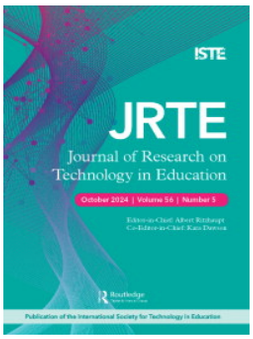

Routledge
Taylor & Francis Group

Journal of Research on Technology in Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ujrt20

The effect of chatbots on learning: a meta-analysis of empirical research

Ecenaz Alemdag

To cite this article: Ecenaz Alemdag (12 Sep 2023): The effect of chatbots on learning: a meta-analysis of empirical research, Journal of Research on Technology in Education, DOI: 10.1080/15391523.2023.2255698

To link to this article: https://doi.org/10.1080/15391523.2023.2255698

JOURNAL OF RESEARCH ON TECHNOLOGY IN EDUCATION
https://doi.org/10.1080/15391523.2023.2255698

The effect of chatbots on learning: a meta-analysis of empirical research

Ecenaz Alemdag

Technische Universität Dresden, Dresden, Germany

ABSTRACT

This meta-analysis aimed to comprehensively review empirical studies on the effect of chatbots on learning and quantitatively synthesize their findings to produce an overall effect size. Searching several databases yielded 28 eligible reports with 31 individual effect sizes. The results revealed a significant and medium effect (g = .48) of chatbots on learning. Further analyses indicated four significant moderators: the type of instruction in the comparison group, experimental duration, chatbot type, and chatbot tasks. The highest effect sizes emerged when the comparison group had no specific support, the experiment lasted only one session, and chatbots were task-focused and took care of frequently asked questions. These results suggest that chatbots can be more effective in certain cases within their overall contribution area to learning.

ARTICLE HISTORY
Received 7 May 2023
Revised 23 August 2023
Accepted 1 September 2023

KEYWORDS
Chatbot; conversational agent; learning; meta-analysis; artificial intelligence

1. Introduction

Conversational agents in education enable learners to interact with computer systems that can stimulate meaningful and individual learner-instructor conversations (Winkler et al., 2020). Intelligent tutoring systems, pedagogical agents, and language practice systems are examples of conversational agents discussed in the literature (Kerly et al., 2009). Nowadays, owing to rapid developments in artificial intelligence (AI), these agents have been transformed into chatbots through which students can more actively control their learning process (Winkler & Söllner, 2018).

Chatbots are communication tools that mimic human dialogue through written or audio means (Yin et al., 2021). There is a burgeoning interest in the use of chatbots in education to facilitate and support teaching and learning, administrative tasks in institutions, assessment, and students' research and development in career paths (Okonkwo & Ade-Ibijola, 2021; Pérez et al., 2020). Current advancements, such as ChatGPT, a large language model trained on vast amounts of data, have also attracted the attention of researchers and practitioners to the chatbots more because they can perform a variety of natural language tasks (e.g. writing essays and computer programs) (Farrokhnia et al., 2023; Kasneci et al., 2023).

The potential benefits and challenges of chatbots have also been synthesized in prior literature reviews. Chatbots enable multiple learners to quickly access information at any time, synchronously interact with chatbots, receive immediate help and guidance, obtain personalized learning content based on their characteristics, and improve their learning performance and motivation (Farrokhnia et al., 2023; Hsu et al., 2023; Huang et al., 2022; Okonkwo & Ade-Ibijola, 2021; Wollny et al., 2021). Empirical and review studies have more pronounced the advantages of chatbots in language education to provide learners with an authentic context to practice

CONTACT Ecenaz Alemdag ecenazalemdag@gmail.com Psychology of Learning and Instruction, Technische Universität Dresden, Dresden, Germany

communication in the target language and obtain learning resources (Huang et al., 2022; Ji et al., 2023; Lee & Hwang, 2022). However, there are also challenges and limitations of using chatbots in education. They include ethical issues (e.g. the privacy of the collected data, abuse of user trust, and encouragement of harmful behaviors), chatbots' limited ability to respond to all questions appropriately and sustain long conversations with visible emotional cues, students' negative attitudes and decreasing interest over time, and distractive and cognitively overloading elements in the design of chatbots (Huang et al., 2022; Kuhail et al., 2023; Okonkwo & Ade-Ibijola, 2021; Pérez et al., 2020). Overall, the use of chatbots in education requires considering both benefits and challenges and analyzing whether the benefits outweigh the challenging issues to determine the worthiness of new learning environments designed with chatbots.

Previous studies have indicated that the effects of chatbots in education are equivocal and rely on different factors to ensure a successful learning process (Deng & Yu, 2023; Fidan & Gencel, 2022; Hsu et al., 2021; Huang et al., 2022; Okonkwo & Ade-Ibijola, 2021; Vázquez-Cano et al., 2021; Zhang, Shan, et al., 2023). Therefore, it is important to determine to what extent chatbots are effective in facilitating learning (Winkler et al., 2020). Meta-analysis studies quantitively synthesizing prior empirical research revealed the medium and large effects of chatbots on language learning (Lee & Hwang, 2022; Zhang, Shan, et al., 2023) and learning achievement (Deng & Yu, 2023), respectively. However, these studies have limitations concerning the generalizability and precision of the findings because they focused on only a specific discipline (i.e. language education) and lacked strong explanations for the heterogeneity of the effect sizes for learning. Therefore, it is necessary to indicate the potential of chatbots in different subject domains (e.g. science and technology) and optimal conditions that can lead to more enhanced learning. To fill these gaps in the literature and provide more rigorous conclusions, this study aimed to conduct a systematic and comprehensive review of chatbots in education and indicate their overall effect size on learning together with six potential moderator variables (education level, subject domain, type of instruction in the comparison group, experimental duration, chatbot type, and chatbot task).

2. Literature review

2.1. Chatbots in education

A chatbot, which stands for chat robot, is a software application that interacts with users *via* audio or text to mimic real human conservation (Hsu et al., 2021). They are also regarded as intelligent agents, conversational agents, or dialog systems (Pérez et al., 2020; Yin et al., 2021). The first chatbot, ELIZA, was launched with a psycho-therapist role in 1956 (Adamopoulou & Moussiades, 2020; Smutny & Schreiberova, 2020). However, the attention of researchers and companies to chatbots increased at the beginning of the twenty-first century (Vázquez-Cano et al., 2021). Currently, there is tremendous interest and curiosity in the affordances of chatbots, especially large language models that can provide natural responses with both textual and visual information to users' free-style inquiries.

Chatbots are also integrated into education by using the available ones for educational purposes or by developing task-focused chatbots. Chatbots can perform different tasks in education and support learners in expert, facilitator, or peer roles (Kuhail et al., 2023). Garcia Brustenga et al. (2018) categorize their tasks into eight: (1) administrative and management, (2) taking care of frequently asked questions (FAQs), (3) mentoring, (4) motivation, (5) the practice of specific skills and abilities, (6) simulation, (7) reflection and metacognitive strategies, and (8) assessment and feedback (Table 1). The interaction between learners and chatbots during these tasks can be considered micro-learning activities that provide small units of content in short intervals with user control (Yin et al., 2021).

Table 1. Chatbot Tasks (Garcia Brustenga et al., 2018).

Task	Description
1. Administrative and management	Facilitating students' coordination of tasks and personal productivity (e.g. schedule management and reminders about submission deadlines)
2. Taking care of FAQs	Responding to students' frequent questions about administration or learning contents and concepts
3. Mentoring	Mentoring students during the learning process (e.g. monitoring students' understanding, providing and adapting learning content, and giving suggestions)
4. Motivation	Providing positive emotional reinforcement to enhance students' learning
5. The practice of specific skills and abilities	Provide the opportunity to exercise dialogues in language learning with simulated conversations
6. Simulation	Simulating specific professional cases and giving support for reflection
7. Reflection and metacognitive strategies	Supporting students for the regulation of their metacognitive processes and reflection on learning
8. Assessment and feedback	Assessing students and giving feedback to support students' learning

In addition to chatbots' contributions to learners' immediate access to personalized information, help, and guidance (Huang et al., 2022; Okonkwo & Ade-Ibijola, 2021; Wollny et al., 2021), empirical studies provide evidence supporting the effect of chatbots on individual and collaborative learning environments. To illustrate, Hsu et al. (2021) used a chatbot that allowed students to practice English-speaking skills in Taiwan through interactive conversations. They revealed that the experimental group with chatbots achieved significantly higher scores on the speaking test than the control group using textbooks and audio learning materials. Fidan and Gencel (2022) used chatbots to provide immediate feedback to pre-service teachers. Those who watched the videos on instructional technologies, answered related questions, and received feedback from the chatbot showed higher learning achievement and intrinsic motivation than the control group. Moreover, Kumar (2021) developed a chatbot to assist students in their group projects in an instructional design course and found greater learning achievement in the group with chatbots, although students' perceptions of learning did not differ significantly between the experimental and control groups.

There are also contradictory findings concerning the effect of chatbots on learning. Kim (2018a, 2018b) found no significant difference between the experimental and control groups in terms of total vocabulary learning and reading skills in English as a Foreign Language (EFL) education. Furthermore, Yin et al. (2021) revealed that the learning performance of the groups that either used chatbots or attended teacher lectures on the "basic college computer" course did not differ significantly after the pretest scores were controlled. Comparisons between chatbots and peer conversations also yielded insignificant results (e.g., Fidan & Gencel, 2022; Kim, 2016). Such findings lead to questions regarding whether and when the use of chatbots in education can lead to enhanced learning.

In addition, ethical issues, such as data privacy, inaccurate, inadequate, and unmeaningful responses, and a lack of visible emotional cues are the limitations of chatbots highlighted in the literature (Huang et al., 2022; Kuhail et al., 2023; Okonkwo & Ade-Ibijola, 2021; Pérez et al., 2020). For example, Han and Lee (2022) revealed that students in massive open online courses (MOOCs), especially those in Asia and Oceania, had low intention of and high resistance to using chatbots for FAQs because of chatbots' unsatisfactory or redundant answers to their questions, inquiries about their personal information, and students' communication with a robot rather than a human. Chatbots' limited ability to understand the questions completely and give totally correct and helpful responses also caused students to feel frustrated (Jasin et al., 2023). Even after new chatbots were trained with a large amount of data, similar concerns were reported, and new issues (e.g. plagiarism, cheating, and different responses to the same questions) arose (Farrokhnia et al., 2023; Tlili et al., 2023). Therefore, it is critical to weigh the benefits and challenges of chatbots before incorporating them into education. At this point, systematic literature reviews and meta-analyses can indicate the overall impact of chatbots on education.

2.2. Previous reviews on chatbots in education

Several literature reviews have explored the use of chatbots in education. They have specified the types, roles, learning activities, benefits, challenges, limitations, interaction styles, and technologies of chatbots (Huang et al., 2022; Hwang & Chang, 2021; Kuhail et al., 2023; Okonkwo & Ade-Ibijola, 2021; Pérez et al., 2020; Wollny et al., 2021; Zhang, Zou, et al., 2023). In addition, learning domains and strategies, and research methods applied in the studies on the educational use of chatbots were determined in the reviews (Hwang & Chang, 2021; Kuhail et al., 2023). However, these literature review studies did not provide the overall effect size of chatbots on learning. Pérez et al. (2020), Huang et al. (2022), Kuhail et al. (2023), and Zhang, Zou, et al. (2023) descriptively explained how chatbots affect learning in certain conditions. Nevertheless, they relied on a limited amount of empirical research and specific learning domains (e.g. language learning) and did not use statistical methods that could quantify the overall effect and account for study-level differences. Therefore, there is a need for a comprehensive meta-analysis that can present robust results concerning the effect of chatbots and the variables that change their effect on learning.

There are recent meta-analyses (Deng & Yu, 2023; Lee & Hwang, 2022; Zhang, Shan, et al., 2023) concerning the impact of chatbots on learning in literature, but these studies have some important limitations. First, Lee and Hwang (2022) searched only Korean databases and limited their meta-analysis to EFL education. They found that the overall effect size of the chatbots was .689. Despite the heterogeneity of the effect sizes, there were no significant moderators in this study (school level, publication type, treatment period, chatbot type, mode of learner interaction, and devices for interaction). Second, Zhang, Shan, et al. (2023) analyzed the impact of chatbot-assisted language learning and determined the effect size of .527. Although this effect did not significantly differ by chatbots' interactional features and interface, instruction duration, and education level, there was a significant moderating influence of target language, language domain, and learning outcome in the study. Although the two meta-analyses are valuable in indicating the contributions of chatbots to language learning, there is a need for further evidence concerning their impact on other subject domains.

Recently, Deng and Yu (2023) conducted a meta-analysis to investigate the overall effect of chatbots on different learning outcomes. They revealed a large effect of chatbots on learning achievement ($d = 1.033$). In addition, the impact of chatbots on explicit reasoning, knowledge retention, and learning interest was significant in the meta-analysis, whereas their effects on critical thinking, learning engagement, and learning motivation were insignificant. Moderator analysis conducted for the impact of chatbots on the combined learning outcomes indicated that intervention duration, chatbot roles, and learning content were not significant moderators. Even though this study is one of the first meta-analyses on the effects of chatbots in education, it has some limitations that warrant careful interpretation of findings. First, only post-test scores were considered in the calculation of individual study effect sizes, even if there were experimental studies with pretest-posttest-control group designs. Ignoring pretest data in the calculation of effect sizes decreases precision in the estimation of experimental impact (Morris, 2008). Second, although different experimental groups were learning with chatbots in one study, Deng and Yu (2023) selected only one group instead of combining them and did not explain the reason for choosing a particular treatment group. Third, the authors performed moderator analysis for the effect of chatbots on cumulative learning outcomes that included both cognitive and affective domains. Therefore, the specific role of moderators in the effect on learning achievement or performance is unknown.

2.3. Potential moderators

This study focused on six potential moderators that can change the effect of chatbots on learning. These were education level, subject domain, type of instruction in the comparison group, experimental duration, chatbot type, and chatbot task.

First, *the education level* of learners might affect how much they can learn from chatbots. Chatbots were mainly used in higher education contexts (Chiu et al., 2023; Huang et al., 2022; Lee & Hwang, 2022). Indeed, previous review research has seldom discussed chatbots at the K-12 level (Hwang & Chang, 2021). Chiu et al. (2023) suggest that compared to higher education students, K-12 students require more precise responses and guidance from chatbots since they have less advanced self-regulation skills and tend to give up more easily when they are confused or experience failure while learning. Therefore, not all chatbots might be impactful for K-12 students. However, Lee and Hwang (2022) found in their meta-analysis that chatbots were significantly effective at the primary, secondary, and university levels in Korean EFL education, and the effect of chatbots did not significantly change by grade level. Nevertheless, there is a need for a comprehensive meta-analysis to reveal how this finding is valid in a global context, covering all countries and subject domains.

Second, *the subject domain* of chatbots was investigated in previous related reviews (e.g. Hwang & Chang, 2021; Kuhail et al., 2023; Wollny et al., 2021; Zhang, Zou, et al., 2023). These reviews indicated that chatbots were mainly used in language learning and computer science fields. Regarding language learning, Huang et al. (2022) revealed the changing effects of chatbots based on specific knowledge and skills targeted in this subject domain. Zhang, Zou, et al. (2023) indicated the highest effect size in vocabulary learning and the lowest effect size in reading and writing. However, the effect of chatbots on other educational domains is not well known (Kumar, 2021). A meta-analysis by Deng and Yu (2023) revealed that learning content was not a significant moderator of the impact of chatbots on the cumulative seven educational outcomes (e.g. engagement, motivation, and explicit reasoning). Nevertheless, the meta-analysis did not perform a moderator analysis only for the learning achievement outcome. This leaves an open question about whether conversation-focused short-term learning activities or micro-learning opportunities with chatbots can benefit learning in all subject areas (Yin et al., 2021).

Third, *the type of instruction in the comparison group* can moderate the effect of chatbots on learning. Meta-analyses on intelligent tutoring systems (ITSs), which also function as conversational agents and have similar roles as chatbots, provide evidence of the potential impact of this moderator. For example, Ma et al. (2014) first defined ITSs as a computer program that provides tutoring functions, creates multidimensional models for each student's psychological state, and adapts their functions based on these models. Then, they revealed that ITSs were not significantly more effective when compared with small-group or individual human instruction, whereas they were more effective than large-group human instruction. Xu et al. (2019) conducted a meta-analysis with 88 individual effect sizes to synthesize empirical evidence regarding the effect of ITSs on K-12 students' reading comprehension. They found that when compared to traditional instruction, ITSs had a large effect size on learning measured with researcher-designed and standardized instruments; on the other hand, they had a small effect when compared to human tutoring. Overall, previous reviews on ITSs suggest that conversational agents including chatbots could be more effective for learning than traditional lecture-based instruction, but their impact might be similar to that of individual human conversations. It is also imperative to compare chatbot-supported learning with other technological tools (Huang et al., 2022).

Fourth, *the experimental duration* of research studies on educational programs is an important factor that educators and policymakers need to consider before deciding to integrate a program into their practice (Slavin, 2008). Interventions lasting at least 12 weeks are recommended for consideration because they can indicate their applicability and impact over extended periods and support the external validity of the studies (Slavin, 2008). Brief studies also suffer from the novelty effect that emerges when the treatment is new to the participants, and its effect is significant only in the initial stages. Correspondingly, Fryer et al. (2017) found that the task interest of students interacting with chatbots for English-speaking tasks decreased during the 12 weeks intervention, and it was significantly lower than that of students interacting with human partners in the second and third measurement times. The authors attributed this finding to the novelty effect that led to high interest at the beginning of the intervention but then declined interest

over time. Therefore, it is important to investigate whether the effect of chatbots on learning remains permanent over long periods by synthesizing all related prior research.

Fifth, the impact of chatbots on learning can differ by *chatbot type*. Two general types of chatbots are task- and non-task-oriented (Yin et al., 2021). While task-oriented chatbots aim to respond to user inquiries about a specific task in brief conversations, non-task-oriented chatbots stimulate causal conservations for entertainment (Yin et al., 2021). Grudin and Jacques (2019) also classify non-task-oriented chatbots into virtual companions and intelligent assistants. Virtual companions (e.g. ELIZA and Cleverbot) can talk about any topic and maintain a deep conversation (Grudin & Jacques, 2019). Similarly, intelligent assistants (e.g. Siri and Google Assistant) can respond to any topic, but their conservations are generally short and limited to performing a task. Concerning the impact of chatbot types on learning, Lee and Hwang (2022) found in their meta-analysis that general-purpose chatbots were less effective in EFL learning than chatbots with the specific purpose of teaching English. General-purpose or non-task-oriented chatbots can be considered open environments that rely on individuals' efforts to identify their learning needs, set learning goals, and be involved in learning activities (Hannafin et al., 1999). Learners with low self-regulation skills and prior knowledge can encounter problems in such environments and experience cognitive load while directing their learning processes with minimal guidance (Kirschner et al., 2006). Hence, not all learners might benefit from non-task-oriented chatbots.

Finally, chatbots can perform different *tasks* (Table 1). Vázquez-Cano et al. (2021) claimed that chatbots can be beneficial resources to practice and improve communication and linguistic skills that necessitate continuous practice and immediate feedback. In addition, Pérez et al. (2020) highlighted the role of chatbots as teaching assistants in facilitating repetitive tasks, such as answering FAQs and relieving the burden on human teachers. Although chatbot tasks may vary, it is unknown whether chatbots can complete each task successfully to enhance learning. Deng and Yu (2023) compared the effects of three chatbot roles (i.e. teaching assistant, tutor, and partner) on learning and found no significant difference in the meta-analysis. However, this categorization of chatbot roles provides inadequate insights into the effectiveness of chatbots in specific tasks.

2.4. Research aim and questions

There is an increasing interest in the use of chatbots in education, especially with the emergence of large language models (Kasneci et al., 2023; Okonkwo & Ade-Ibijola, 2021; Pérez et al., 2020; Smutny & Schreiberova, 2020). However, "our understanding of the effects of chatbot on student outcomes is still limited" (Huang et al., 2022, p. 253) because of greater emphasis on chatbot development rather than its use (Han & Lee, 2022) and inconsistent findings in the literature (Deng & Yu, 2023; Fidan & Gencel, 2022; Hsu et al., 2021; Huang et al., 2022; Okonkwo & Ade-Ibijola, 2021; Vázquez-Cano et al., 2021; Zhang, Shan, et al., 2023). It is crucial to determine the circumstances in which chatbots can respond to certain pedagogical needs and support human teaching (Hsu et al., 2023; Kumar, 2021). The current study was a meta-analysis that aimed to combine the findings of related empirical research and provide more precise answers concerning the influence of chatbots on learning. Moderator analysis was also employed to explain the causes of variations in study effects and determine conditions that are more conducive to enhanced learning with chatbots. More specifically, this study aimed to answer the following research questions:

1. What is the effect of students' use of chatbots on learning?
2. How does the effect of chatbots on learning differ by moderators (education level, subject domain, type of instruction in the comparison group, experimental duration, chatbot type, and chatbot task)?

3. Methods

This study is a meta-analysis that aims to provide a quantitative summary of results concerning the effect of chatbots on learning. In contrast to narrative reviews that synthesize primary studies and discuss the variability between studies descriptively, meta-analyses benefit from statistical techniques to generate quantitative estimates for the overall effect size and heterogeneity of individual effect sizes (Petticrew & Robert, 2006). More particularly, meta-analysis studies use explicit and systematic steps for study selection and data collection, assign the weights to the studies based on mathematical criteria, and apply statistical techniques to account for differences between studies, which leads to a transparent, impartial, and repeatable basis for discussion of findings (Borenstein et al., 2009; Lipsey & Wilson, 2001). They are especially helpful in resolving controversies concerning contradictory studies and quantitatively identifying the reasons for different results (Deeks et al., 2022). Because of these advantages, the current research used meta-analysis and followed the updated guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021) to present findings more clearly, thoroughly, and precisely.

3.1. Eligibility criteria

The sources were screened based on the following criteria to select eligible ones for this meta-analysis:

- The study is an empirical research article, a dissertation, or a conference paper.
- The study had at least one experimental group with a chatbot used for educational purposes and a comparison group without a chatbot.
- The participants were students receiving formal education in an institution.
- The chatbot was developed by researchers or other professionals, not students.
- The study measured individual and immediate learning scores or performance.
- The study provides the effect size or sufficient information to obtain the effect size (means, standard deviations, sample size, *F*, *t*).
- The language of the study is English.
- The full text of the source is accessible.

3.2. Information sources and selection process

Information sources for this meta-analysis were searched on the following databases on July 8, 2022: Web of Science, Scopus, ERIC, PsycArticles, Education Source, and ProQuest. Titles, abstracts, and keywords of the sources were screened by using these terms and Boolean operators: (chatbot* or "conversational agent*" or "conversational tutor*" or "chatterbot*") and (success* or achievement* or performance* or learning* or outcome* or education*) and (group* OR *experiment* OR treatment*). The search was limited to conference papers, (early access) articles, and thesis/dissertations in English. A total of 1659 records were identified from the aforementioned databases, and 13 documents were found when citations of relevant studies and prior systematic reviews on educational chatbots (e.g. Huang et al., 2022; Hwang & Chang, 2021) were examined. Screening of titles and abstracts and then full texts resulted in 28 eligible reports that included 31 effect sizes related to the effect of chatbots on learning. The procedure followed to identify relevant studies is shown in Figure 1.

3.3. Data collection process and data items

This meta-analysis examined studies on the impact of chatbots on learning outcomes. Data were collected from empirical studies with experimental and control groups that immediately measured

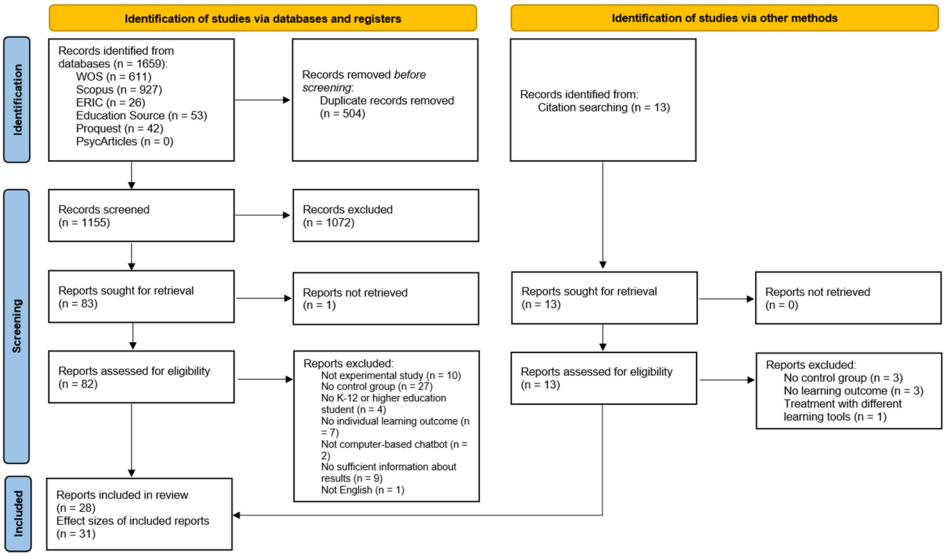

Figure 1. PRISMA flow diagram of the meta-analysis.

learners' knowledge or skills after the intervention with chatbots. In addition, data regarding the variables that could moderate the effect of chatbots were gathered. Categories of the moderator variables listed in Table 2 were identified based on previous reviews and classifications (e.g. Garcia Brustenga et al., 2018; Grudin & Jacques, 2019; Ma et al., 2014). Then, the author coded each study in terms of these variables (Table 3). It is important to note that the most emphasized and dominant task of chatbots in the reviewed studies was selected while determining chatbot tasks according to Garcia Brustenga et al.'s (2018) classification.

A second coder, who was a research and design expert in instructional systems technology, also analyzed all studies based on the list in Table 2. According to Landis and Koch's (1977) benchmarks, Cohen's Kappa values for intercoder reliability between the author and second coder were found to be almost perfect (>.80) except that for the moderator of the chatbot main task (k = .76). The coders also discussed disagreements and resolved them until they reached a consensus.

3.4. Effect measures and synthesis methods

The standardized mean difference was selected as an effect size statistic for the studies with two independent groups in this meta-analysis. Moreover, Hedge's g formula was applied to correct for bias based on a small sample size (Hedges & Olkin, 1985; Lipsey & Wilson, 2001). In the case of missing information regarding group means and standard deviations, F and t values were used to obtain the effect size (Borenstein, 2009). For the studies that used a pretest-posttest-control group design and reported pretest scores with means and standard deviations, effect size estimates using pooled pretest SD were calculated (Morris, 2008). Moreover, when the effect of the pretest was controlled through an ANCOVA design, the effect size was calculated with adjusted means by following Schmidt and Hunter's (2015) formula. *The Meta-Essentials* tool (Suurmond et al., 2017) was used to record the effect sizes for selected studies and perform a meta-analysis.

Some studies included multiple comparison or experimental groups and multiple learning outcomes. In the studies with more than one comparison group (Fidan & Gencel, 2022; Graesser et al., 2003; Kim, 2016; Xu et al., 2021), only one group was selected because the type of instruction in the comparison group was a moderating variable in this meta-analysis.

Table 2. Moderating Variables and Their Categories.

Variable	Categories
Education level	K12
	Higher education (undergraduate)
Subject domain	Computer science
	Foreign language education
	Health sciences
	Instructional design (Native) language and literacy
	Natural science (e.g. physics and chemistry)
	Psychology
	General
Type of instruction in the comparison group	Individual human (e.g. adult and peer) conversation
	Individual computer-based instruction (CBI)
	Individual learning with text- or audio-based sources
	Large group teacher lecture
	No specific support (after the instruction or during the task performance)
Experimental duration	One session
	Two sessions-<1 week
	1–4 weeks
	5–11 weeks
	12 and above weeks
Chatbot type	Task-focused chatbot
	Virtual companion
	Intelligent assistant
Chatbot main task	Administrative and management tasks
	Taking care of FAQs
	Student mentoring
	Motivation
	Practice of specific skills and abilities
	Simulation
	Reflection and metacognitive strategies
	Assessment and feedback

Comparison groups providing alternative types of instruction (e.g. individual human conversation and individual learning with text- or audio-based sources) were selected to vary the comparison condition. However, when there were multiple experimental groups (Jeon, 2021; Kim et al., 2021; Tegos et al., 2016), these groups were combined with Higgins et al.'s (2021) formula to prevent the violation of the independence of effect size that could emerge from multiple comparisons of the same control group with different experimental groups. Nevertheless, the experimental groups were not combined in the studies that did not provide adequate information about the scores of experimental groups (Hoque et al., 2013) or included groups using chatbots with limited interaction capabilities, as in the non-scaffolding group in the study by Winkler et al. (2020).

In addition, two studies (Hayashi, 2020; Kim, 2016) included subgroups that provided independent information about the effect of chatbots in different cases (e.g. the gaze visibility of the learning partner and listening proficiency level). These subgroups were treated as independent studies, as suggested by Borenstein et al. (2009). Finally, some studies (Jeon, 2021; Kim et al., 2021; Ruan et al., 2021; Winkler et al., 2020) presented results for multiple learning outcomes. Since the same participants were used in these studies to measure different outcomes, the effect sizes were not independent (Borenstein et al., 2009). Therefore, a single value which was the mean of the learning outcomes in these studies was calculated if the report included adequate information for each outcome without total scores.

The random effects model was selected for this meta-analysis because it was assumed that effect sizes might vary among studies applying different methods, such as in the selection of participants and experimental treatment (Borenstein et al., 2009). Accordingly, heterogeneity analysis was also performed using Q and I^2 statistics to indicate a statistically significant variation in the true effect sizes and quantify the amount of variance. In addition, a subgroup analysis with the proposed moderator variables (Table 2) was conducted to explain the potential reasons

Table 3. Studies Included in the Meta-Analysis and Data on Moderator Variables.

References	Type of source	Education level	Subject domain	Comparison group	Duration	Chatbot type	Chatbot main task
Brachten et al. (2020)	Article	Higher education	General	Individual text-or audio-based sources	One session	Task-focused	Taking care of FAQs
Chang, Hwang, et al. (2022)	Article	Higher education	Health sciences	No specific support	One session	Task-focused	Taking care of FAQs
Chang, Kuo, et al. (2022)	Article	Higher education	Health sciences	Large group teacher lecture	One session	Task-focused	Assessment and feedback
Deng et al. (2018)	Conference	Higher education	(Native) language and literacy	Individual CBI	NA	Task-focused	Mentoring
Deveci Topal et al. (2021)	Article	K-12	Natural science	No specific support	1–4 weeks	Task-focused	Taking care of FAQs
Fidan and Gencel (2022), peer feedback *vs.* chatbot feedback	Article	Higher education	Instructional design	Individual human conversation	1–4 weeks	Task-focused	Assessment and feedback
Graesser et al. (2003), read-textbook *vs.* AutoTutor	Conference	Higher education	Natural science	Individual text-or audio-based sources	1–4 weeks	Task-focused	Mentoring
Hayashi (2020)	Article	Higher education	Psychology	No specific support	One session	Task-focused	Reflection and metacognitive strategies
Hoque et al. (2013), control *vs.* MACH + video + feedback	Conference	Higher education	General	Individual CBI	One session	Task-focused	Assessment and feedback
Hsu et al. (2021)	Article	Higher education	(Native) language and literacy	Individual text-or audio-based sources	12 and above weeks	Task-focused	Practice of specific skills and abilities
Jeon (2021), control *vs.* CA-DA + CA-NDA	Article	K-12	Foreign language education	No specific support	Two sessions-<1 week	Task-focused	Assessment and feedback
Kim (2016), peer chat *vs.* chatbot	Article	Higher education	Foreign language education	Individual human conversation	12 and above weeks	Companion	Practice of specific skills and abilities
Kim (2018a), total score (listening and reading)	Article	Higher education	Foreign language education	No specific support	12 and above weeks	Companion	Practice of specific skills and abilities
Kim (2018b)	Article	Higher education	Foreign language education	No specific support	5–11 weeks	Companion	Practice of specific skills and abilities
Kim et al. (2021), face-to-face *vs.* AI-text and voice	Article	Higher education	Foreign language education	Individual human conversation	12 and above weeks	Mixed	Practice of specific skills and abilities
Kron et al. (2017)	Article	Higher education	Health sciences	Individual CBI	Two sessions-<1 week	Task-focused	Simulation
Lin and Chang (2020)	Article	Higher education	Psychology	No specific support	1–4 weeks	Task-focused	Mentoring
Mageira et al. (2022)	Article	K-12	Foreign language education	Individual CBI	Two sessions-<1 week	Task-focused	Assessment and feedback
Mejbri et al. (2017)	Conference	K-12	Natural science	Individual CBI	One session	Task-focused	Assessment and feedback
Nghi et al. (2019)	Article	Higher education	Foreign language education	No specific support	NA	Task-focused	Mentoring
Ruan et al. (2021)	Conference	Higher education	Foreign language education	Individual text- or audio-based sources	Two sessions-<1 week	Task-focused	Practice of specific skills and abilities

(*Continued*)

Table 3. Continued.

References	Type of source	Education level	Subject domain	Comparison group	Duration	Chatbot type	Chatbot main task
Song and Kim (2021)	Article	Higher education	Instructional design	Individual text- or audio-based sources	12 and above weeks	Task-focused	Reflection and metacognitive strategies
Tegos and Demetriadis (2017)	Article	Higher education	Computer science	No specific support	One session	Task-focused	Mentoring
Tegos et al. (2016), control and U + D treatment	Article	Higher education	Computer science	No specific support	One session	Task-focused	Mentoring
Wambsganss et al. (2021)	Conference	Higher education	(Native) language and literacy	Individual CBI	One session	Task-focused	Assessment and feedback
Winkler et al. (2020), control (video only) *vs.* SARA	Conference	Higher education	Computer science	No specific support	One session	Task-focused	Mentoring
Xu et al. (2021), human conversation *vs.* CA-conversation	Article	K-12	(Native) language and literacy	Individual human conversation	One session	Task-focused	Assessment and feedback
Yin et al. (2021)	Article	Higher education	Computer science	Large group teacher lecture	One session	Task-focused	Mentoring

CBI: computer-based instruction; NA: not available.

for variance (Lipsey & Wilson, 2001). The magnitudes of the overall and group effect sizes were interpreted using Cohen's (1988) thresholds (small: $d = 0.20$; moderate: $d = 0.50$; large: $d = 0.80$).

A sensitivity analysis was also conducted to evaluate the robustness of the synthesized findings (Page et al., 2021). To this end, outliers not within the range of −2 and +2 of the overall effect size were first identified. Then, the effect sizes with and without each outlier were compared to determine influential cases (Viechtbauer & Cheung, 2010). Consequently, one study (Hsu et al., 2021) that substantially changed the overall effect size was omitted from the meta-analysis, and 31 effect sizes were considered in the meta-analysis.

3.5. Bias assessment

Because studies with higher effect sizes tend to be published more frequently, meta-analysis research can have a risk of including mostly these studies, which can cause publication bias in meta-analysis (Borenstein et al., 2009). Several methods were applied to check for publication bias in this meta-analysis. First, the distribution of studies around the mean effect size in the funnel plot was evaluated in terms of symmetry (Borenstein et al., 2009). Second, Egger's regression and Begg and Mazumdar's rank correlation tests were conducted. Third, Rosenthal's Fail-safe N was calculated to obtain the number of studies to make the overall effect size insignificant and check whether it is $>5k + 10$ (k is the number of effect sizes) (Rosenthal, 1979). Finally, the observed effect size and adjusted size with imputed studies were compared with Duval and Tweedie's trim-and-fill procedure.

4. Findings

4.1. Study characteristics

This study included 28 eligible reports (21 journal articles and seven conference papers) that produced 31 effect sizes related to the effect of chatbots on learning. Most effect sizes ($k = 26$) were calculated from the studies conducted with higher education students, specifically undergraduate students. Although the subject domains varied, chatbots were mostly used for foreign language education ($k = 11$). The groups compared to the experimental group with chatbots usually provided no specific support ($k = 12$). The experimental duration of many studies was one session ($k = 13$). While the chatbots were generally task-based ($k = 25$), they performed different tasks (e.g. taking care of FAQs, mentoring, and practice of specific skills and abilities). Assessment and feedback ($k = 10$) were the most frequent tasks of the chatbots.

4.2. The overall effect of chatbots on learning

The overall effect of chatbots on learning was significant in the random-effects model (Table 4). The effect size was Hedges' g = .48, which corresponded to a nearly medium effect threshold (.50) (Cohen, 1988). The forest plot of individual and overall effect sizes is available in the Appendix. The heterogeneity test also yielded a significant value ($Q = 120.80$, p = .00). I^2 = 75.17% indicated high heterogeneity in effect sizes (Higgins et al., 2003). Hence, moderator analysis was conducted to identify the possible reasons for variance across the effect sizes.

Table 4. The Overall Effect of Chatbots on Learning.

		Effect size					Heterogeneity		
	k	Hedge's *g*	*SE*	95% CI	*Z*	*p*	*Q*	*p*	I^2
Fixed	31	.37	.04	[.29, .45]	9.39	.00	120.80	.00	75.17
Random	31	.48	.10	[.27, .69]	4.69	.00			

4.3. Moderator analysis

The six variables listed in Table 5 were considered in the moderator analysis. Categories including only one effect size (e.g. 5–11 weeks experimental duration, mixed chatbot type, and simulation role) were not included in the analysis.

Regarding *education level*, the effect sizes were .57 and .47 for K-12 and higher education students. Q_B was insignificant (Q_B = .17, p = .676), revealing that the effect of chatbots on learning did not differ in terms of education level.

Concerning *the subject domain*, the largest effect size was found in health sciences (g = 1.13), followed by general (g = .77), psychology (g = .73), computer science (g = .53), natural science (g = .47), instructional design (g = .35), foreign language learning (g = .31), and (native) language and literacy (g = .20). However, the moderating effect of the subject domain on learning with chatbots was not significant (Q_B = 6.72, p = .459).

In terms of the *type of instruction in the comparison group*, effect sizes varied considerably. The effect size of chatbots was large compared to the group without specific support (g = .75) and large group teacher lecture (g = .68). Chatbots also had a moderate effect on learning when students in the comparison group individually learned with text- and audio-based sources (g = .48) and CBI (g = .36). However, there was no effect of chatbots (g = .02) when the control group involved individual human conservation. The Q_B was also significant (Q_B = 9.76, p = .045), which showed that the effect size of chatbots changed based on the instruction in the comparison group.

Regarding the *experimental duration*, the effect size of the chatbots tended to decrease when the treatment lasted longer. The highest effect size was obtained in the experiments completed in one session (g = .78), followed by two sessions-<1 week (g = .55), 1–4 weeks (g = .24), and

Table 5. Analysis of Moderator Variables.

Variable	Categories	k	g	95% CI	I^2 (%)	Q_B	p
Education level	K-12	5	.57	[−0.18, 1.33]	85.18	.17	.676
	Higher education	26	.47	[.26, .67]	73.16		
Subject domain	Health sciences	3	1.13	[.06, 2.19]	92.47	6.72	.459
	General	2	.77	[.63, .91]	.00		
	Psychology	3	.73	[−0.27, 1.73]	91.47		
	Computer science	4	.53	[.19, .88]	63.19		
	Natural science	3	.47	[.31, .63]	.00		
	Instructional design	2	.35	[.15, .56]	.00		
	Foreign language education	11	.31	[−0.08, .70]	78.06		
	(Native) language and literacy	3	.20	[−0.16, .56]	35.79		
Type of instruction in the comparison group	No specific support	12	.75	[.38, 1.13]	84.95	9.76	.045*
	Large group teacher lecture	2	.68	[−0.94, 2.30]	92.79		
	Individual learning with text- or audio-based sources	5	.48	[.22, .75]	.00		
	Individual CBI	6	.36	[.14, .59]	18.60		
	Individual human conversation	6	.02	[−0.25, .29]	35.23		
Experimental duration	One session	13	.78	[.46, 1.09]	73.28	9.78	.044*
	Two sessions-<1 week	5	.55	[−0.17, 1.27]	85.07		
	NA	2	.39	[−0.23, 1.01]	86.20		
	1–4 weeks	4	.24	[.06, .42]	.00		
	12 and above weeks	6	.08	[−0.22, .38]	43.37		
Chatbot type	Task-focused chatbot	25	.59	[.37, .81]	66.57	6.58	.010*
	Virtual companion	5	−0.02	[−0.33, .30]	3.53		
Chatbot main task	Taking care of FAQs	3	.96	[.11, 1.80]	79.50	10.84	.028*
	Reflection and metacognitive strategies	3	.93	[.14, 1.73]	77.73		
	Assessment and feedback	10	.60	[.18, 1.03]	77.26		
	Mentoring	8	.41	[.18, .65]	73.60		
	Practice of specific skills and abilities	6	.04	[−0.23, .31]	29.66		

*A p-value ≤0.05 denotes significance.

12 and above weeks (g = .08). The Q_B also revealed that the effects of chatbots on learning differed significantly according to experimental duration (Q_B = 9.78, p = .044), favoring shorter treatments.

In terms of *chatbot type*, task-focused chatbots had a positive effect on learning (g = .59), but the chatbots in the virtual companion type did not produce any substantial impact (g = −0.02). The difference between the effects of task-focused and virtual companion chatbots on learning was also significant (Q_B = 6.58, p = .01).

Regarding *chatbot tasks*, the effect of chatbots on learning ranged from .04 and .96. The effect sizes were g = .96 for taking care of FAQs, g = .93 for reflection and metacognitive strategies, g = .60 for assessment and feedback, g = .41 for mentoring, and g = .04 for the practice of specific skills and abilities. While the effect of chatbots taking care of FAQs on learning was the highest, chatbots for the practice of specific skills and abilities had the lowest impact. The Q_B was also significant (Q_B = 10.84, p = .028), indicating that chatbot roles differentiated the effect of chatbots on learning.

4.4. Bias assessment

The first bias assessment with inspection of the funnel plot revealed a higher concentration of smaller studies on the bottom right side of the plot (Figure 2). This indicated an asymmetric distribution of studies and the possibility of publication bias. Second, while the Egger regression test showed publication bias (t = 2.41, p = .023), Begg and Mazumdar's rank correlation test resulted in an insignificant correlation value (Kendall's Tau a = 0.21, z = 1.65, p = .099) and implied low publication bias. Third, Rosenthal's Fail-safe N was found to be 1196. Because this number largely exceeds $5k + 10$ (it is equal to 5*31 + 10 = 165 for this study), the robustness of the synthesized results can be suggested (Rosenthal, 1979). Finally, using Duval and Tweedie's trim-and-fill procedure, the effect size was recalculated with three imputed missing studies (Figure 2). Since both the observed (g = .48) and adjusted effect sizes (g = .29) were in the small to medium range according to Cohen's (1988) criteria, it was more likely to propose the validity of the observed effect (Borenstein et al., 2009).

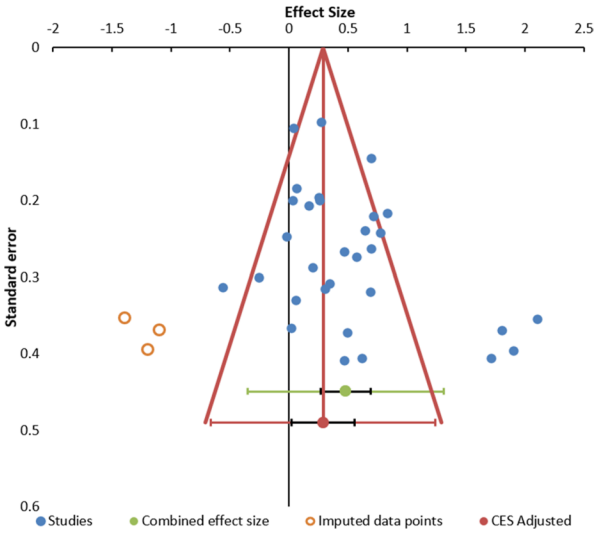

Figure 2. Funnel plot of both observed and imputed studies.

5. Discussion and recommendations

5.1. Discussion

This study quantitatively synthesized the findings of empirical research on the impact of chatbots on learning. The meta-analysis yielded a significant and moderate effect size. Contradictory findings regarding the effect of chatbots have posed questions about the educational value of this technology and the conditions conducive to learning with chatbots (Deng & Yu, 2023; Fidan & Gencel, 2022; Hsu et al., 2021; Huang et al., 2022; Okonkwo & Ade-Ibijola, 2021; Vázquez-Cano et al., 2021; Zhang, Shan, et al., 2023). Several literature reviews and commentaries on the use of chatbots in education have also drawn attention to both their advantages (e.g. quick access to information and personalized support) and disadvantages (e.g. data privacy and inaccurate and unmeaningful responses). This meta-analysis indicated that the overall impact of chatbots on learning was significant and positive, but the heterogeneity test revealed significant variation in effect sizes, which led to further analyses with potential moderator variables.

First, regarding education level, moderator analysis revealed no significant difference in effect sizes between the K-12 and higher education (undergraduate) student groups. Likewise, Lee and Hwang (2022) and Zhang, Shan, et al. (2023) found that chatbots had similar effects at different educational levels in the language learning context. Kasneci et al. (2023) suggest that chatbots trained with large language models can facilitate the development of reading, writing, language, and problem-solving skills of K-12 students. However, Xia et al. (2023) claim that chatbots might not be competent enough to facilitate the learning of young students with limited disciplinary knowledge. Moreover, undergraduate and graduate students can benefit from chatbots with different degrees. Sáiz-Manzanares et al. (2023) found that master's students used chatbots integrated with the learning management system more frequently, achieved higher learning outcomes, and perceived more satisfaction than undergraduate students. On the other hand, comparing the effect of chatbots at the K-12 and higher education levels, this study suggests that chatbots can be effective agents independent of education level. However, it is important to note the limited number of studies conducted at the K-12 level.

Second, the impact of chatbots on learning did not differ by subject. Similarly, Deng and Yu (2023) found that learning content was not a significant moderator of chatbots' effect on educational outcomes. Intriguingly, the lowest effect sizes belonged to the domains of foreign language learning (g = .31) and (native) language and literacy (g = .20). Lee and Hwang (2022) revealed a medium effect size of chatbots in the Korean EFL education context. Bibauw et al. (2022) also indicated the medium impact of overall dialog systems on second language development. However, Huang et al. (2022) highlighted that the effects of chatbots can change based on the specific knowledge and skills targeted in language learning. Correspondingly, Zhang, Shan, et al. (2023) revealed the low impact of chatbots on the reading and writing domains of language learning. In the current study, focusing on the use of chatbots in overall foreign and native language learning in the global context, small-to-medium effect sizes were obtained. Although chatbots enable students to frequently and easily practice language skills in an authentic language environment without a human partner and to feel less anxiety and worry (Adamopoulou & Moussiades, 2020; Huang et al., 2022; Jeon, 2022; Ji et al., 2023; Kim et al., 2021), they have some limitations and challenges that can decrease the positive influence of their affordances on learning. For example, some Korean primary school students in Jeon's (2022) research drew attention to the exhausting speech recognition problems of chatbots and favored human partners for more active engagement with language activities. Failures of students with low language skills to maintain spoken conversations were also reported in the research. Similar problems were pronounced in the study by Kim et al. (2021), especially when voice chatbots were used by participants. In addition, Kim et al. (2021) indicated that unclear and irrelevant chatbot messages and abrupt changes in conversation topics were the disadvantages of this technology in language learning. On the whole, though chatbots seem promising to contribute to language learning, the

results imply the need for more amendments in their development and implementation to optimize their benefits.

Third, the type of instruction in the comparison group significantly changed the effect size of the chatbots on learning. The chatbots had the highest effect on learning when compared to the group without specific support. Their effect was also greater than those of large-group teacher lectures and individual learning with text- and audio-based resources. However, chatbots did not outperform individual human (e.g. adult and peer) conversations. Likewise, meta-analysis studies on ITSs (Ma et al., 2014; Xu et al., 2019) indicated the lack of effect of this type of conversational agent when compared to small-group or individual human instruction or tutoring. Despite the latest advancements in chatbot technology that can perform various natural language tasks, Kasneci et al. (2023) underscore that "they can only serve as assistive tools to human learners and educators and cannot replace the teacher" (p. 5). The lack of a deep understanding of the words, the risk of provision of inaccurate, biased, discriminated, and outdated information, and less ability to generate content requiring higher-order thinking skills are current weaknesses of these advanced technologies (Farrokhnia et al., 2023). Teachers who are still the main agents in the design, evaluation, and decision-making process can overcome these challenges by collaborating with chatbots (Ji et al., 2023). This study also suggests that chatbots might not be powerful enough to replace productive individual conversations between learners and teachers or peers now.

Fourth, experimental duration was a significant moderator of the effect of chatbots on learning. The results indicated that the effect tended to decrease when the treatment in the reviewed studies lasted longer. The highest effect size was obtained in studies in which the participants used chatbots for only one session. Similarly, in the meta-analysis by Lee and Hwang (2022), the effect of chatbots on EFL learning was noticeably less when they were implemented for more than eight weeks. The novelty effect, indicated as one of the challenges of using chatbots in education (Huang et al., 2022), could have played a role in the initially high but decreasing impact of chatbots over time. Winkler et al. (2020) and Zhang, Zou, et al. (2023) drew attention to the students who had higher motivation when they started using chatbots for the first time. Moreover, Fryer et al. (2017) indicated that students' task interest in conversations with chatbots decreased during a 12-week intervention. Such changes in students' motivation that occur in the integration of new technologies into education might have also caused less permanent learning benefits of chatbots.

Fifth, in terms of chatbot type, task-focused chatbots had a significantly higher impact on learning than virtual companions. In contrast to task-focused chatbots designed for a specific learning activity, virtual companions can make conversations with users on any topic. However, learning with virtual companions depends mainly on students' initiative in determining their learning needs and directing their learning process with self-generated prompts. Although such an open learning environment provides autonomy for students, those with low self-regulation skills and prior knowledge might have difficulty in following an effective learning path (Lin, 2019; Terras et al., 2013). For learners with limited language proficiency, maintaining goal-oriented conversations can be even more challenging (Jeon, 2022). In addition, low digital literacy in the effective use of language while giving prompts to AI applications (e.g. ChatGPT) can prevent learners from obtaining accurate and thorough explanations from these applications (Bozkurt, 2023). Hence, instructor guidance might become more important for learners with inadequate knowledge and skills to benefit from general-purpose chatbots as well. Correspondingly, Chiu et al. (2023) found that when students were novices in terms of self-regulated learning and digital literacy, teacher support during learning with chatbots (e.g. explanations about tasks and feedback on chatbot usage) became a significant variable in producing higher student motivation and competence at this learning activity. Moreover, undergraduate students in the research by Sáiz-Manzanares et al. (2023) attributed their low usage of chatbots to the little guidance in the learning environment, although task-focused chatbots were incorporated into learning management systems. All these studies corroborate the need for specific guidance and support in the use of general-purpose chatbots, especially for novice students.

Finally, the effect of chatbots significantly differed by their tasks. The highest effect was found with the chatbots taking care of FAQs. Compared to the static content search on the web pages of FAQs, chatbots can be perceived as more friendly and attractive by users (Adamopoulou & Moussiades, 2020). The use of chatbots for this task is also crucial to provide immediate assistance to the students, especially in large classes or MOOCs, if chatbots provide quality and barrier-free content (Han & Lee, 2022). To illustrate, Jasin et al. (2023) indicated that students utilizing chatbots for their questions in the chemistry course admired their accessibility at any time and speed for answering the questions despite some unhelpful responses. The intriguing finding concerning chatbot tasks pertained to the lowest effect of chatbots developed for the practice of specific skills and abilities. It is important to note that the main task of these chatbots was providing students with the opportunity to practice EFL skills (Kim, 2016, 2018a, 2018b; Kim et al., 2021). However, they were general-purpose chatbots without specific assessment and feedback mechanisms. Feedback based on recursive dialogs between learners and significant others (e.g. teacher and peers) is a critical component of foreign language learning for students to understand their strengths and weaknesses in their performance and ways to improve their learning (Vattøy, 2020; Vattøy & Smith, 2019). Correspondingly, in one of the reviewed studies (Kim et al., 2021), some participants highlighted the lack of feedback and unidirectional communication as a disadvantage of both text-based and voice-based conversations with chatbots for speaking practice. These chatbots might be more beneficial for students with good English proficiency who can meet autonomy, competence, and relatedness needs and regulate their learning while using chatbots (Xia et al., 2023). Therefore, it could be important to pay more attention to the feedback elements and learner characteristics in the design of chatbots for the practice of skills and abilities to enhance learning.

5.2. Practical implications

This study shows that chatbots can contribute to learning at a medium level. Although their effect is independent of education level and subject domain, some factors concerning the implementation of chatbots can change their impact on learning. Since instructors play a critical role in ensuring the successful integration of chatbots into education (Winkler et al., 2020), this study gives some practical suggestions that can help them in their design decisions.

First, chatbots can provide more effective learning opportunities when students do not have access to specific support, or instructional options are limited to only large group teacher lectures and individual learning with text- or audio-based sources. Especially when the practitioners do not have enough time to offer one-to-one guidance in large classes, they can benefit from chatbots that can respond to students' inquiries and support their learning process. Second, the use of chatbots might be restricted to short durations, such as 1 week, since their contribution to learning appears to decrease significantly over time. Third, it is more advisable for practitioners to prefer task-focused chatbots designed for specific learning content and activities than virtual companions. Finally, chatbots can be more beneficial for learning when their tasks are taking care of FAQs, reflection and metacognitive strategies, assessment and feedback, and mentoring because effect sizes for these tasks were found above .40, a hinge-point for an intervention to be regarded as worthy of implementing (Hattie, 2009). However, concerning the chatbots for the practice of specific skills and abilities, educators can give more attention to whether the feedback is a component of these chatbots to inform students' performance levels and provide suggestions for improvement. Otherwise, their impact on learning might not be visible.

5.3. Limitations and future directions

There are some limitations of the study that need to be addressed. First, this study accessed only 31 effect sizes to calculate the overall effect of chatbots on learning. The use of chatbots

in education is a developing area that needs more experimental studies to investigate its impact on learning (Deveci Topal et al., 2021; Vázquez-Cano et al., 2021; Yin et al., 2021). There has been a sharp increase in related publications since 2020 (Hwang & Chang, 2021; Kuhail et al., 2023) owing to the availability of different visual chatbot development tools that do not require programming skills and the emergence of large language models. With the accumulation of a considerable amount of new empirical evidence, a further meta-analysis can be conducted in the following years. Second, data for this meta-analysis were collected before ChatGPT, an advanced language model performing various complex tasks, was launched. The next meta-analysis can also include empirical studies on ChatGPT and compare its effect with that of prior chatbots. Third, some categories (e.g. the practice of skills and abilities and virtual companions) in moderator analysis included only the studies published by a specific author (Kim, 2016, 2018a, 2018b; Kim et al., 2021). There is a need for more studies conducted by different researchers on these categories to draw more generalizable and strong conclusions about the effect of these moderators. Fourth, moderator analysis in this study focused on six variables. The effect of more variables, especially concerning the learner characteristics of the students (e.g. attitudes toward technology, trait emotions, and self-regulated skills) and quality of interaction between chatbots and learners, can be investigated in future studies (Winkler & Söllner, 2018).

The characteristics of the reviewed studies also indicate several research gaps in the use of chatbots in education. First, a low number of studies integrated chatbots into the K-12 education level, and experimental periods were mostly limited to one session. This meta-analysis calls for more research that can investigate the long-term effects of chatbots, especially with young learners or children. Second, most of the reviewed studies were related to foreign language education. The affordances of the chatbots in providing a conversation environment where learners can practice the target language by receiving individualized feedback and feeling less anxiety might have brought about more use of this technology in the foreign language context (Ji et al., 2023). However, conversation-based learning activities with chatbots can also be an impetus to enhance learning in other subject domains. Guiding inquiry activities in science education (Chang et al., 2023), stimulating communication with patients in medical education (Kron et al., 2017), and helping to debug erroneous codes in programming education (Yilmaz & Yilmaz, 2023), chatbots can be a learning partner in different subject domains. However, more empirical studies are needed to explore and discuss the growing potential of chatbots in these domains. Third, most chatbots were task-based and developed for specific learning content. Owing to the emergence of large language models that can aid learning in different tasks, further studies can provide more empirical evidence about the impact of general-purpose chatbots on learning to go beyond current commentaries on their benefits and limitations. Finally, due to the limited number of chatbots that mainly serve for motivation and simulation, future studies can focus on how well chatbots perform these tasks to enhance learning.

6. Conclusion

The focus of chatbot research in education is shifting from chatbot development to the evaluation of its use and effectiveness (Han & Lee, 2022). Existing empirical evidence provides conflicting results regarding the impact of chatbots on learning, which leads to questions concerning to what extent and when chatbots can contribute to learning (Deng & Yu, 2023; Huang et al., 2022; Zhang, Shan, et al., 2023). This comprehensive meta-analysis study quantitively synthesizing prior related research reveals the medium effect of chatbots on learning. In addition, it shows that their effect can change based on the type of instruction in the comparison group, experimental duration, chatbot type, and chatbot task. Specifically, moderator analysis results suggest that the highest learning benefits can be obtained when students (1) do not have specific support for task performance or access to complementary instruction, (2) use chatbots for only one learning session, (3) utilize task-focused chatbots, and (4) receive responses from chatbots for

FAQs. Overall, this study has enhanced the understanding of the overall effect of chatbots on learning and optimal conditions to promote their impact. Practical and research implications of the study can also guide future directions to contribute to the growing body of evidence on chatbots in education.

Acknowledgements

The author completed this research during her postdoctoral fellowship funded by the Alexander von Humboldt Foundation. She is grateful to this foundation for providing financial support for her studies. She also expresses her gratitude to Dr. Merve Basdogan, who provided enormous help in the analysis stage of this study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Ecenaz Alemdag is a postdoctoral fellow at Technische Universität Dresden in Germany. She received her Ph.D. degree in Computer Education and Instructional Technology from Middle East Technical University in Turkey. Her research interests include feedback, interactive online learning, multimedia learning, teacher education, and user experience.

ORCID

Ecenaz Alemdag http://orcid.org/0000-0003-2645-4732

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Bibauw, S., Van den Noortgate, W., François, T., & Desmet, P. (2022). Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology*, *26*(1), 1–25. Retrieved from https://hdl.handle.net/10125/73488

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Bozkurt, A. (2023). Generative artificial intelligence (AI) powered conversational educational agents: The inevitable paradigm shift. *Asian Journal of Distance Education*, *18*(1), 198–204. https://doi.org/10.5281/zenodo.7716416

*Brachten, F., Brünker, F., Frick, N. R., Ross, B., & Stieglitz, S. (2020). On the ability of virtual agents to decrease cognitive load: An experimental study. *Information Systems and e-Business Management*, *18*(2), 187–207. https://doi.org/10.1007/s10257-020-00471-7

*Chang, C. Y., Hwang, G. J., & Gau, M. L. (2022). Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. *British Journal of Educational Technology*, *53*(1), 171–188. https://doi.org/10.1111/bjet.13158

*Chang, C. Y., Kuo, S. Y., & Hwang, G. H. (2022). Chatbot-facilitated nursing education. *Educational Technology & Society*, *25*(1), 15–27.

Chang, J., Park, J., & Park, J. (2023). Using an artificial intelligence chatbot in scientific Inquiry: Focusing on a guided-inquiry activity using Inquirybot. *Asia-Pacific Science Education*, *9*(1), 44–74. https://doi.org/10.1163/23641177-bja10062

Chiu, T. K., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2023). Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot. *Interactive Learning Environments*, 1–17. https://doi.org/10.1080/10494820.2023.2172044

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2022). Analysing data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions version 6.3*. Cochrane.

*Deng, H., Liu, M., Su, R., & Dang, X. (2018). The design and empirical study of an online dialogic teaching model. In *Proceeding of the 2018 Seventh International Conference of Educational Innovation through Technology (EITT)* (pp. 234–238). https://doi.org/10.1109/EITT.2018.00054

Deng, X., & Yu, Z. (2023). A meta-analysis and systematic review of the effect of chatbot technology use in sustainable education. *Sustainability*, *15*(4), 2940. https://doi.org/10.3390/su15042940

*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/s10639-021-10627-8

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 1–15. https://doi.org/10.1080/14703297.2023.2195846

*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

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior*, *75*, 461–468. https://doi.org/10.1016/j.chb.2017.05.045

Garcia Brustenga, G., Fuertes-Alpiste, M., & Molas-Castells, N. (2018). *Briefing paper: Chatbots in education*. eLearn Center. Universitat Oberta de Catalunya.

*Graesser, A. C., Jackson, G. T., Mathews, E. C., Mitchell, H. H., Olney, A., & Ventura, M. (2003). Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In R. Alterman & D. Hirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1–5) Cognitive Science Society.

Grudin, J., & Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). https://doi.org/10.1145/3290605.3300439

Han, S., & Lee, M. K. (2022). FAQ chatbot and inclusive learning in massive open online courses. *Computers & Education*, *179*, 104395. https://doi.org/10.1016/j.compedu.2021.104395

Hannafin, M., Land, S. M., & Oliver, K. (1999). Open learning environments: Foundations, methods, and models. In C. Reigeluth (Ed.), *Instructional design theories and models* (pp. 115–140). Lawrence Erlbaum Associates.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of 800 meta-analyses relating to achievement*. Routledge.

*Hayashi, Y. (2020). Gaze awareness and metacognitive suggestions by a pedagogical conversational agent: An experimental investigation on interventions to support collaborative learning process and performance. *International Journal of Computer-Supported Collaborative Learning*, *15*(4), 469–498. https://doi.org/10.1007/s11412-020-09333-3

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Higgins, J. P. T., Li, T., & Deeks, J. J. (2021). Choosing effect measures and computing estimates of effect. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions version 6.2*. Cochrane.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical Research ed.)*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

*Hoque, M. E., Courgeon, M., Martin, J.-C., Mutlu, B., & Picard, R. W. (2013). Mach: My automated conversation coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 697–706). https://doi.org/10.1145/2493432.2493502

*Hsu, M. H., Chen, P. S., & Yu, C. S. (2021). Proposing a task-oriented chatbot system for EFL learners speaking practice. *Interactive Learning Environments*, 1–12. https://doi.org/10.1080/10494820.2021.1960864

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*, 1–14. https://doi.org/10.1080/10494820.2021.1952615

Jasin, J., Ng, H. T., Atmosukarto, I., Iyer, P., Osman, F., Wong, P. Y. K., Pua, C. Y., & Cheow, W. S. (2023). The implementation of chatbot-mediated immediacy for synchronous communication in an online chemistry course. *Education and Information Technologies*, *28*, 10665–10690. https://doi.org/10.1007/s10639-023-11602-1

*Jeon, J. (2021). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2021.1987272

Jeon, J. (2022). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning*, 1–26. https://doi.org/10.1080/09588221.2021.2021241

Ji, H., Han, I., & Ko, Y. (2023). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, *55*(1), 48–63. https://doi.org/10.1080/15391523.2022.2142873

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

*Kerly, A., Ellis, R., & Bull, S. (2009). Conversational agents in e-learning. In T. Allen, R. Ellis, & M. Petridis (Eds.), *Applications and innovations in intelligent systems XVI* (pp. 169–182). Springer.

*Kim, N. Y. (2016). Effects of voice chat on EFL learners' speaking ability according to proficiency levels. *Multimedia-Assisted Language Learning*, *19*(4), 63–88.
*Kim, N. Y. (2018a). A study on chatbots for developing Korean college students' English listening and reading skills. *Journal of Digital Convergence*, *16*(8), 19–26.
*Kim, N. Y. (2018b). Chatbots and Korean EFL students' English vocabulary learning. *Journal of Digital Convergence*, *16*(2), 1–7.
Kim, H.-S., Kim, N. Y., & Cha, Y. (2021). Is it beneficial to use ai chatbots to improve learners' speaking performance? *The Journal of AsiaTEFL*, *18*(1), 161–178. https://doi.org/10.18823/asiatefl.2021.18.1.10.161
Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1
*Kron, F. W., Fetters, M. D., Scerbo, M. W., White, C. B., Lypson, M. L., Padilla, M. A., Gliva-McConvey, G. A., Belfore, L. A.II, West, T., Wallace, A. M., Guetterman, T. C., Schleicher, L. S., Kennedy, R. A., Mangrulkar, R. S., Cleary, J. F., Marsella, S. C., & Becker, D. M. (2017). Using a computer simulation for teaching communication skills: A blinded multisite mixed methods randomized controlled trial. *Patient Education and Counseling*, *100*(4), 748–759. https://doi.org/10.1016/j.pec.2016.10.024
Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 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*(1), 65. https://doi.org/10.1186/s41239-021-00302-w
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310
Lee, J. Y., & Hwang, Y. (2022). A meta-analysis of the effects of using AI chatbot in Korean EFL education. *Studies in English Language & Literature*, *48*(1), 213–243. https://doi.org/10.21559/aellk.2022.48.1.011
Lin, H. (2019). Teaching and learning without a textbook: Undergraduate student perceptions of Open Educational Resources. *The International Review of Research in Open and Distributed Learning*, *20*(3), 1–18. https://doi.org/10.19173/irrodl.v20i4.4224
*Lin, M. P.-C., & Chang, D. (2020). Enhancing post-secondary writers' writing skills with a chatbot: A mixed-method classroom study. *Educational Technology & Society*, *23*(1), 78–92.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE.
Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901–918. https://doi.org/10.1037/a0037123
*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), 3239. https://doi.org/10.3390/app12073239
*Mejbri, N., Essalmi, F., & Rus, V. (2017). Educational system based on simulation and intelligent conversation. In *Proceedings of the 2017 6th International Conference on Information and Communication Technology and Accessibility* (pp. 1–6). https://doi.org/10.1109/ICTA.2017.8336020
Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/1094428106291059
Nghi, T. T., Phuc, T. H., & Thang, N. T. (2019). Applying Al chatbot for teaching a foreign language: An empirical research. *International Journal of Scientific & Technology Research*, *8*(12), 897–902.
Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, *2*, 100033. https://doi.org/10.1016/j.caeai.2021.100033
Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ (Clinical Research ed.)*, *372*, n160. https://doi.org/10.1136/bmj.n160
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*(6), 1549–1565. https://doi.org/10.1002/cae.22326
Petticrew, M., & Robert, H. (2006). *Systematic reviews in the social sciences A practical guide*. Blackwell Publishing.
Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638
*Ruan, S., Jiang, L., Xu, Q., Liu, Z., Davis, G. M., Brunskill, E., & Landay, J. A. (2021). Englishbot: An AI-powered conversational system for second language learning. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (pp. 434–444). https://doi.org/10.1145/3397481.3450648
Sáiz-Manzanares, M. C., Marticorena-Sánchez, R., Martín-Antón, L. J., Díez, I. G., & Almeida, L. (2023). Perceived satisfaction of university students with the use of chatbots as a tool for self-regulated learning. *Heliyon*, *9*(1), e12843. https://doi.org/10.1016/j.heliyon.2023.e12843

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Sage.
Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5–14. https://doi.org/10.3102/0013189X08314117
Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education*, *151*, 103862. https://doi.org/10.1016/j.compedu.2020.103862
*Song, D., & Kim, D. (2021). Effects of self-regulation scaffolding on online participation and learning outcomes. *Journal of Research on Technology in Education*, *53*(3), 249–263. https://doi.org/10.1080/15391523.2020.1767525
Suurmond, R., van Rhee, H., & Hak, T. (2017). Introduction, comparison, and validation of Meta-Essentials: A free and simple tool for meta-analysis. *Research Synthesis Methods*, *8*(4), 537–553. https://doi.org/10.1002/jrsm.1260
*Tegos, S., & Demetriadis, S. (2017). Conversational agents improve peer learning through building on prior knowledge. *Educational Technology & Society*, *20*(1), 99–111.
*Tegos, S., Demetriadis, S., Papadopoulos, P. M., & Weinberger, A. (2016). Conversational agents for academically productive talk: A comparison of directed and undirected agent interventions. *International Journal of Computer-Supported Collaborative Learning*, *11*(4), 417–440. https://doi.org/10.1007/s11412-016-9246-2
Terras, M. M., Ramsay, J., & Boyle, E. (2013). Learning and Open Educational Resources: A psychological perspective. *E-Learning and Digital Media*, *10*(2), 161–173. https://doi.org/10.2304/elea.2013.10.2.161
Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*(1), 1–24. https://doi.org/10.1186/s40561-023-00237-x
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), 1–20. https://doi.org/10.1186/s41239-021-00269-8
Vattøy, K. D. (2020). Teachers' beliefs about feedback practice as related to student self-regulation, self-efficacy, and language skills in teaching English as a foreign language. *Studies in Educational Evaluation*, *64*, 100828. https://doi.org/10.1016/j.stueduc.2019.100828
Vattøy, K. D., & Smith, K. (2019). Students' perceptions of teachers' feedback practice in teaching English as a foreign language. *Teaching and Teacher Education*, *85*, 260–268. https://doi.org/10.1016/j.tate.2019.06.024
Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11
*Wambsganss, T., Kueng, T., Söllner, M., & Leimeister, J. M. (2021). ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). https://doi.org/10.1145/3411764.3445781
*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. *Academy of Management Proceedings*, *2018*(1), 15903. https://doi.org/10.5465/AMBPP.2018.15903abstract
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*, 654924. https://doi.org/10.3389/frai.2021.654924
Xia, Q., Chiu, T. K., Chai, C. S., & Xie, K. (2023). The mediating effects of needs satisfaction on the relationships between prior knowledge and self-regulated learning through artificial intelligence chatbot. *British Journal of Educational Technology*, *54*(4), 967–986. https://doi.org/10.1111/bjet.13305
*Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner. *Computers & Education*, *161*, 104059. https://doi.org/10.1016/j.compedu.2020.104059
Xu, Z., Wijekumar, K., Ramirez, G., Hu, X., & Irey, R. (2019). The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology*, *50*(6), 3119–3137. https://doi.org/10.1111/bjet.12758
Yilmaz, R., & Yilmaz, F. G. K. (2023). Augmented intelligence in programming learning: Examining student views on the use of ChatGPT for programming learning. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100005. https://doi.org/10.1016/j.chbah.2023.100005
*Yin, J., Goh, T. T., Yang, B., & Xiaobin, Y. (2021). Conversation technology with micro-learning: The impact of chatbot-based learning on students' learning motivation and performance. *Journal of Educational Computing Research*, *59*(1), 154–177. https://doi.org/10.1177/0735633120952067
Zhang, S., Shan, C., Lee, J. S. Y., Che, S., & Kim, J. H. (2023). Effect of chatbot-assisted language learning: A meta-analysis. *Education and Information Technologies*, 1–21. https://doi.org/10.1007/s10639-023-11805-6
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*, 1–29. https://doi.org/10.1080/10494820.2023.2202704

Appendix A. Forest plot of effect sizes

#	Study name	Hedges' g	CI Lower limit	CI Upper limit	Weight
2	Chang et al. (2022a)	1.90	1.14	2.75	2.38%
3	Chang et al. (2022b)	1.71	0.93	2.59	2.32%
4	Deng et al. (2018)	0.07	-0.30	0.43	3.92%
5	Deveci-Topal et al. (2021)	0.34	-0.27	0.97	2.96%
6	Fidan and Gencel (2022), peer feedback vs. chatbot feedback	0.26	-0.13	0.65	3.84%
7	Graesser et al. (2003), read-textbook vs. AutoTutor	0.47	-0.35	1.32	2.30%
8	Hayashi (2020), no gaze	1.80	1.09	2.59	2.54%
9	Hayashi (2020), visible gaze	0.69	0.06	1.35	2.88%
10	Hoque et al. (2013), control vs. MACH+video+feedback	0.70	0.18	1.23	3.31%
11	Jeon (2021), control vs. CA-DA+CA-NDA	2.10	1.42	2.84	2.64%
12	Kim (2016), advanced level, peer vs. chatbot	-0.25	-0.87	0.35	3.02%
13	Kim (2016), beginning level, peer vs. chatbot	0.31	-0.33	0.95	2.91%
14	Kim (2016), intermediate level, peer vs. chatbot	-0.56	-1.21	0.06	2.93%
15	Kim (2018a), total score (listening and reading)	0.17	-0.24	0.58	3.75%
16	Kim (2018b)	0.20	-0.37	0.79	3.12%
17	Kim et al. (2021), face-to-face vs. AI-text and voice	0.26	-0.14	0.66	3.80%
18	Kron et al. (2017)	0.27	0.08	0.46	4.51%
19	Lin and Chang (2020)	0.04	-0.17	0.25	4.46%
20	Mageira et al. (2022)	0.06	-0.61	0.73	2.80%
21	Mejbri et al. (2018)	0.62	-0.20	1.48	2.32%
22	Nghi et al. (2019)	0.70	0.41	0.99	4.21%
23	Ruan et al. (2021, Study 1), vocabulary	0.50	-0.26	1.28	2.52%
24	Ruan et al. (2021, Study 2), vocabulary	0.02	-0.74	0.77	2.56%
25	Song and Kim (2021)	0.47	-0.06	1.01	3.27%
26	Tegos and Demetriadis (2017)	0.64	0.17	1.13	3.49%
27	Tegos et al. (2016), control and U+D treatment	0.71	0.28	1.16	3.64%
28	Wambsganss et al. (2021)	0.57	0.03	1.13	3.22%
29	Winkler et al. (2020), control (video only) vs. SARA	0.78	0.30	1.27	3.47%
30	Xu et al. (2021), human conversation vs. CA-conversation	-0.02	-0.52	0.47	3.43%
31	Yin et al. (2021)	0.03	-0.37	0.43	3.81%
32					
33					
34					
35