundation ensures that chatbots align with cognitive, motivational, or social learning theories, which can enhance their effectiveness.

Theoretical frameworks can also prevent issues related to "technological determinism", where tools such as chatbots are adopted simply because they are available or convenient rather than being aligned with specific learning goals of pedagogical design (Banihashem & Macfadyen, 2021). For example, Mageira et al. (2022) developed a chatbot to teach a foreign language that lacked a theoretical foundation and was not effective. In contrast, a foreign language teaching chatbot developed by Hew et al. (2023), with a theoretical foundation based on personal development and motivational theories, and social learning theories, proved effective. The theoretical foundation likely ensured that the chatbot was aligned with key educational principles, supporting student motivation and engagement.

In this study, all chatbots classified as ineffective lacked a theoretical foundation. However, some effective chatbots also lacked a theoretical basis, suggesting that while theory can significantly enhance the design and implementation of chatbots, it is not the sole factor influencing outcomes. Grounding chatbot design in theory ensures alignment with educational principles, such as incremental knowledge scaffolding and promoting student engagement through motivational strategies. Without this foundation, chatbots might overwhelm learners with excessive information or fail to engage them adequately, resulting in poor learning outcomes. But even among the reviewed studies that have a theoretical foundation, an in-depth explanation of how the theory is utilised is sometimes lacking. Without such details, it becomes challenging for other researchers to replicate these approaches or build upon them. Providing clearer insights into the operationalisation of theories would enhance transparency and support the broader adoption of theoretically informed chatbot designs.

Other considerations, such as the technical architecture and design, also play a critical role in determining effectiveness. For example, two AI-based chatbots designed to teach a foreign language, both lacking a theoretical foundation, produced different results: one was effective (Chae et al., 2023), while the other was not (Oralbayeva et al., 2022). The contrasting outcomes were likely due to technical differences, indicating that other factors also influence effectiveness. Nonetheless, a solid theoretical grounding can help ensure that chatbots are not only technically sound but also pedagogically meaningful. Theories of learning guide the selection and design of chatbots, ensuring that they provide relevant, structured, and meaningful interactions for students, which could otherwise be missed without such a foundation. Furthermore, chatbots grounded in theory are more likely to succeed because they align with established learning principles, reducing the risk of ineffectiveness.

Interestingly, other literature reviews on educational chatbots and AI applications, such as Chiu et al. (2023) and Ouyang et al. (2022), did not discuss the theoretical foundations of these technologies. While some reviews address the ethical challenges of educational chatbots (e.g., Okonkwo & Ade-Ibijola, 2021a; Pérez et al., 2020), they rarely address the theoretical frameworks guiding the integration of these technologies. This lack is surprising given the importance of theories in informing the integration of chatbots in education. However, despite this oversight in the literature, several well-established theories could provide a robust foundation for the design and integration of chatbots in education. Zhang et al. (2023) suggest eight theoretical frameworks that may be useful in analysing and supporting chatbot-assisted learning. For example, their research highlights constructivism (Winkler & Söllner, 2018) and cognitive theories of multimedia learning (Mayer, 2014) as potential frameworks. Beyond Zhang et al.'s suggestions, additional theoretical frameworks that may also be relevant include cognitive load theory (Sweller et al., 1998), and gamification and motivation (Deterding et al., 2011). Constructivism emphasises learning as an active process in which learners build their understanding through interactions with others and their environment (Winkler & Söllner, 2018). Cognitive load theory (Sweller et al., 2011) and cognitive theory of multimedia learning (Mayer, 2014) both suggest that learners have limited cognitive capacity and that educational tools should be designed to manage this load effectively. Chatbots, when designed well, could offer this by breaking down complex information or offering just-in-time support (Sweller et al., 2011). Lastly, theories related to gamification and motivation emphasise that using game-like elements can enhance engagement and motivation in learning environments (Deterding et al., 2011). Despite the availability of these and other relevant theories, most educational chatbots are still integrated without a theoretical foundation.

Hwang and Chang (2021) is one prior review that briefly identified the theoretical frameworks used in educational chatbots, though their primary emphasis was on learning strategies. However, their analysis focused only on which strategies were employed, without exploring the underlying theories that informed them. Furthermore, the study did not assess whether chatbot designs were explicitly grounded in theoretical frameworks or simply assumed to align with a particular strategy. Therefore, our study conducts a more comprehensive analysis of the educational theories that inform chatbot integration in education, offering a broader theoretical perspective.

The trend analysis further highlighted the increasing gap between chatbots with theoretical underpinning and those without. One explanation for this trend is the accelerated adoption of distance and online education (Shams et al., 2022). Chatbots are introduced to help bridge the gap between teachers and students due to asynchronous learning and the round-the-clock availability of chatbots (Pérez et al., 2020). Another possible reason is the increasing teacher-student ratio, which contributes to a higher workload for teachers. Chatbots are introduced to handle repetitive tasks, such as answering frequently asked questions. If the answers to these questions are pre-programmed, it might diminish the perceived need for a robust theoretical foundation. Practical considerations, such as managing high teacher-student ratios and alleviating teacher workloads, often outweigh theoretical motivations.

Another explanation for the lack of theoretical reasoning is that some papers enhance existing technologies, for which the benefits have already been shown. If the effectiveness of a technology has already been demonstrated, developers might see little need for further theoretical justification. This reasoning can also be applied when papers are referring to other successful chatbots on the same topic. Finally, it is also possible that the chatbots are developed solely by computer scientists with no or limited knowledge of educational theories, focusing primarily on the functionality and performance of their AI models and algorithms.

4.4. Evaluation criteria of chatbots in education

The results of the fourth research question revealed an abundance of unique criteria used to evaluate chatbots. To further investigate these criteria, they have been categorised into perceptual, behavioural, technical and analytical, and psychological factors. Our categorisation aligns with Hobert (2019), who kept the most common criteria as their category and added categories for technical correctness and further psychological factors. This alignment indicates that our categorisation is well-reasoned and grounded in the existing literature.

Our analysis found that perceptual factors are the most commonly measured, aligning with the findings of Jeon, Lee, and Choi (2023), who focused on speech-recognition chatbots for language learning. One possible reason is that the reviewed articles discuss chatbots that are still in their infancy. Before conducting experiments to measure the effectiveness of chatbots, researchers prioritise understanding if the chatbot is perceived well. This follows the design science research methodology (Wieringa & Heerkens, 2006).

Moreover, our analysis revealed that perceptual factors are often measured with questionnaires, which are less time-consuming and intrusive compared to evaluating learning gains through tests. Self-administered questionnaires, after interacting with a chatbot, can ensure faster and more accurate data collection (Malhotra, 2006). Comparing this to measuring learning gain, where participants need to perform a test, it is more time-consuming for both the participants and researchers. The ease and convenience of questionnaires allow for several factors to be measured simultaneously, and they can be conducted after a test, leading to the frequent measurement of perceptual factors. Our analysis showed that fourteen papers combined perceptual and behavioural factors.

However, a critical question arises regarding the sufficiency of evaluating chatbots solely based on subjective user perceptions, particularly for teaching-oriented chatbots. Does the proposed chatbot fulfil its educational purpose? Similarly, relying solely on technical factors may be insufficient. While technical criteria measure quality on paper, such as the ability to recognise user intent, they do not guarantee overall effectiveness. For question-answer chatbots, technical factors might seem sufficient, but this assumption holds only if responses are pre-programmed. Without pre-programmed answers, there is no guarantee of accurate responses, even if the chatbot recognises the intended question. For example, Pereira et al. (2023) exclusively measured technical factors, such as answer accuracy, but did not evaluate whether the chatbot improved students' clinical interview abilities. This narrow approach limited their ability to assess the chatbot's broader educational impact. Many of these reviewed studies that solely rely on perceptual or technical factors did not critically reflect on the limitations of relying on these factors, leaving unanswered questions about the broader educational impact of their chatbots. This could be accomplished by integrating behavioural or psychological evaluation factors, such as knowledge retention and motivation, to provide a more comprehensive assessment of chatbot effectiveness.

Addressing these gaps is important, it is encouraging to see that our trend analysis showed that, despite being the most measured criteria since 2019, there has been a decrease in measuring perceptual factors and a decrease in measuring technical factors. This coincides with an increase in behavioural and psychological factors, reflecting an evolving understanding of chatbot effectiveness beyond perceptual and technical factors. This shift indicates a more holistic approach to evaluating chatbots, recognising the importance of user behaviour and psychological impact in determining overall effectiveness.

4.5. Effectiveness of chatbots in education

Effectiveness in this context refers to the extent to which chatbots achieve their intended objectives, the performance based on automatic metrics, or the user experience of interacting with the chatbot. When examining the effectiveness of chatbots in education, the majority showed a positive effect, some were only partly effective, a minority were ineffective, and a few were unclear due to insufficient explanations in the respective studies. While this finding is new compared to most existing literature reviews, which primarily discuss the tools used to evaluate chatbots rather than their overall effectiveness, we also acknowledge that systematic reviews like ours and meta-analyses can sometimes lead to different results due to their different methodological approaches (Gough et al., 2017).

A systematic review qualitatively analyses a wide range of studies, identifies trends and gaps, and considers factors like user experience, context, and implementation. It allows for a broader, more descriptive overview but may not provide quantifiable effect sizes. In contrast, a meta-analysis focuses on statistically analysing the results of multiple studies to produce an overall effect size, often emphasising quantitative outcomes like learning gains (Gough et al., 2017). While meta-analyses offer the strength of providing a clearer, numerical assessment of effectiveness, they may overlook important qualitative factors, such as user experience, which are important in designing chatbots in education. These differences in approach can lead to varying conclusions. For example, our analysis accounts for perceptual factors like user experience, which are key to the long-term adoption and effectiveness of chatbots in education. Meta-analyses, by contrast, might report lower or different effect sizes because they exclude these factors and focus primarily on measurable outcomes. Thus, a meta-analysis may find that chatbots have only a small or moderate effect on learning outcomes, whereas a systematic review might highlight broader positive impacts. Nevertheless, a recent meta-analysis conducted by Alemdag (2023) found a medium effect ($g = 0.48$) of chatbots on learning outcomes, aligning with our conclusion that chatbots can positively influence education, though this effect may appear more evident when considering both qualitative and quantitative factors.

The evaluation criteria play a crucial role in determining the perceived effectiveness of chatbots. Chatbots assessed solely on perceptual factors generally claim to be effective, indicating a positive user perception. Conversely, most chatbots evaluated on technical factors are also deemed effective, although there was one instance where the technical analysis did not indicate the chatbot's performance. This raises the question of whether perceptual and technical factors alone are sufficient to conclude a chatbot's effectiveness. Many reviewed studies did not explicitly justify their choice of evaluation criteria, making it difficult to determine whether these criteria were well-suited to the chatbots' objectives. This lack of justification, combined with the absence of a standardised evaluation framework, limits the comparability and replicability of findings across studies. Addressing these gaps by integrating both qualitative and quantitative factors into a unified framework is essential for advancing the field. For instance, combining behavioural and perceptual factors could provide insights into how user satisfaction translates into measurable learning gains, offering a more comprehensive evaluation of chatbot effectiveness.

Finally, our trend analysis indicated an increase in the number of effective chatbots over the years and a decrease in the less effective categories. One explanation could be a growing body of research on educational chatbots, which informs the development of new chatbots. However, there is also the possibility of publication bias, where studies reporting effective chatbots are more likely to be published than those reporting ineffective ones, potentially skewing the trend analysis (Joober et al., 2012). If studies reporting less effective chatbots are underrepresented, the observed trend may not fully reflect reality. The potential outliers in the trend analysis are the chatbots classified as ineffective. Upon closer examination, it is notable that these chatbots were ineffective, as for each ineffective chatbot, there is another chatbot in the same technology category and with the same objective that was effective. For example, the AI-based chatbot developed by Oralbayeva et al. (2022) to teach a foreign language was not effective, while a similar AI-based chatbot developed by Chae et al. (2023) was effective. However, differences in the evaluation methods complicate direct comparisons. While Oralbayeva et al. (2022) focused on perceptual and behavioural factors, Chae et al. (2023) emphasised a technical factor. This underscores the need for an integrated approach to evaluating chatbots that accounts for both sets of factors. Even though not all chatbots were effective, this review highlights the potential of educational chatbots to enhance learning experiences. By integrating multiple evaluation criteria, their design and application can be optimised to fully realise these benefits.

4.6. Interaction of the five components: objective, technology, theory, evaluation criteria, and effectiveness

The conducted social network analysis on effective chatbots in education revealed significant insights into the interactions between objectives, technologies, theories, and evaluation criteria. The analysis showed that interactions involving the most prevalent category within each element were the most common. This suggests a current focus on practical, development-driven approaches with a strong emphasis on user perception and educational outcomes. Notably, there was a lack of representation of rule-based chatbots, psychological evaluation factors, and chatbots employing social or personal theories, pointing to potential gaps or underexplored areas in chatbot research and development. This underrepresentation might imply either a perceived lack of importance or challenges in integrating these elements into effective chatbot design.

The prominence of chatbots developed using platforms and the frequent use of perceptual factors for evaluation can be attributed to the ease of development and the convenience of using questionnaires to measure user perception. Educational researchers, often lacking coding expertise, tend to prefer development platforms, making these the most common underlying technology. This practical approach is further reinforced by the ease of deploying perceptual evaluations, which are straightforward and readily interpretable. However, many reviewed studies did not explain why perceptual evaluations were prioritised over psychological or behavioural factors, limiting the depth of their analyses and the understanding of broader educational impacts.

Interestingly, the analysis highlighted a significant gap in the interactions between the theory elements and other components. Despite the general level of effectiveness of the reviewed chatbots, researchers seem to integrate them into educational contexts without a solid theoretical framework. This contradicts the assumption that educational researchers, who typically value theoretical foundations, primarily develop these chatbots. The absence of a standardized approach to chatbot development is evident, suggesting a need for more structured methodologies that incorporate theoretical perspectives. The lack of theoretical grounding raises concerns about whether these chatbots align with educational principles, potentially limiting their ability to achieve meaningful learning outcomes.

The absence of rule-based chatbots from the analysis is particularly notable. Historically significant chatbots, such as the Socratic System (Suppes, 1971) and ELIZA (Weizenbaum, 1966), are rule-based, but this category now appears not to be strongly represented, indicating a shift in chatbot development towards other approaches. However, this shift does not necessarily lean towards deep learning chatbots in educational contexts, as they are also underrepresented in the analysis, but towards platform-based chatbots, possibly indicating the ease of use of these platforms. The AI chatbots identified are primarily teaching-oriented, evaluated by technical factors, and lack a theoretical foundation. This might suggest that AI chatbots are being developed predominantly by computer scientists without substantial input from educational researchers. Another analysis needs to be conducted to evaluate why AI chatbots tend to be teaching-oriented chatbots. One potential reason is that they focus on frequently asked questions on specific domains or topics rather than general university-related questions. Therefore, evaluating these AI chatbots with solely technical and analytical factors might be sufficient.

Aligning these findings with existing literature is difficult because the interactions between objectives, technologies, theories, and evaluation criteria in effective educational chatbots have not been investigated in depth before. While several studies have explored the effectiveness of educational chatbots, user perceptions, and technological developments individually, a comprehensive analysis of how the objectives, technologies, theories, and evaluation criteria in effective educational chatbots interact has not been conducted. Hwang and Chang (2021) partially address this gap through the Technology-based Learning Review (TLR) model, which categorises studies based on learning domains, learning strategies, research design, and analysis methods. In contrast, our study focuses on objectives, technologies, theories, evaluation criteria, and effectiveness. Key distinctions between Hwang and Chang's (2021) approach and ours include the following.

- Categorisation differences: The TLR model emphasises research methodology and learning domains, whereas our study investigates conceptual and functional elements crucial to chatbot implementation.
- Evaluation focus: While Hwang and Chang (2021) review analysis methods such as ANCOVA and t-tests, our study examines specific evaluation criteria used to measure chatbot effectiveness
- Interaction mapping: Hwang and Chang (2021) do not explicitly analyse how their identified categories relate to one another and do not provide a comprehensive overview of their analysis. In contrast, our study provides a structured mapping of the interconnections between objectives, technologies, theories, and evaluation criteria and offers a more comprehensive overview of our analyses (Table 4).

By exploring these interactions, our study offers a pioneering view of the complex relationships in chatbot development and evaluation, laying the foundation for a more integrated and systematic approach to future research.

5. Limitations and implications for future research

The systematic literature review has revealed several limitations that need to be acknowledged. First, the included studies showed heterogeneity. The variations in sample size, educational context, and methodologies pose challenges to generalising the findings. This is shown by the many different interactions between all the components. To enrich future research, a deep dive in specific directions can be taken to find detailed information. Furthermore, future research could also benefit from a more standardised approach to study design and reporting to better facilitate comparisons and meta-analyses. Second, publication bias is another limitation that could have influenced this literature review. The majority of included articles had a positive effect, possibly skewing the perceived effectiveness of chatbots in educational settings. Efforts should be made to include studies that were not effective, providing a more balanced view of evidence. Third, a majority of the included studies focus only on the short-term impact of chatbots, therefore, the long-term benefits and effects on learning and engaging remain underexplored. Therefore, future research should include longitudinal studies to determine whether the positive effects of chatbots are sustained over time and how they influence long-term outcomes.

Fourth, the four categories introduced to classify chatbots based on their underlying technology are not fully conceptually distinct. In particular, hybrid chatbots, by definition, combine elements of both rule-based and AI-based approaches, while platform chatbots may also incorporate one or both of these techniques. The decision to include platform chatbots as a separate category was not based on a technical distinction but rather on a practical limitation: Many studies did not specify whether the chatbots developed using platforms such as Dialogflow or Rasa relied on AI, rule-based logic, or a combination of both. To avoid making assumptions about the underlying implementation, these cases were grouped pragmatically into a separate platform category. While this introduces some conceptual asymmetry into the categorisation scheme, it allows for a more methodologically transparent and rigorous analysis, particularly in light of the variability in reporting quality across the included studies. Fifth, the evaluation criteria varied significantly across studies, making it difficult to assess the effectiveness of chatbots, especially since approximately 50 per cent only measure perceptual and/or technical factors. This variability complicates finding what contributes to an effective chatbot. Future research should aim to standardize the evaluation criteria for specific objectives, enabling a more consistent and comparable assessment of a chatbot's effectiveness.

Sixth, this review focused on recent literature, coinciding with the introduction of advanced models such as GPT-2. However, this field is ever-changing, especially with the launch of ChatGPT, other foundational models, and other commercial chatbots. Chatbots based on ChatGPT were excluded from this review. ChatGPT, being relatively new, has not yet undergone extensive long-term evaluation in educational settings. Its inclusion could skew findings, given its current prominence and rapid adoption. Additionally, limited information is available about the technical aspects of ChatGPT, aligning with another exclusion criterion. Furthermore, ethical concerns surrounding its use, such as data privacy, bias, and potential for misuse, are still being debated and addressed (Stahl & Eke, 2024). Additionally, commercial chatbots could have been excluded as well, as they might not have complied with our inclusion and exclusion criteria since information on technical aspects was often lacking. Therefore, as more robust studies on commercial chatbots and ChatGPT become available (Banihashem et al., 2024), future reviews will be able to incorporate these insights, providing a more comprehensive understanding of their impact on education. Therefore, when more information is available and researchers reach a consensus, future research can include studies based on ChatGPT to gain foundational insights into its impact on education (e.g., Banihashem et al., 2024).

Finally, systematic reviews themselves have limitations that must be acknowledged. While they are effective at analysing existing studies and identifying trends, they are constrained by the scope and quality of the available literature, and the inclusion and exclusion criteria used to select studies (Gough et al., 2017). This can make it difficult to fully explore alternative interpretations of the data, as systematic reviews depend on the included studies' theoretical and methodological approaches. Future research could consider meta-analyses or a mixed-methods approach, incorporating both quantitative and qualitative data, to provide alternative interpretations and offer more diverse perspectives on the effectiveness of chatbots in education.

The results of this systematic literature review also revealed some implications for educational practices. First, a notable gap identified in this review is the limited integration of educational theories in the design and implementation of chatbots. Many studies develop chatbots based on practical needs without grounding them in robust theoretical frameworks. Therefore, future studies should strive to incorporate established educational theories to enhance the pedagogical soundness of chatbot implementations. For instance, cognitive load theory (Sweller et al., 2011) can inform the timing and structure of feedback provided by chatbots. Chatbots grounded in cognitive load theory could provide timely support when students struggle, preventing cognitive overload and improving retention by offering step-by-step guidance. For example, in a mathematics task, the chatbot might detect repeated incorrect answers and offer step-by-step guidance when needed. In a language learning task, the chatbot could prompt the student to correct grammatical errors in sentences based on the frequency of mistakes. This approach aligns with research on effective feedback (Hattie & Timperley, 2007) and can optimise student engagement by ensuring that feedback is constructive, specific, and provided at the optimal time. For educators and practitioners, this means that they should prioritise selecting chatbots that are not only functional but also grounded in relevant educational theories, ensuring that the technology actively supports learning by delivering timely, theory-based interventions tailored to students' needs.

This review highlights the need for improved evaluation methods and reveals that many publications evaluated their chatbots solely on perceptual and technical factors, such as usability and accuracy. These factors are not necessarily sufficient to measure the effectiveness of chatbots. Therefore, future studies should also strive to measure the effectiveness of chatbots based on behavioural and/or psychological factors. For instance, chatbots could integrate self-report surveys that prompt students to reflect on their progress following each learning task, aligning with motivational theories (Deci & Ryan, 1985). This could be done by letting the chatbot ask reflective questions after each completed module in a course, helping educators see how students perceive their progress, motivation, or areas of difficulty. Additionally, behavioural data, such as the frequency of chatbot interactions or requests for hints, could provide valuable insights into students' persistence and engagement. Aligning behavioural and psychological evaluation strategies, such as questionnaires and interaction data analysis, with educational theories allows educators to better assess a chatbot's impact on learning processes and outcomes. This focus on psychological outcomes can directly improve student engagement and long-term academic success.

This review demonstrates the existence of various methods and approaches to develop chatbots. It also highlights the availability of a multitude of different development or chatbot builder platforms that facilitate the process. Future studies should select the development platform not solely based on technical capabilities but rather on how well it aligns with educational objectives. For example, a chatbot for collaborative learning in high school could use a platform that supports peer feedback and group discussions and integrates with existing learning management systems or similar tools used by the educational institution. In contrast, platforms with AI features may suit chatbots for personalised learning in large university courses, as they can deliver tailored feedback and adaptive learning paths. By aligning platform selection with educational goals, educators can ensure that the technology meaningfully supports student learning outcomes, whether by fostering collaboration or enhancing personalised instruction.

Finally, we present a comprehensive example that incorporates all of the aforementioned implications: An educator wants to develop a chatbot designed to support students during group work. Such a chatbot could facilitate group discussions by prompting students to share ideas, ask questions, and provide peer feedback, thereby promoting collaborative problem-solving skills. Social learning theories, particularly those related to Computer-Supported Collaborative Learning (CSCL) (e.g., Stahl et al., 2006), could inform the design of the chatbot to ensure it effectively facilitates group interactions and enhances collaborative learning among students (first implication). Additionally, the chatbot could provide real-time feedback when students collaborate effectively or offer suggestions if a group appears to be struggling. For instance, if the chatbot detects low engagement or frequent misunderstandings within a group, it could alert the educator, who might intervene by directly facilitating the group discussion or adjusting the task to offer clearer instructions. Educators could also use this data to tailor future interventions, either by modifying the chatbot's responses or by providing targeted feedback to specific students during subsequent activities. This dual approach, combining automated feedback with human guidance, helps ensure that both immediate and long-term learning objectives are met. By applying these theories and interventions, educators can optimise the chatbot's ability to facilitate meaningful peer interactions and enhance students' collaborative problem-solving skills.

To implement this, educators or practitioners should consult relevant research (e.g., in the case of collaborative learning interventions, Michos et al., 2020) to evaluate the technologies utilised in previous studies. Platform selection should focus on aligning technology with educational goals and ensuring seamless integration with institutional systems, as discussed in the third implication.

To assess the effectiveness of the chatbot, educators should move beyond evaluating students' perceptions of its usability. Instead, they could analyse chatlogs and administer questionnaires to determine whether the chatbot is successfully improving collaborative learning behaviours and whether students are developing better collaborative skills (second implication). For example, educators might analyse the frequency of meaningful peer interactions triggered by the chatbot or assess the quality of peer feedback over time. This method provides a more comprehensive understanding of how the chatbot impacts both the processes and outcomes of group work, offering insights into its educational value.

By acknowledging and addressing these limitations and implications, future research can build on the findings of this review to advance the field of educational chatbots, ensuring more reliable, valid, and comprehensive insights into their use and effectiveness in various educational contexts.

6. Conclusion

This review identified five key components of educational chatbots - objectives, technology, theory, evaluation criteria, and effectiveness - and analysed their interactions to provide a comprehensive overview of their role in overall effectiveness. This study enhances our understanding of chatbots in education and offers valuable insight and guidance for educators, practitioners, and researchers in the field.

From a practical standpoint, by understanding the objective of the chatbot, the available development methods, how to theoretically ground the chatbot's integration in education, and how to evaluate the chatbot accordingly, educators and practitioners can better optimise its use in their educational practices. As the focus is shifting towards chatbots utilising foundational models, this review shows accessible alternatives in the form of development platforms. While development platforms offer accessible solutions, foundational models can complement these platforms by enabling nuanced conversational interactions and personalised learning experiences. By aligning these elements and leveraging the insights from this study, educators can make informed pedagogical decisions, design more effective chatbots, and enhance user satisfaction.

From a research perspective, our findings deepen the understanding of the role of the underlying theory in applying chatbots to education. Building upon our findings, we suggest that chatbots should be integrated more deeply into pedagogical frameworks to ensure they are not merely add-ons but play an integral role in enhancing their pedagogical effectiveness. Researchers should investigate the long-term impacts of such integrations and develop best practices for implementing chatbots within various educational contexts. Additionally, our comprehensive overview highlights several areas that are underexplored or overlooked, indicating the need for further research. This review provides a clear guideline for future research and practical applications, ensuring that chatbots contribute significantly to educational outcomes.

CRediT authorship contribution statement

Tim Debets: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Seyyed Kazem Banihashem:** Writing – review & editing, Writing – original draft, Conceptualization. **Desirée Joosten-Ten Brinke:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Tanja E.J. Vos:** Writing – review & editing, Conceptualization. **Gideon Maillette de Buy Wenniger:** Conceptualization. **Gino Camp:** Writing – review & editing, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve the readability and language of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors are not aware of any potential conflicts of interest.

Appendix

Table A.1
General Characteristics of Included Studies

Authors and year	Journal/Conference Name	Context	Country (first author)	Domain	Methodology
Abdelhamid and Katz (2020)	Engineering Experience Proceedings	Higher education (N = 53)	USA	Engineering	Quantitative
Adair et al. (2023)	Artificial Intelligence in Education	Secondary education (N = 70)	USA	Science	Quantitative
Al-Abdullatif et al. (2023)	Frontiers in Psychology	Higher education (N = 60)	Saudi Arabia	Behavioural Sciences (Motivation & Learning strategies)	Quantitative
Ali et al. (2022)	International Conference on Agents	Higher education (N = 50)	Pakistan	Admission	Quantitative
Aljameel et al. (2019)	Advances in Intelligent Systems and Computing	Primary & Secondary education (N = 24)	Saudi Arabia	Science	Quantitative
Aloqayli and Abdelhafez (2023)	International Journal of Information and Education Technology	Higher education (N = 22)	Saudi Arabia	Admission	Quantitative
Alqahtani and Alrwais (2023)	International Journal of Advanced Computer Science and Applications	Higher education (N = 0)	Saudi Arabia	LMS (Blackboard)	Quantitative
Aujogue and Aussem (2019)	International Joint Conference on Neural Networks	Higher education (N = 0)	France	Admission	Quantitative
Cai et al. (2021)	Machine Learning	Online education (N = 1063)	USA	Math	Qualitative
Chae et al. (2023)	Conference on Artificial Intelligence	Higher education (N = 0)	South-Korea	Language (English)	Quantitative
Chen et al. (2021)	Journal of Pacific Rim Psychology	Primary & Secondary education (N = 628)	China	Moral education (Need deficiencies)	Quantitative
Chen et al. (2023)	Information Systems Frontiers	Higher education (N = 410)	USA	1) Computer Science (AI) 2) interviewing	Mixed-method
Deveci Topal et al. (2021)	Education and Information Technologies	Secondary education (N = 41)	Turkey	Science	Mixed-method
Draxler et al. (2022)	Designing Interactive Systems Conference	Higher education (N = 72)	Germany	Behavioural Sciences (Students' well-being)	Qualitative
Durall and Kapros (2020)	HCI International Conference	Secondary & Higher education (N = 32)	Finland	Core competencies	Mixed-method
El Hefny et al. (2021)	International Conference on Human-Agent Interaction	Higher education (N = 40)	Egypt	Course support	Quantitative
Essel et al. (2022)	International Journal of Educational Technology in Higher Education	Higher education (N = 68)	Ghana	Computer Science (Multimedia programming)	Mixed-method
Fidan and Gencel (2022)	Journal of Educational Computing Research	Higher education (N = 144)	Turkey	Education (Instructional technologies)	Mixed-method

(continued on next page)

Table A.1 (*continued*)

Authors and year	Journal/Conference Name	Context	Country (first author)	Domain	Methodology
Giler et al. (2023)	Trends in Artificial Intelligence and Computer Engineering	Higher education (N = 134)	Ecuador	Faculty Information	Quantitative
Gonçalves et al. (2022)	International Journal of Innovation	Higher education (8309 conversations)	Brazil	University Information	Quantitative
González et al. (2022)	Computer Applications in Engineering Education	Higher education (N = 311)	Chile	Computer Science (Software Engineering)	Mixed-method
Han et al. (2022)	BMC Medical Education	Higher education (N = 61)	South-Korea	Healthcare (Electronic fetal monitoring)	Mixed-method
Hew et al. (2023)	Journal of Computing in Higher Education	Higher education (N = 67)	China	1) Behavioural Sciences (Goal setting) 2) Language	Mixed-method
Hirose et al. (2021)	International Conference on Agents	Higher education (N = 27)	Japan	Behavioural Sciences (Learning strategies)	Quantitative
Hsu et al. (2023)	Educational Technology & Society	Secondary education (N = 70)	Taiwan	Geography	Quantitative
Iku-Silan et al. (2023)	Computers and Education	Secondary education (N = 71)	Taiwan	Healthcare (Indigenous traditional medicine)	Mixed-method
Jariwala et al. (2021)	HCI International Conference	Secondary education (N = 0)	USA	Math	Quantitative
Khalil and Rambech (2022)	HCI International Conference	Higher education (N = 42)	Norway	Course support	Mixed-method
Khin and Soe (2020)	International Conference on Speech Database and Assessments	Higher education (N = 0)	Myanmar	University Information	Quantitative
Kohnke (2023)	International Journal of Mobile Learning and Organisation	Higher education (N = 128)	China	Language (English)	Mixed-method
Kumar (2021)	International Journal of Educational Technology in Higher Education	Higher education (N = 60)	Malaysia	Group work support	Mixed-method
Lam et al. (2023)	IEEE Access	Higher education (N = 0)	Vietnam	Multi-domain	Quantitative
Lee and Yeo (2022)	Computers and Education	Higher education (N = 23)	USA	Math	Mixed-method
Leonardi and Torchiano (2023)	Methodologies and Intelligent Systems for Technology-Enhanced Learning	Higher education (N = 0)	Italy	Computer Science (Object-oriented programming)	Quantitative
Liaw et al. (2023a)	Journal of Medical Internet Research	Higher education (N = 64)	Singapore	Healthcare (Sepsis care & interprofessional communication)	Mixed-method
Liaw et al. (2023b)	Nurse Education Today	Higher education (N = 32)	Singapore	Healthcare (Interprofessional communication)	Mixed-method
Lin and Ye (2023)	Journal of Internet Technology	Secondary education (N = 34)	Taiwan	Biology	Quantitative
Mageira et al. (2022)	Applied Sciences	Secondary education (N = 61)	Greece	Language & Culture (Content and language integrated learning)	Mixed-method
Mai et al. (2022)	International Conference on Human-Computer Interaction	Higher & Secondary education (N = 182)	Germany	Exam anxiety	Mixed-method
Mamani et al. (2019)	Conference on Electronics, Electrical Engineering and Computing	Higher education (N = 50)	Peru	Academic information	Quantitative
Martha et al. (2023)	Transactions on Learning Technologies	Higher education (N = 92)	Indonesia	Behavioural Sciences (Self- and co-regulation skills)	Mixed-method
Memon et al. (2021)	Scientific Programming	Several levels of education (N = 0)	Pakistan	Multi-domain	Quantitative
Mendoza et al. (2020)	HCI International Conference	Secondary education (N = 8)	Mexico	Student Support	Quantitative
Michos et al. (2020)	Conference on Technology Enhanced Learning	Higher education (N = 54)	Spain	Computer-supported collaborative learning	Mixed-method
Mohamed et al. (2021)	Conference on Smart City Applications	Several levels of education (N = 0)	Morocco	Multi-domain	Quantitative
Mokmin and Ibrahim (2021)	Education and Information Technologies	Higher education (N = 75)	Malaysia	Healthcare (Health literacy)	Mixed-method
Mzwri and Turcsányi-Szabo (2023)	Applied Sciences	Higher education (N = 35)	Hungary	Multi-domain	Quantitative
Nasharuddin et al. (2021)	International Conference on Information Retrieval and Knowledge Management	Higher education (N = 89)	Malaysia	Computer Science (AI)	Mixed-method
Nguyen et al. (2021)	Sensors and Materials	Higher education (N = 0)	Vietnam	Computer science	Quantitative

(*continued on next page*)

Table A.1 (*continued*)

Authors and year	Journal/Conference Name	Context	Country (first author)	Domain	Methodology
Nguyen et al. (2022)	Communications in Computer and Information Science	Higher education (N = 27)	Vietnam	Computer Science	Quantitative
Okonkwo and Ade-Ibijola (2021b)	Engineering Letters	Higher education (N = 205)	South-Africa	Computer Science	Quantitative
Oralbayeva et al. (2022)	International Conference on Human-Robot Interaction	Higher education (N = 20)	Kazakhstan	Language (Kazakh Latin)	Quantitative
Palasundram et al. (2019)	International Journal of Emerging Technologies in Learning	Higher education (N = 0)	Malaysia	Computer science	Quantitative
Paschoal et al. (2023)	Informatics in Education	Higher education (N = 38)	Brazil	Computer Science (Software Testing)	Quantitative
Pereira et al. (2023)	Medical Education Online	Higher education (N = 13)	Portugal	Healthcare (Clinical interview)	Quantitative
Raiche et al. (2023)	Frontiers in Psychology	Higher education (N = 112)	Canda	Behavioural Sciences (Risk assessment)	Quantitative
Ruan et al. (2019)	Conference on Human Factors in Computing Systems Proceedings	Higher education (N = 112)	USA	Science, safety and English Vocabulary	Mixed-method
Sáiz-Manzanares et al. (2023)	Heliyon	Higher education (N = 57)	Spain	Health Science	Mixed-method
Salas-Velasco (2023)	Journal of Economic Education	Higher education (N = 538)	Spain	Economics	Mixed-method
Salazar (2023)	International Journal of Engineering Trends and Technology	Higher education (N = 46)	Philippines	Healthcare (Pharmacology)	Quantitative
Schmitt et al. (2022)	Proceedings of the ACM on Human-Computer Interaction	Higher education (N = 95)	Switzerland	Course support	Mixed-method
Sharma et al. (2023)	Journal of Maritime Affairs	Higher education (N = 18)	Norway	Science	Quantitative
Tanana et al. (2019)	Journal of Medical Internet Research	Higher education (N = 151)	USA	Behavioural sciences (Interview and counselling skills)	Quantitative
Vázquez-Cano et al. (2021)	International Journal of Educational Technology in Higher Education	Higher education (N = 103)	Spain	Language (Spanish)	Mixed-method
Wambsganss et al. (2020)	International Conference on Wirtschaftsinformatik	Higher education (N = 44)	Germany	Course evaluation	Mixed-method
Wambsganss et al. (2022)	Proceedings of the ACM on Human-Computer Interaction	Higher education (N = 208)	Germany	Course evaluation	Quantitative
Wijayawardena et al. (2022)	International Conference for Convergence in Technology	Secondary education (N = 0)	Sri Lanka	Multi-domain	Quantitative
Winkler et al. (2020)	Conference on Human Factors in Computing Systems	Higher education (N = 182)	Switzerland	Computer science	Mixed-method
Winkler et al. (2021)	Computers and Education	Secondary education (N = 44)	Switzerland	Law, morality and ethics	Mixed-method
Xie et al. (2021)	Transactions on Learning Technologies	Higher education (N = 40)	China	Computer-supported collaborative learning	Quantitative
Xu et al. (2022)	Conference on Human Factors in Computing Systems	Kindergarten (N = 20)	USA	Science	Mixed-method

Data availability

Data will be made available on request.

References

Abdelhamid, S., & Katz, A. (2020). Using chatbots as smart teaching assistants for first-year engineering students. In *2020 first-year engineering experience proceedings*. https://doi.org/10.18260/1-2–35782

Adair, A., Pedro, M. S., Gobert, J., & Segan, E. (2023). Real-time AI-driven assessment and scaffolding that improves students' mathematical modeling during science investigations. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 202–216). Cham: Springer Nature Switzerland.

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial intelligence applications and innovations* (pp. 373–383). Cham: Springer International Publishing.

AERA. (2006). Standards for reporting on empirical social science research in AERA publications. American educational research association. *Educational Researcher, 35*(6), 33–40.

Agarwal, R., & Wadhwa, M. (2020). Review of state-of-the-art design techniques for chatbots. *SN Computer Science*. https://doi.org/10.1007/s42979-020-00255-3

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*. https://doi.org/10.3389/fpsyg.2023.1129070

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education, 3*, 60. https://doi.org/10.1007/s44217-024-00138-2

Alemdag, E. (2023). The effect of chatbots on learning: A meta-analysis of empirical research. *Journal of Research on Technology in Education*, 1–23. https://doi.org/10.1080/15391523.2023.2255698

Ali, M., Azam, F., Safdar, A., & Anwar, M. (2022). Intelligent agents in educational institutions: NEdBOT - NLP-based chatbot for administrative support using DialogFlow. In *2022 IEEE international conference on agents (ICA)* (pp. 30–35). https://doi.org/10.1109/ICA55837.2022.00012

Aljameel, S., O'Shea, J., Crockett, K., Latham, A., & Kaleem, M. (2019). LANA-I: An Arabic conversational intelligent tutoring system for children with ASD. In K. Arai, R. Bhatia, & S. Kapoor (Eds.), *Intelligent computing* (pp. 498–516). Cham: Springer International Publishing.

Aloqayli, A., & Abdelhafez, H. (2023). Intelligent chatbot for admission in higher education. *International Journal of Information and Education Technology*. https://doi.org/10.18178/ijiet.2023.13.9.1937

Alqahtani, Q., & Alrwais, O. (2023). Building a machine learning powered chatbot for KSU Blackboard users. *International Journal of Advanced Computer Science and Applications, 14*(2). https://doi.org/10.14569/IJACSA.2023.0140290

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*, 167–207. https://doi.org/10.1207/s15327809jls0402_2

Aujogue, J.-B., & Aussem, A. (2019). Hierarchical recurrent attention networks for context-aware education chatbots. In *2019 international joint conference on neural networks (IJCNN)* (pp. 1–8). https://doi.org/10.1109/IJCNN.2019.8852445

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Banihashem, K., & Macfadyen, L. P. (2021). Pedagogical design: Bridging learning theory and learning analytics. *Canadian Journal of Learning and Technology, 47*(1).

Cai, W., Grossman, J., Lin, Z. J., Sheng, H., Wei, J. T.-Z., Williams, J. J., & Goel, S. (2021). Bandit algorithms to personalize educational chatbots. *Machine Learning, 110*(9), 2389–2418. https://doi.org/10.1007/s10994-021-05983-y

Chae, H., Kim, M., Kim, C., Jeong, W., Kim, H., Lee, J., & Yeo, J. (2023). TUTORING: Instruction-grounded conversational agent for language learners. *arXiv [Cs.AI]*. Retrieved from http://arxiv.org/abs/2302.12623.

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers, 25*(1), 161–182. https://doi.org/10.1007/s10796-022-10291-4

Chen, P., Lu, Y., Yu, S., Xu, Q., & Liu, J. (2021). A dialogue system for identifying need deficiencies in moral education. *Journal of Pacific Rim Psychology, 15*, Article 1834490921998589. https://doi.org/10.1177/1834490921998589

Chen, X., Xie, H., Zou, D., & Hwang, G.-J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers & Education: Artificial Intelligence, 1*, Article 100002. https://doi.org/10.1016/j.caeai.2020.100002

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers & Education: Artificial Intelligence, 4*, Article 100118. https://doi.org/10.1016/j.caeai.2022.100118

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer. https://doi.org/10.1007/978-1-4899-2271-7

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems, 19*(4), 9–30. https://doi.org/10.1080/07421222.2003.11045748

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining 'gamification'. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future Media environments* (pp. 9–15). Presented at the Tampere, Finland. https://doi.org/10.1145/2181037.2181040.

Deveci Topal, A., Dilek Eren, C., & Kolburan Geçer, A. (2021). Chatbot application in a 5th grade science course. *Education and Information Technologies, 26*(5), 6241–6265. https://doi.org/10.1007/s10639-021-10627-8c

DialogFlow. (n.d.). Retrieved June 18, 2024,from https://cloud.google.com/dialogflow?hl=nl.

Draxler, F., Hirsch, L., Li, J., Oechsner, C., Völkel, S. T., & Butz, A. (2022). Flexibility and social disconnectedness: Assessing university students' well-being using an experience sampling chatbot and surveys over two years of COVID-19. In *Proceedings of the 2022 ACM designing interactive systems conference* (pp. 217–231). https://doi.org/10.1145/3532106.3533537

Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers & Education: Artificial Intelligence, 6*, Article 100230. https://doi.org/10.1016/j.caeai.2024.100230

Durall, E., & Kapros, E. (2020). Co-Design for a competency self-assessment chatbot and survey in science education. In P. Zaphiris, & A. Ioannou (Eds.), *Learning and collaboration technologies. Human and technology ecosystems* (pp. 13–24). Cham: Springer International Publishing.

El Hefny, W., El Bolock, A., Herbert, C., & Abdennadher, S. (2021). Applying the character-based chatbots generation framework in education and healthcare. In *Proceedings of the 9th international conference on human-agent interaction* (pp. 121–129). https://doi.org/10.1145/3472307.3484172

Ertmer, P. A., & Newby, T. J. (2013). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance Improvement Quarterly, 26*(2), 43–71.

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), 57. https://doi.org/10.1186/s41239-022-00362-6

Fidan, M., & Gencel, N. (2022). Supporting the instructional videos with chatbot and peer feedback mechanisms in online learning: The effects on learning performance and intrinsic motivation. *Journal of Educational Computing Research, 60*(7), 1716–1741. https://doi.org/10.1177/07356331221077901

Gao, X., Noroozi, O., Gulikers, J., Biemans, H. J. A., & Banihashem, S. K. (2024). A systematic review of the key components of online peer feedback practices in higher education. *Educational Research Review, 42*, Article 100588. https://doi.org/10.1016/j.edurev.2023.100588

Gerritsen-van Leeuwenkamp, K. J., Joosten-ten Brinke, D., & Kester, L. (2017). Assessment quality in tertiary education: An integrative literature review. *Studies In Educational Evaluation, 55*, 94–116. https://doi.org/10.1016/j.stueduc.2017.08.001

Giler, M., Cedeño, E., Zambrano, W., Zambrano, M., & Zambrano, D. (2023). Chatbots and its impact on the information support service for students of the faculty of computer science of the technical university of manabí. *Trends in artificial intelligence and computer engineering*. Cham: Springer Nature Switzerland.

Goel, A. K., & Polepeddi, L. (2018). Jill Watson. *Learning Engineering for Online Education*. https://doi.org/10.4324/9781351186193-7

Gonçalves, G. S., Ribeiro, T., de, L. S., Teixeira, J. E. V., & Costa, B. K. (2022). The deployment of chatbot to improve customer service in higher education institutions during COVID-19. *International Journal of Innovation, 10*(1), 178–203. https://doi.org/10.5585/iji.v10i1.20652

González, L. A., Neyem, A., Contreras-McKay, I., & Molina, D. (2022). Improving learning experiences in software engineering capstone courses using artificial intelligence virtual assistants. *Computer Applications in Engineering Education, 30*(5), 1370–1389. https://doi.org/10.1002/cae.22526

Google Cloud. (n.d.). Generators. Dialogflow CX. Retrieved January 21, 2025, from https://cloud.google.com/dialogflow/cx/docs/concept/generators.

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews* (2nd ed.). SAGE Publications.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialog. *IEEE Transactions on Education, 48*, 612–618. https://doi.org/10.1109/TE.2005.856149

Han, J.-W., Park, J., & Lee, H. (2022). Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: A quasi-experimental study. *BMC Medical Education, 22*(1), 830. https://doi.org/10.1186/s12909-022-03898-3

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hew, K. F., Huang, W., Du, J., & Jia, C. (2023). Using chatbots to support student goal setting and social presence in fully online activities: Learner engagement and perceptions. *Journal of Computing in Higher Education, 35*(1), 40–68. https://doi.org/10.1007/s12528-022-09338-x

Hirose, N., Shiramatsu, S., & Okuhara, S. (2021). Development of chatbot to support student learning strategies in design education. In *2021 IEEE international conference on agents (ICA)* (pp. 1–6). https://doi.org/10.1109/ICA54137.2021.00007

Hobert, S. (2019). How are you, chatbot? Evaluating chatbots in educational settings - results of a literature review. *Fachtagung 'e-Learning' Der Gesellschaft Für Informatik*. https://doi.org/10.18420/delfi2019_289. Retrieved from.

Hobert, S., & von Wolff, R. M. (2019). *Say hello to your new automated tutor - a structured literature review on pedagogical conversational agents*. Wirtschaftsinformatik. Retrieved from https://api.semanticscholar.org/CorpusID:201114924.

Hsu, T. C., Huang, H. L., Hwang, G. J., & Chen, M. S. (2023). Effects of incorporating an expert decision-making mechanism into chatbots on students' achievement, enjoyment, and anxiety. *Educational Technology & Society, 26*(1), 218–231. https://doi.org/10.30191/ETS.202301_26(1).0016

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning, 38*(1), 237–257. https://doi.org/10.1111/jcal.12610

Hwang, G. J., & Chang, C. Y. (2021). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments, 31*(7), 4099–4112. https://doi.org/10.1080/10494820.2021.1952615

Iku-Silan, A., Hwang, G.-J., & Chen, C.-H. (2023). Decision-guided chatbots and cognitive styles in interdisciplinary learning. *Computers & Education, 201*, Article 104812. https://doi.org/10.1016/j.compedu.2023.104812

Jariwala, A., Marghitu, D., & Chapman, R. (2021). A multimodal platform to teach mathematics to students with vision-impairment. In M. Antona, & C. Stephanidis (Eds.), *Universal access in human-computer interaction. Access to Media, learning and assistive environments* (pp. 109–117). Cham: Springer International Publishing.

Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. *Computers & Education, 206*, Article 104898. https://doi.org/10.1016/j.compedu.2023.104898

Jeon, J., Lee, S., & Choi, S. (2023). A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interactive Learning Environments, 32*, 4613–4631. https://doi.org/10.1080/10494820.2023.2204343

Jonassen, D. H. (1991). Objectivism versus constructivism: Do we need a new philosophical paradigm? *Educational Technology Research & Development, 39*(3), 5–14. https://doi.org/10.1007/BF02296434

Joober, R., Schmitz, N., Annable, L., & Boksa, P. (2012). Publication bias: What are the challenges and can they be overcome? *Journal of Psychiatry & Neuroscience: JPN, 37*(3), 149–152. https://doi.org/10.1503/jpn.120065

Khalil, M., & Rambech, M. (2022). Eduino: A Telegram learning-based platform and chatbot in higher education. In P. Zaphiris, & A. Ioannou (Eds.), *Learning and collaboration technologies. Novel technological environments* (pp. 188–204). Cham: Springer International Publishing.

Khin, N. N., & Soe, K. M. (2020). Question answering based university chatbot using sequence to sequence model. In *23rd conference of the oriental COCOSDA international committee for the Co-ordination and standardisation of speech databases and assessment techniques (O-COCOSDA)* (pp. 55–59). https://doi.org/10.1109/O-COCOSDA50338.2020.9295021

Kohnke, L. (2023). L2 learners' perceptions of a chatbot as a potential independent language learning tool. *International Journal of Mobile Learning and Organisation, 17* (1–2), 214–226. https://doi.org/10.1504/ijmlo.2023.128339

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*, 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Kumar, J. A. (2021). Educational chatbots for project-based learning: Investigating learning outcomes for a team-based design course. *International Journal of Educational Technology in Higher Education, 18*, 65. https://doi.org/10.1186/s41239-021-00302-w

Lai, W. Y. W., & Lee, J. S. (2024). A systematic review of conversational AI tools in ELT: Publication trends, tools, research methods, learning outcomes, and antecedents. *Computers & Education: Artificial Intelligence, 7*, Article 100291. https://doi.org/10.1016/j.caeai.2024.100291

Lam, K. N., Nguy, L. H., Le, V. L., & Kalita, J. (2023). A transformer-based educational virtual assistant using diacriticized Latin script. *IEEE Access, 11*, 90094–90104. https://doi.org/10.1109/ACCESS.2023.3307635

Laurillard, D. (2002). *Rethinking university teaching: A conversational framework for the effective use of learning technologies* (2nd ed.). Routledge. https://doi.org/10.4324/9781315012940

Lee, D., & Yeo, S. (2022). Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Computers & Education, 191*, Article 104646. https://doi.org/10.1016/j.compedu.2022.104646

Leonardi, S., & Torchiano, M. (2023). Educational chatbot to support question answering on Slack. In M. Temperini, V. Scarano, I. Marenzi, M. Kravcik, E. Popescu, R. Lanzilotti, & P. Vittorini (Eds.), *Methodologies and intelligent systems for technology enhanced learning, 12th international conference* (pp. 20–25). Cham: Springer International Publishing.

Liaw, S. Y., Tan, J. Z., Bin Rusli, K. D., Ratan, R., Zhou, W., Lim, S., … Chua, W. L. (2023a). Artificial intelligence versus human-controlled doctor in virtual reality simulation for sepsis team training: Randomized controlled study. *Journal of Medical Internet Research, 25*, Article e47748. https://doi.org/10.2196/47748

Liaw, S. Y., Tan, J. Z., Lim, S., Zhou, W., Yap, J., Ratan, R., … Chua, W. L. (2023b). Artificial intelligence in virtual reality simulation for interprofessional communication training: Mixed method study. *Nurse Education Today, 122*, Article 105718. https://doi.org/10.1016/j.nedt.2023.105718

Lin, Y.-T., & Ye, J.-H. (2023). Development of an educational chatbot system for enhancing students' biology learning performance. *Journal of Internet Technology, 24* (2), 275–281. Retrieved from https://jit.ndhu.edu.tw/article/view/2867.

Lo, C. K., Hew, K. F., & Jong, M. S.-Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education, 219*, Article 105100. https://doi.org/10.1016/j.compedu.2024.105100

Mageira, K., Pittou, D., Papasalouros, A., Kotis, K., Zangogianni, P., & Daradoumis, A. (2022). Educational AI chatbots for content and language integrated learning. *Applied Sciences, 12*(7). https://doi.org/10.3390/app12073239

Mai, V., Bauer, A., Deggelmann, C., Neef, C., & Richert, A. (2022). AI-based coaching: Impact of a chatbot's disclosure behavior on the working alliance and acceptance. In J. Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *HCI international 2022 – late breaking papers: Interacting with eXtended reality and artificial intelligence* (pp. 391–406). Cham: Springer Nature Switzerland.

Malhotra, N. K. (2006). *Questionnaire design and scale development*. https://doi.org/10.4135/9781412973380.n5

Mamani, J. R. C., Álamo, Y. J. R., Aguirre, J. A. A., & Toledo, E. E. G. (2019). Cognitive services to improve user experience in searching for academic information based on chatbot. In *2019 IEEE XXVI international conference on electronics, electrical engineering and computing (INTERCON)* (pp. 1–4). https://doi.org/10.1109/INTERCON.2019.8853572

Martha, A. S. D., & Santoso, H. B. (2019). The design and impact of the pedagogical agent: A systematic literature review. *The Journal of Educators Online*. https://doi.org/10.9743/JEO.2019.16.1.8. Retrieved from.

Martha, A. S. D., Santoso, H. B., Junus, K., & Suhartanto, H. (2023). The effect of the integration of metacognitive and motivation scaffolding through a pedagogical agent on self- and co-regulation learning. *IEEE Transactions on Learning Technologies, 16*(4), 1–12. https://doi.org/10.1109/TLT.2023.3266439

Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press.

Mckie, I. A. S., & Narayan, B. (2019). Enhancing the academic library experience with chatbots: An exploration of research and implications for practice. *Journal of the Australian Library and Information Association, 68*, 268–277. https://doi.org/10.1080/24750158.2019.1611694

Memon, Z., Aghian, H., Sarfraz, M. S., Hussain Jalbani, A., Oskouei, R. J., Jalbani, K. B., & Hussain Jalbani, G. (2021). Framework for educational domain-based multichatbot communication system. *Scientific Programming, 2021*, Article 5518309. https://doi.org/10.1155/2021/5518309

Mendoza, S., Hernández-León, M., Sánchez-Adame, L. M., Rodríguez, J., Decouchant, D., & Meneses-Viveros, A. (2020). Supporting student-teacher interaction through a chatbot. In P. Zaphiris, & A. Ioannou (Eds.), *Learning and collaboration technologies. Human and technology ecosystems* (pp. 93–107). Cham: Springer International Publishing.

Michos, K., Asensio-Pérez, J. I., Dimitriadis, Y., García-Sastre, S., Villagrá-Sobrino, S., Ortega-Arranz, A., … Topali, P. (2020). Design of conversational agents for CSCL: Comparing two types of agent intervention strategies in a university classroom. In C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, & S. M. Dennerlein (Eds.), *Addressing global challenges and quality education* (pp. 215–229). Cham: Springer International Publishing.

Mohamad Noor, N. (2023). The impact of educational technology on distance learning in the era of post-COVID-19. *International Journal on E-Learning Practices (IJELP), 6*(1). https://doi.org/10.51200/ijelp.v6i1.4301

Mohamed, B. A., Abdelhakim, B. A., & Youness, S. (2021). A deep learning model for an intelligent chat bot system: An application to E-learning domain. In M. Ben Ahmed, İ. Rakıp Karaş, D. Santos, O. Sergeyeva, & A. A. Boudhir (Eds.), *Innovations in smart cities applications* (Vol. 4, pp. 165–179). Cham: Springer International Publishing.

Mokmin, N. A. M., & Ibrahim, N. A. (2021). The evaluation of chatbot as a tool for health literacy education among undergraduate students. *Education and Information Technologies, 26*(5), 6033–6049. https://doi.org/10.1007/s10639-021-10542-y

Mzwri, K., & Turcsányi-Szabo, M. (2023). Internet Wizard for enhancing open-domain question-answering chatbot knowledge base in education. *Applied Sciences, 13*. https://doi.org/10.3390/app13148114

Nasharuddin, N. A., Sharef, N. M., Mansor, E. I., Samian, N., Murad, M. A. A., Omar, M. K., … Marhaban, M. H. (2021). Designing an educational chatbot: A case study of CikguAIBot. In *2021 Fifth international conference on information retrieval and knowledge management (CAMP)* (pp. 19–24). https://doi.org/10.1109/CAMP51653.2021.9498011

Nguyen, D. C., Dinh, N. H. D., Pham-Nguyen, C., Le Dinh, T., & Nguyen Hoai Nam, L. (2022). ITCareerBot: A personalised career counselling chatbot. *Recent challenges in intelligent information and database systems* (pp. 423–436). Singapore: Springer Nature Singapore.

Nguyen, H. D., Tran, T.-V., Pham, X.-T., Huynh, A. T., & Do, N. V. V. (2021). Ontology-based integration of knowledge base for building an intelligent searching chatbot. *Sensors and Materials, 33*(9), 3101–3123, 2021.

Okonkwo, C. W., & Ade-Ibijola, A. (2021a). Chatbots applications in education: A systematic review. *Computers & Education: Artificial Intelligence, 2*, Article 100033. https://doi.org/10.1016/j.caeai.2021.100033

Okonkwo, C. W., & Ade-Ibijola, A. (2021b). Python-bot: A chatbot for teaching python programming. *Engineering Letters, 29*(1), 25–34.

Oralbayeva, N., Shakerimov, A., Sarmonov, S., Kantoreyeva, K., Dadebayeva, F., Serkali, N., & Sandygulova, A. (2022). K-Qbot: Language learning chatbot based on reinforcement learning. In *Proceedings of the 2022 ACM/IEEE international conference on human-robot interaction* (pp. 963–967). Sapporo, Hokkaido, Japan: IEEE Press.

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies, 27*, 7893–7925. https://doi.org/10.1007/s10639-022-10925-9

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

Palasundram, K., Sharef, N. M., Nasharuddin, N. A., Kasmiran, K. A., & Azman, A. (2019). Sequence to sequence model performance for education chatbot. *International Journal of Emerging Technologies Learning, 14*, 56–68. Retrieved from https://api.semanticscholar.org/CorpusID:212620854.

Paschoal, L. N., Melo, S. M., de Oliveira Neves, V., Conte, T. U., de Souza, S., & d, R. S. (2023). An experimental study on a conversational agent in software testing lessons. *Informatics in Education, 22*(1), 99–120. https://doi.org/10.15388/infedu.2023.01

Pereira, D. S. M., Falcão, F., Nunes, A., Santos, N., Costa, P., & Pêgo, J. M. (2023). Designing and building OSCEBot ® for virtual OSCE - performance evaluation. *Medical Education Online, 28*(1), Article 2228550. https://doi.org/10.1080/10872981.2023.2228550

Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education, 28*, 1549–1565. https://doi.org/10.1002/cae.22326

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*.

Raiche, A.-P., Dauphinais, L., Duval, M., De Luca, G., Rivest-Hénault, D., Vaughan, T., … Guay, J.-P. (2023). Factors influencing acceptance and trust of chatbots in juvenile offenders' risk assessment training. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1184016

Ruan, S., Jiang, L., Xu, J., Tham, B. J.-K., Qiu, Z., Zhu, Y., … Landay, J. A. (2019). QuizBot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI conference on human factors in computing systems, 1–13. Presented at the glasgow, scotland UK*. https://doi.org/10.1145/3290605.3300587

Sáiz-Manzanares, M. C., Marticorena-Sánchez, R., Martín-Antón, L. J., González Díez, I., & Almeida, L. (2023). Perceived satisfaction of university students with the use of chatbots as a tool for self-regulated learning. *Heliyon, 9*(1), Article e12843. https://doi.org/10.1016/j.heliyon.2023.e12843

Salas-Velasco, M. (2023). Economic and financial education for investment and financing decision-making in a graduate degree: Experimental evaluation of the effectiveness of two delivery methods. *The Journal of Economic Education, 54*(3), 225–242. https://doi.org/10.1080/00220485.2023.2191594

Salazar, C. F. (2023). Using cloud-based chatbot builder in developing pedagogical conversational agent. *International Journal of Engineering Trends and Technology*. https://doi.org/10.14445/22315381/IJETT-V71I7P229

Schmitt, A., Wambsganss, T., & Leimeister, J. M. (2022). Conversational agents for information retrieval in the education domain: A user-centered design investigation. *Proceedings ACM Human and Computer Interaction, 6*(CSCW2). https://doi.org/10.1145/3555587

Shams, M. S., Niazi, M. M., Gul, H., Mei, T. S., & Khan, K. U. (2022). E-Learning adoption in higher education institutions during the COVID-19 pandemic: A multigroup analysis. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.783087

Sharma, A., Undheim, P. E., & Nazir, S. (2023). Design and implementation of AI chatbot for COLREGs training. *WMU Journal of Maritime Affairs, 22*(1), 107–123. https://doi.org/10.1007/s13437-022-00284-0

Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education, 151*, Article 103862. https://doi.org/10.1016/j.compedu.2020.103862

Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). *Quality in qualitative evaluation: A framework for assessing research evidence*.

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – exploring the ethical issues of an emerging technology. *International Journal of Information Management, 74*, Article 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 409–426). Cambridge University Press.

Suppes, P. (1971). *Computer assisted instruction at stanford*. https://doi.org/10.1159/000393845

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2019). Development and evaluation of ClientBot: Patient-like conversational agent to train basic counseling skills. *Journal of Medical Internet Research, 21*(7), Article e12529. https://doi.org/10.2196/12529

Theelen, H., van den Beemt, A., & den Brok, P. (2019). Classroom simulations in teacher education to support preservice teachers' interpersonal competence: A systematic literature review. *Computers & Education, 129*, 14–26. https://doi.org/10.1016/j.compedu.2018.10.015

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

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540

Wambsganss, T., Winkler, R., Söllner, M., & Leimeister, J. M. (2020). A conversational agent to improve response quality in course evaluations. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1–9). https://doi.org/10.1145/3334480.3382805

Wambsganss, T., Zierau, N., Söllner, M., Käser, T., Koedinger, K. R., & Leimeister, J. M. (2022). Designing conversational evaluation tools: A comparison of text and voice modalities to improve response quality in course evaluations. *Proceedings ACM Human and Computer Interaction, 6*(CSCW2). https://doi.org/10.1145/3555619

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wieringa, R. J., & Heerkens, J. M. G. (2006). The methodological soundness of requirements engineering papers: A conceptual framework and two case sstudies. *Requirements Engineering, 11*, 295–307. https://doi.org/10.1007/s00766-006-0037-6

Wijayawardena, G. C. S., Subasinghe, S. G. T. S., Bismi, K. H. P., & Gamage, A. (2022). AI and machine learning based E – learning system for secondary education. In *2022 IEEE 7th international conference for convergence in technology (I2CT)* (pp. 1–6). https://doi.org/10.1109/I2CT54291.2022.9824643

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., & Leimeister, J. M. (2020). Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–14). https://doi.org/10.1145/3313831.3376781

Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of management proceedings*. https://doi.org/10.5465/AMBPP.2018.15903abstract

Winkler, R., Söllner, M., & Leimeister, J. M. (2021). Enhancing problem-solving skills with smart personal assistant technology. *Computers & Education, 165*, Article 104148. https://doi.org/10.1016/j.compedu.2021.104148

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? - a systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence, 4*. https://doi.org/10.3389/frai.2021.654924

Xie, T., Liu, R., Chen, Y., & Liu, G. (2021). MOCA: A motivational online conversational agent for improving student engagement in collaborative learning. *IEEE Transactions on Learning Technologies, 14*, 653–664. https://doi.org/10.1109/TLT.2021.3129800

Xu, Y., Vigil, V., Bustamante, A. S., & Warschauer, M. (2022). "Elinor's talking to me!":integrating conversational AI into children's narrative science programming. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3491102.3502050

Yigci, D., Eryilmaz, M., Yetisen, A. K., Tasoglu, S., & Ozcan, A. (2024). Large Language model-based chatbots in higher education. *Advanced Intelligent Systems*. , Article 2400429. https://doi.org/10.1002/aisy.202400429

Zhang, R., Zou, D., & Cheng, G. (2023). A review of chatbot-assisted learning: Pedagogical approaches, implementations, factors leading to effectiveness, theories, and future directions. *Interactive Learning Environments, 32*, 4529–4557. https://doi.org/10.1080/10494820.2023.2202704