Journal of Computer Assisted Learning

WILEY

Journal of Computer Assisted Learning

REVIEW ARTICLE

ChatGPT Interventions in Higher Education: A Systematic Review of Experimental Studies

Yiwen Jin | Lies Sercu

Faculty of Arts, K.U. Leuven, Leuven, Belgium

Correspondence: Lies Sercu (lies.sercu@kuleuven.be)

Received: 29 September 2024 | **Revised:** 22 April 2025 | **Accepted:** 15 May 2025

Funding: This work was supported by the Chinese Government Scholarship (CSC), 202208510021.

Keywords: academic outcomes | ChatGPT | experimental intervention studies | higher education | research design

ABSTRACT

Background: Artificial intelligence has been reshaping many industries, and education is no exception. When ChatGPT burst onto the scene in late 2022, it ignited conversations among educators and researchers alike. Despite growing interest and experimentation with this technology in university classrooms, the field has been missing a comprehensive analysis of the actual research examining how ChatGPT is being used in higher education settings.

Objectives: This review addresses that gap by focusing on experimental designs and interventions that incorporate ChatGPT in higher education settings, with particular attention to its impact on academic outcomes.

Methods: This study followed the PRISMA guidelines to identify and analyse relevant experimental studies. Through systematic coding and analysis of 21 selected studies, this review examined research design, intervention types and academic outcomes.

Results and Conclusions: This review found that research on ChatGPT-based interventions in higher education has focused on language and STEM domains. Although many of the studies are true experimental designs, they often lack large, diverse samples. Two main types of interventions are identified: task assistance and general learning support, each with unique strengths and challenges. ChatGPT was generally effective for knowledge acquisition, but its impact on skills development varied. Medium-term interventions showed the best results, while short-term effects were limited and long-term outcomes were mixed.

1 | Introduction

1.1 | Background

The rapid advancement of artificial intelligence (AI) technologies has sparked significant interest in their potential applications across various fields, including education (Dempere et al. 2023; Labadze et al. 2023). Since the release of ChatGPT on November 30, 2022, institutions and educators have been exploring ways to integrate it into their curricula and pedagogical practices (Lu et al. 2024; Shang and Geng 2024). ChatGPT's ability to generate human-like text, answer complex queries, and engage in nuanced conversations has triggered both excitement and debate (Kasneci et al. 2023; Ansari et al. 2024). This growing interest is reflected in the increasing number of pilot projects, research and scholarly discussions focused on its integration into educational settings (Adeshola and Adepoju 2023; Urban et al. 2024).

The review of Lo (2023) highlighted both the great potential of ChatGPT in various educational contexts and the challenges associated with its use. Several critical concerns have emerged regarding the integration of ChatGPT in educational settings. First, there are concerns about privacy and ethical data collection, as interactions with ChatGPT may enable the collection of personal data to build user profiles, potentially leading to privacy breaches and surveillance issues

Summary

- What is already known about this topic:
  - ChatGPT is increasingly integrated into higher education.
  - Previous reviews lack emphasis on experimental intervention studies.
- What does this paper add:
  - Gives a comprehensive overview of experimental intervention studies on ChatGPT in higher education.
  - Identifies two main intervention types: task assistance and general learning support.
  - Finds that ChatGPT enhances knowledge acquisition, but its impact on skill development varies.
  - Shows that medium-term interventions (1 day–3 months) are most effective.
- Implications for practice/or policy:
  - Improve research design with larger and more diverse samples.
  - Expand applications beyond language and STEM to other disciplines.
  - Refine interventions by reducing ChatGPT reliance and adding non-textual tasks.
  - Bridge knowledge and skills by integrating hands-on learning.

(Dempere et al. 2023). Second, the risk of bias and discrimination presents a serious challenge, as AI systems trained on existing data may perpetuate systemic biases and unfairness, particularly affecting students from historically marginalised groups (Dempere et al. 2023). Third, a major challenge lies in ChatGPT's tendency to generate false or inaccurate information with apparent confidence, as it may present incorrectly formatted information or fabricated content rather than acknowledging knowledge gaps (Meyer et al. 2023). Finally, the over-reliance on AI can lead to diminished creativity and critical thinking abilities (Zhai et al. 2024).

Alongside these concerns, researchers have also identified considerable benefits of ChatGPT in higher education, with several key roles and functions it could fulfil. Ansari et al. (2024) summarised four main roles of ChatGPT in higher education: teaching assistant, personalised tutor, assessment partner and co-researcher. Similarly, Kasneci et al. (2023) summarised that large language models can support university students by aiding in research, writing and the development of critical thinking and problem-solving skills, through generating summaries, organising ideas and providing valuable insights and resources on various topics. In specific teaching practices, for example, ChatGPT offers assistance in language classes by helping students detect and correct grammar and vocabulary mistakes while also proposing alternative sentence constructions to improve both the quality and coherence of their writing (Song and Song 2023; Boudouaia et al. 2024). In STEM fields, it has been used to explain difficult concepts and solve problems (Kosar et al. 2024; Sun et al. 2024). Although it is generally recognised that AI has great potential, the question of how to effectively integrate it into the field of education is still in the initial exploration stage (Alnaqbi and Fouda 2023).

Yet, several reviews have explored ChatGPT and AI-generated content (AIGC) in education, each with a distinct focus (see Table 1). Montenegro-Rueda et al. (2023) analysed 12 articles to examine the effects of ChatGPT in education since its launch, focusing on aspects such as geographical distribution, methodology, main findings and emerging trends. Lo (2023) conducted a rapid review of 50 articles, summarising ChatGPT's performance across different subject domains, its potential to enhance teaching and learning, and the associated challenges and solutions. Zhang and Tur (2024) provided a SWOT analysis of ChatGPT's use in K-12 education, discussing current practices, future directions and recommendations for its implementation. Chen et al. (2024) took a broader approach, reviewing 134 publications on AIGC's educational application to map out the development status and future trends. Ansari et al. (2024) reviewed 69 published studies on the use of ChatGPT in higher education from contextual, methodological and disciplinary perspectives and found that only 19% of these studies were empirical. Baig and Yadegaridehkordi (2024) also focused on higher education, reviewing 57 articles published between January 2023 and mid-January 2024. They present an overview of ChatGPT trends, key applications and constraints in the field of higher education. Nevertheless, despite the existence of many related review studies, none of them have a particular focus on experimental studies with an educational intervention.

This systematic review differs from previous studies in that: (1) It narrows the focus to intervention-based experimental studies in higher education, emphasising the analysis of research designs and the outcomes of these interventions, whereas previous reviews have often taken a broader approach, including both theoretical studies, experimental intervention studies, or opinion-based studies; and (2) it employs a targeted examination of the impact of one particular too, namely ChatGPT on academic outcomes, with a specific focus on whether ChatGPT-based interventions lead to more favourable outcomes compared to traditional interventions; (3) it covers research published between 2022 and August 2024, a period that allows for the inclusion of more mature experimental studies, as sufficient time has passed for data collection, analysis and publication. Given this timely focus, we believe now is an appropriate time for a systematic review centered on ChatGPT related experimental intervention studies.

By offering a more focused and in-depth analysis of the actual use of ChatGPT in higher education, this review not only addresses an important gap in the existing literature but also equips educators and researchers with actionable insights into how to implement ChatGPT more effectively within academic settings. Specifically, this systematic review conducts comparative analyses of research designs, categorises intervention types, and—most importantly—assesses academic outcomes by analysing whether students in ChatGPT-based interventions perform better or worse compared with those in control groups using traditional methods. Therefore, this systematic review was guided by the following research questions:

RQ1. How are experimental studies on ChatGPT in higher education designed?

RQ2. How is ChatGPT used in educational interventions?

TABLE 1 | Overview of the previous related review studies.

Authors	Types of review	Focus	Level of education	Time frame
Montenegro-Rueda et al. (2023)	Systematic review	Summarising geographical distribution, methodology, findings and emerging trends.	General education	Until June 2023
Lo (2023)	Rapid review	Summarising ChatGPT's performance across subject domains, teaching/learning potential and challenges/solutions	General education	February 2023
Zhang and Tur (2024)	Systematic review	Provided a SWOT analysis of ChatGPT's use in K-12 education, discussing current practices, future directions and implementation recommendations.	K12 education	Until August 2023
Chen et al. (2024)	Systematic review	Reviewed 134 publications on AIGC's educational application, mapping development status and future trends.	General education	November 2022–August 2023
Ansari et al. (2024)	Systematic review	Focused on higher education core processes (teaching, learning, assessment and research)	Higher education	Until May 2023
Baig and Yadegaridehkordi (2024)	Systematic review	Analysed the application, research trends, adoption strategies, implementation and limitations of ChatGPT in higher education.	Higher education	January 2023—mid-January 2024

RQ3. How do ChatGPT-based interventions compare to traditional methods in terms of their effects on academic outcomes?

2 | Method

This study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al. 2021) to identify and analyse relevant experimental studies. Through systematic coding and analysis of 21 selected studies, this review examined research design, intervention types and academic outcomes.

2.1 | Search Strategy

In order to identify the relevant studies for this systematic review, a thorough search was carried out in the two leading academic databases: Web of Science and Scopus. These databases were chosen because they index a wide range of high-impact, peer-reviewed journals to ensure that the selected studies meet rigorous academic standards. The selection aligns with best practices in systematic reviews, as evidenced by Heck et al. (2024), who found that a combination of high-quality multidisciplinary databases provides the most effective and comprehensive coverage for interdisciplinary topics.

The search was performed using the following keywords: ('ChatGPT') AND ('Education' OR 'Teach*' OR 'Learn*') AND ('Student' OR 'Undergraduate' OR 'College' OR 'University'). We deliberately did not use 'GenAI' or other 'AI' tools, such as Claude, because we wanted to limit the retrieved texts to literature on ChatGPT, which is the most frequently used tool in education. No temporal constraints were imposed, to ensure the retrieval of all pertinent literature published since the release of ChatGPT. Furthermore, filters were applied to limit the search to 'articles' and 'English-language publications'. The restriction to 'articles' was set to exclude conference papers, book chapters and other types of publications, although limiting the language to English ensured accessibility.

2.2 | Inclusion and Exclusion Criteria

The studies included in this review were selected based on the inclusion and exclusion criteria in Table 2. These criteria were designed to focus exclusively on experimental studies with well-defined educational interventions that aim to improve academic achievement and provide measurable learning outcomes. The rationale behind this is as follows: Experimental studies allow for stronger causal inferences regarding the outcomes of ChatGPT interventions, distinguishing them from merely correlational or observational studies. This approach also aligns with our research objective, which seeks to evaluate the direct impact of ChatGPT-based interventions on students' academic performance (RQ3). Based on these considerations, studies had to focus on higher education students and use ChatGPT during the educational intervention. Furthermore, the presence of a control group not using ChatGPT was required to effectively assess its impact.

TABLE 2 | Inclusion and exclusion criteria for study selection.

Inclusion criteria	Exclusion criteria
1. Studies that involve a specific intervention aimed at improving students' academic outcomes.	1. Studies that do not involve a specific intervention aimed at improving academic outcomes.
2. Studies focusing on higher education students.	2. Studies not focusing on higher education students.
3. Studies using ChatGPT in educational interventions.	3. Studies using tools other than ChatGPT in interventions.
4. Studies reporting measurable academic learning outcomes.	4. Studies without measurable academic outcomes.
5. Studies including at least one control group not using ChatGPT.	5. Experiments without a non-ChatGPT control group.

Together, these selection criteria allowed for a systematic assessment of ChatGPT's application.

2.3 | Screening Process

The search process was finalised in August 2024, yielding a total of 1584 articles (536 from Web of Science and 1048 from Scopus). After removing 434 duplicate records, 1150 unique articles remained.

The 1150 articles underwent a two-stage screening process. First, the abstracts of the articles were reviewed, resulting in the exclusion of 1101 articles that did not meet the inclusion criteria. This left 49 articles for full-text screening. The full texts of the remaining 49 articles were assessed, leading to the exclusion of 27 additional articles that did not meet the inclusion criteria.

Ultimately, 21 articles met all inclusion criteria and were included in the final review. The PRISMA flowchart in Figure 1 illustrates the entire search and screening process.

2.4 | Quality Appraisal

The 21 included articles were quality assessed using the Mixed Methods Appraisal Tool (MMAT) for the quality appraisal process (Hong et al. 2018). The MMAT is a tool for assessing the quality of empirical studies, specifically primary research based on experiments, observations, or simulations (Hong et al. 2018). The included articles were divided into quantitative randomised controlled trials and quantitative nonrandomised studies according to the experimental design and were evaluated according to the two screening questions and five methodological quality criteria. The results showed that these 21 articles met the minimum quality criteria in terms of methodology and analysis and could be included for further review. The detailed results are included in Appendix A Tables A1 and A2.

2.5 | Coding Scheme

After having decided on whether to include or exclude the articles given their quality, the 21 articles were coded according to a systematic coding scheme. This coding scheme was developed deductively, based on the predefined research questions (RQ1–RQ3), in order to ensure that the analysis remained closely aligned with the main objectives of the review (Bingham and Witkowsky 2022). Figure 2 presents the coding scheme, illustrating the relationship between the research questions and the corresponding coding categories. Appendix B provides a detailed mapping of each article to its assigned codes.

3 | Results

3.1 | RQ1: How Are Experimental Studies on ChatGPT in Higher Education Designed?

To understand the research design of the included studies, three types of codes were applied: types of experimental design, sample size and assessment timing.

3.1.1 | Research Design

As shown in Table 3, the classification of experimental designs is based on Creswell's guidelines (2015, 307). True experiments, which constitute 57.14% of the included studies, represent the most rigorous and robust experimental approach due to their implementation of random assignment, ensuring group equivalence (Creswell 2015, 309). In contrast, quasi-experiments were used in 42.86% of studies. These designs lack random assignment of participants but are often set in natural educational settings where randomisation is impractical.

3.1.2 | Sample Size

The range of the sample sizes and the corresponding number of articles are displayed in Table 4. According to Creswell (2015), 146, in educational experimental research, each group should have at least 15 participants, and preferably more for reasons of reliability. However, not all of the studies reviewed met this requirement. Araji and Brooks (2024) had only 19 participants in total, with 9 and 10 participants in each group, respectively. Dasari et al. (2024) and Li et al. (2024) also had fewer than 15 participants in each group.

The total sample sizes varied across studies. Small sample sizes (1–50) were found in 7 studies, reflecting common constraints in educational research such as limited access to participants and resources. Medium sample sizes (51–100) were more common, occurring in 11 studies, indicating a balance between feasibility and statistical power. Large samples (101–500) were rare, with only 2 studies in this category, and very large samples (501+)

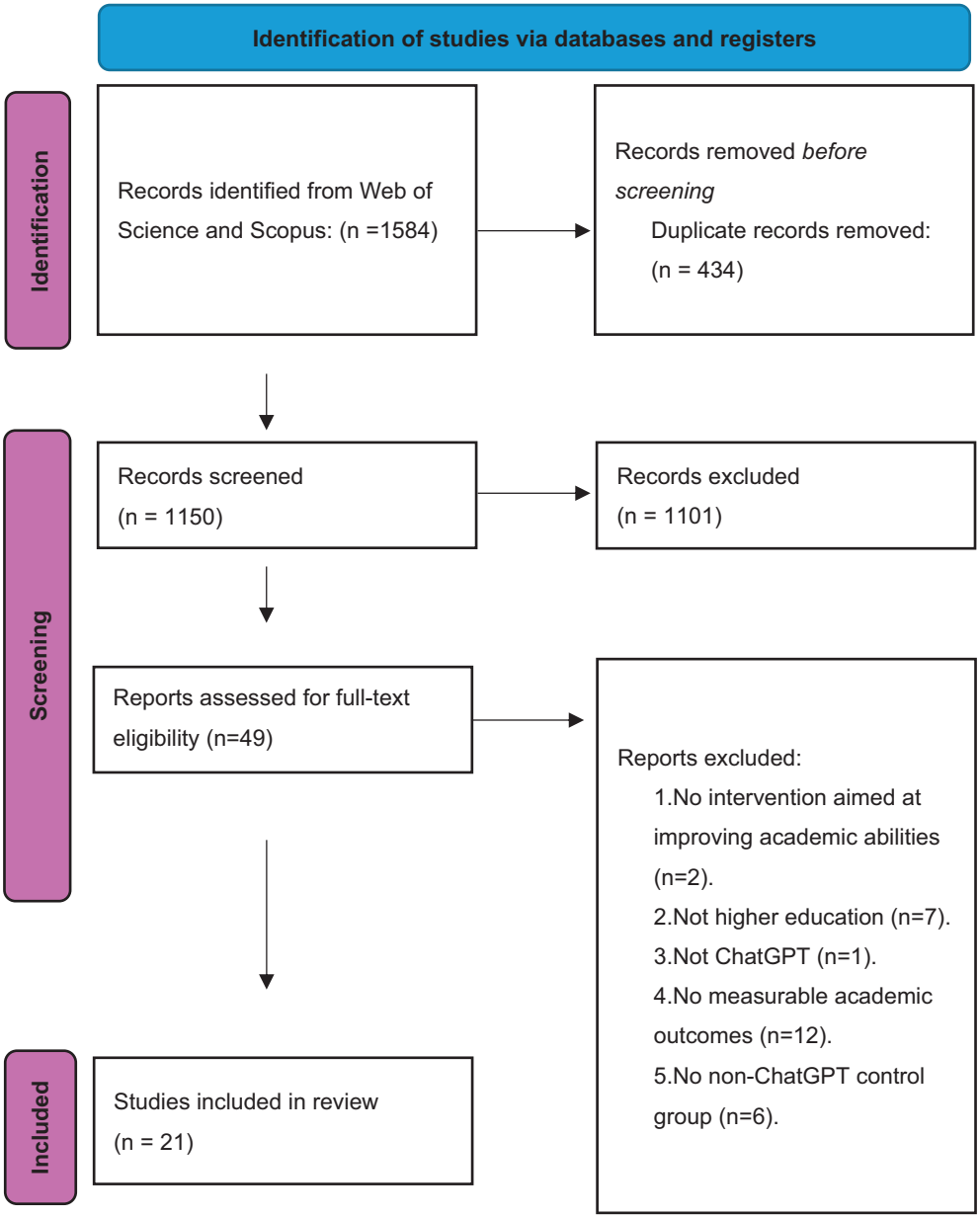

FIGURE 1 | The PRISMA flowchart.

were equally rare, represented by a single study. These findings suggest that while larger samples can provide more generalizable results, they are often difficult to achieve in educational settings due to logistical and resource constraints. However, when a non-randomised quasi-experimental study uses a small sample size, doubt can be raised about the validity of such studies.

3.1.3 | Assessment Timing

From Table 5 it can be seen that most studies (71.43%) used a pretest–posttest design, while only 28.57% adopted a posttest-only design. This indicates that, regarding experimental design, researchers generally prefer pretest–posttest approaches to capture changes before and after the intervention, thereby providing stronger evidence for the intervention's effect. The pretest–posttest design helps control for individual baseline differences and enhances internal validity, whereas the posttest-only design may be more vulnerable to confounding factors.

3.2 | RQ2: How Is ChatGPT Used in Educational Interventions?

The educational interventions discussed in the included articles represent diverse application scenarios. To better understand them, this review coded and analysed the interventions across five dimensions: discipline, duration, types of interventions, role of ChatGPT and interaction mode.

3.2.1 | Discipline Focus

ChatGPT-based interventions in higher education have been applied across four main disciplines: Language and Literature,

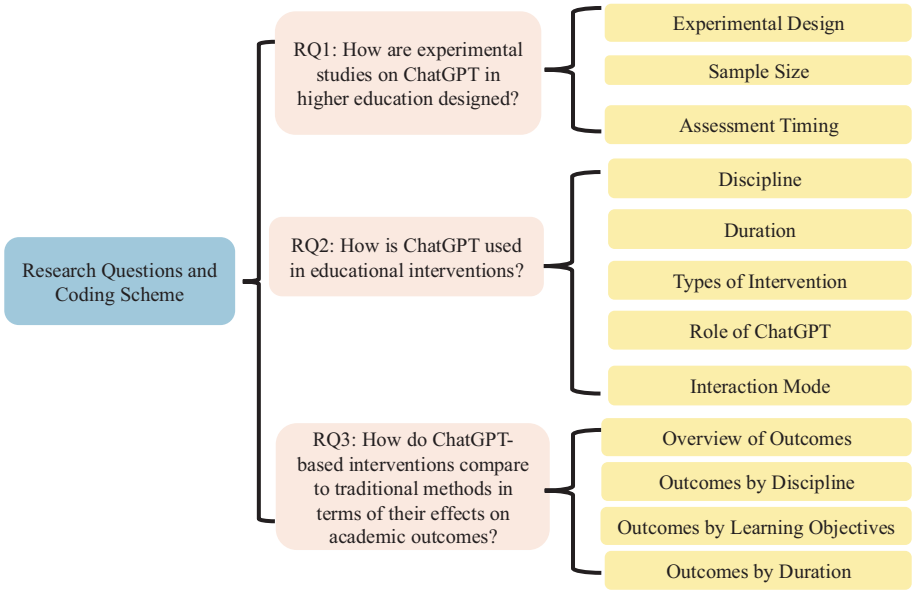

FIGURE 2 | Research questions and coding scheme.

TABLE 3 | Types of experimental design.

Experimental design	Total (%)
True experiment	12 (57.14%)
Quasi-experiment	9 (42.86%)

TABLE 4 | Sample size.

Sample size	Total (%)
Small (1–50)	7 (33.33%)
Medium (51–100)	11 (52.38%)
Large (101–500)	2 (9.52%)
Very large (501+)	1 (4.76%)

TABLE 5 | Assessment timing.

Assessment timing	Total (%)
Pretest and posttest	15 (71.43%)
Posttest only	6 (28.57%)

Science, Technology, Engineering, and Mathematics (STEM), Health Sciences and Music. Table 6 shows the number of studies in each category. In the language and literature category ($n = 8$), subjects included English writing, English grammar, Thai language writing, digital storytelling and creative writing. The STEM fields ($n = 7$) covered mathematics, programming, educational technology, STEM teaching training and statistical reasoning. In health education ($n = 5$), the studies focused on dental, pharmacy, surgery, nursing, clinical medicine and medical terminology. Finally, there was one study in the field of music ($n = 1$), focusing on xyz. These findings indicate that the current body of research is primarily focused on language and literature as well as STEM education.

3.2.2 | Types of Intervention

The interventions in the included studies are categorised into two types: task assistance and general learning support. Table 7 presents a detailed comparison of these two intervention types, although Table 8 lists the studies classified under each category.

The first type, task assistance, including 13 papers, refers to ChatGPT assisting students in completing specific tasks or learning activities. This type of intervention focuses on the process of students solving subject-specific problems that lead to immediate, concrete outcomes. These tasks are closely related to the course content, including preparing a presentation (Kavadella et al. 2024), composing an essay (Niloy et al. 2024), completing a radar chart task (Sun et al. 2024), searching for information and completing tasks related to surgical topics (Araji and Brooks 2024), or completing exercises focused on digital storytelling (Avello et al. 2024). In addition, many studies on second language writing instruction used task assistance intervention. Song and Song (2023) provided students with AI-assisted writing instruction using ChatGPT during a 12-week period. Tsai et al. (2024) asked their students to use ChatGPT to revise their original essays, then submitting a revised version and a detailed revision report after receiving AI assistance.

The second type, general learning support, means that ChatGPT supports students' individual learning needs and helps them to be autonomous in their learning process. This type of intervention usually allows students to use ChatGPT freely to learn according to their individual progress and

TABLE 6 | Discipline focus.

Discipline	Authors	Total (%)
Language and literature	Avello et al. (2024); Boudouaia et al. (2024); Escalante et al. (2023); Kucuk (2024); Niloy et al. (2024); Song and Song (2023); Tsai et al. (2024); Wiboolyasarin et al. (2024)	8 (38.10%)
STEM	Dasari et al. (2024); Huesca et al. (2024); Kosar et al. (2024); Li (2023); Li et al. (2024); Sun et al. (2024); Wahba et al. (2024)	7 (33.33%)
Health education	Araji and Brooks (2024); Hsu (2024); Kavadella et al. (2024); Svendsen et al. (2024); Wu et al. 2025)	5 (23.81%)
Music	Zhou and Kim (2024)	1 (4.76%)

TABLE 7 | Comparison of task assistance and general learning support interventions.

Criteria	Task assistance	General learning support
Definition	ChatGPT assists students in completing specific tasks or learning activities	ChatGPT supports students' individual learning needs and autonomous learning process
Focus	Process of solving subject-specific problems with immediate, concrete outcomes	Supporting personalised learning according to individual progress and needs
Structure	Structured, task-oriented interventions with clear objectives	Flexible, student-driven learning with less structure
Duration	Often shorter, task-specific timeframes	Generally longer-term, continuous support
Student autonomy	More guided, teacher-directed	Higher degree of student autonomy

TABLE 8 | Types of intervention with ChatGPT.

Types of intervention	Authors	Total (%)
Task assistance	Araji and Brooks (2024); Avello et al. (2024); Kavadella et al. (2024); Li et al. (2024); Kosar et al. (2024); Sun et al. (2024); Wahba et al. (2024); Boudouaia et al. (2024); Escalante et al. (2023); Niloy et al. (2024); Song and Song (2023); Tsai et al. (2024); Wiboolyasarin et al. (2024)	13 (61.90%)
General learning support	Dasari et al. (2024); Hsu (2024); Kucuk (2024); Svendsen et al. (2024); Huesca et al. (2024); Li (2023); Zhou and Kim (2024); Wu et al. (2025)	8 (38.10%)

needs. For example, in Kucuk's (2024) study, the experimental group used ChatGPT on their mobile devices for grammar learning for 7 weeks to seek assistance with grammar questions and difficulties. Or in flipped classrooms, students may interact with ChatGPT before the class to study new concepts and to obtain examples (Huesca et al. 2024) or ChatGPT could provide students with learning tasks and personalised guidance (Li 2023).

3.2.3 | Role of ChatGPT

The functions of ChatGPT used in the intervention were coded into the following five categories, representing the five roles of ChatGPT: content generator, information retriever, feedback and evaluator, learning companion and personalised learning guide. Table 9 shows the example of the role and the corresponding sample article. In one intervention, multiple roles of ChatGPT may be involved, so we only listed representative examples of each role. These roles show that ChatGPT is flexible enough to support a variety of tasks, from generating writing drafts and code to providing writing feedback and assisting with data analysis.

3.2.4 | Duration

According to Creswell (2015), 324, interventions must last long enough and be strong enough to really have an impact on outcomes. Table 10 therefore summarises the duration of the intervention in the included studies. The durations were categorised as short-term, medium-term and long-term based on the intervention period. Six studies were categorised as short-term studies involving either a single class session or conducted within 1 day. The majority of the studies (57.14%) were medium-term studies, with durations ranging from a few weeks to 3 months, including 3 weeks, 6 weeks, 2 months, 1 month, 10 weeks and 12 weeks. Finally, 3 studies were

TABLE 9 | Role of ChatGPT in interventions.

Role	Example	Example
Content generator	In academic and creative writing, students use ChatGPT to generate first drafts, revise paragraphs and adjust grammar and structure.	Boudouaia et al. (2024); Niloy et al. (2024); Tsai et al. (2024)
	In the programming task, students use ChatGPT to generate code and perform debugging and optimization	Kosar et al. (2024); Sun et al. (2024)
	In content creation, ChatGPT is used to generate PowerPoint presentations or project reports.	Kavadella et al. (2024)
Information retriever	In medical terminology learning, students use ChatGPT to search for term definitions and obtain relevant background information.	Hsu (2024)
	In programming and data analysis tasks, students search for steps and methods to solve problems through ChatGPT.	Kosar et al. (2024); Wahba et al. (2024)
Feedback and evaluator	In academic writing, students use ChatGPT to receive multiple rounds of writing feedback and get advice on grammar, sentence structure and paragraph development, among others.	Boudouaia et al. (2024); Escalante et al. (2023); Tsai et al. (2024)
	In language learning, ChatGPT provides suggestions for improving grammar, style and expression.	Wiboolyasarin et al. (2024)
	In programming courses, ChatGPT provides code feedback to help students optimise program designs.	Kosar et al. (2024)
Learning companion	In medical internships, students use ChatGPT to discuss cases, simulate diagnostic conversations and help analyse conditions and design treatment plans.	Wu et al. (2025)
	In collaborative learning projects, students interact with ChatGPT to explore problems and validate ideas.	Li et al. (2024)
Personalised learning guide	In a flipped classroom, students use ChatGPT to preview new concepts and plan pre-class learning content.	Huesca et al. (2024); Li (2023)
	In a STEM teaching design project, students use ChatGPT to get guidance and adjust the order and completion of tasks according to project needs.	Li et al. (2024)

TABLE 10 | Duration of interventions.

Duration	Authors	Total (%)
Short-term (within 1 day)	Araji and Brooks (2024); Avello et al. (2024); Niloy et al. (2024); Sun et al. (2024); Svendsen et al. (2024); Tsai et al. (2024)	6 (28.57%)
Medium-term (more than 1 day, up to 3 months)	Boudouaia et al. (2024); Escalante et al. (2023); Hsu (2024); Huesca et al. (2024); Kavadella et al. (2024); Kucuk (2024); Li (2023); Li et al. (2024); Song and Song (2023); Wahba et al. (2024); Wiboolyasarin et al. (2024); Zhou and Kim (2024)	12 (57.14%)
Long-term (more than 3 months)	Dasari et al. (2024); Kosar et al. (2024); Wu et al. (2025)	3 (14.29%)

long-term studies, each spanning one semester. This categorisation suggests that most of the experimental studies were conducted over a medium-term period.

3.2.5 | Interaction Mode

Based on the scenarios in which students interact with ChatGPT, the interaction modes were divided into four categories: in class, out of class, blended and indirect, as shown in Table 11.

The 'in-class mode' ($n = 8$) involves the use of ChatGPT within a classroom setting, where interventions are conducted under the direct guidance and supervision of teachers. The 'out-of-class mode' means the students use ChatGPT for self-study outside of class. In this mode, students have more autonomy, but also face problems of academic integrity and over-reliance. 'Blended mode' ($n = 11$), meaning a combination of in- and out-of-classroom modes. Students can use ChatGPT in or out of the classroom during a period of time. The final category, is 'indirect', which refers to the fact that students do not interact directly with ChatGPT, but only receive feedback from ChatGPT

TABLE 11 | Interaction mode.

Mode	Authors	Total
In class	Araji and Brooks (2024); Avello et al. (2024); Huesca et al. (2024); Niloy et al. (2024); Sun et al. (2024); Svendsen et al. (2024); Tsai et al. (2024); Wiboolyasarin et al. (2024)	8 (38.10%)
Out of class	Hsu (2024)	1 (4.76%)
Blended	Boudouaia et al. (2024); Dasari et al. (2024); Kavadella et al. (2024); Kosar et al. (2024); Kucuk (2024); Li (2023); Li et al. (2024); Song and Song (2023); Wahba et al. (2024); Wu et al. (2025); Zhou and Kim (2024)	11 (52.38%)
Indirect	Escalante et al. (2023)	1 (4.76%)

provided by instructors. Only one study, Escalante et al. (2023), falls into this category.

3.3 | RQ3: How Do ChatGPT-Based Interventions Compare to Traditional Methods in Terms of Their Effects on Academic Outcomes?

The outcomes of the included articles were coded, with a special focus on a comparison of the performance of the ChatGPT intervention group to that of the control group. The analysis was conducted across four key dimensions: overview of outcomes, outcomes by discipline, outcomes by learning objectives and outcomes by duration.

3.3.1 | Overview of Outcomes

To classify the learning outcomes, we examined whether the experimental groups using ChatGPT interventions demonstrated statistically significant differences compared with control groups employing traditional methods. We used four codes to classify the effects of ChatGPT interventions in the analysed studies: higher, lower, no and mixed.

1. Higher: The ChatGPT group showed a statistically significant improvement compared with the control group.
2. Lower: The ChatGPT group performed significantly worse than the control group.
3. No: There was no statistically significant difference between the ChatGPT and control groups.
4. Mixed: The results varied across different tests or measures within the same study.

This classification was applied consistently across all analysed studies to ensure clarity in reporting the outcomes of ChatGPT interventions.

Figure 3 demonstrates the outcomes among the included studies. Among the 21 experimental studies analysed, ChatGPT demonstrated mainly positive effects on learning outcomes. 52.4% of studies reported significantly higher performance in ChatGPT groups, 28.6% found no significant between-group differences, 9.5% observed significantly lower performance in ChatGPT groups, and 9.5% showed mixed results, where one test indicated significant improvement while another showed no improvement.

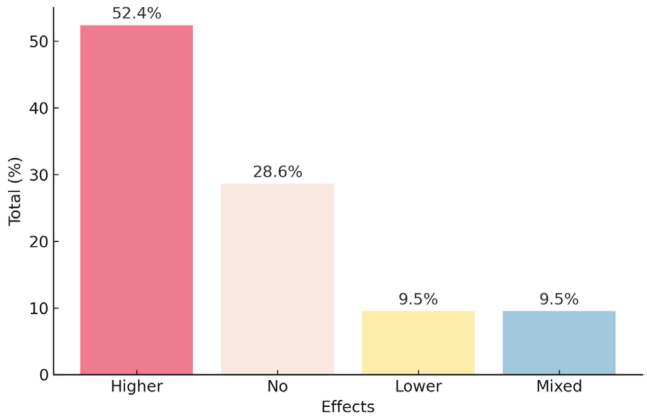

FIGURE 3 | Overview of outcomes.

3.3.2 | Outcomes by Discipline

As shown in Figure 4, this section examines the outcomes of ChatGPT-based intervention across different academic disciplines. In language and literature education, five out of eight studies reported significantly higher performance in ChatGPT-integrated groups. These positive outcomes were particularly evident in EFL writing, grammar learning and Thai language writing. One study (Niloy et al. 2024) found significantly lower performance in creative writing tasks, while two other studies (Avello et al. 2024; Escalante et al. 2023) reported no significant difference in digital storytelling and writing outcomes.

In STEM disciplines, three out of seven studies reported significantly higher performance with ChatGPT integration. In mathematics, Dasari et al. (2024) found that the group that only used ChatGPT to study showed significantly lower performance. In programming studies focusing on self-directed learning, both Kosar et al. (2024) and Sun et al. (2024) reported no significant differences in student performance. Li et al. (2024) reported mixed outcomes in STEM teaching training, with no significant difference in final proposal assessment but significantly higher performance in task-based learning activities.

In health education, two out of five studies reported significantly higher performance with ChatGPT integration (Hsu 2024;

Kavadella et al. 2024). Two studies found no significant differences, one in surgical medicine (Araji and Brooks 2024) and the other in pharmacy education (Svendsen et al. 2024). Wu et al. (2025) conducted a study in clinical medicine, assessing both final exam performance and clinical skills after the intervention and reported mixed outcomes.

The single study in music education demonstrated significantly higher performance in music knowledge acquisition with ChatGPT-enhanced learning.

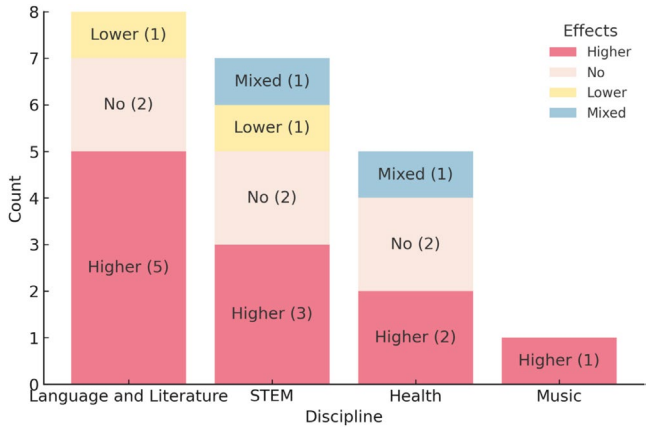

FIGURE 4 | Outcomes by discipline.

3.3.3 | Outcomes by Learning Objectives

Based on learning objectives, we divided the included studies into two major categories: knowledge acquisition and skills building. Knowledge acquisition refers to interventions using ChatGPT to help students master theoretical knowledge and concepts in specific domains (e.g., medical terminology, grammar and music theory). Skills building involves using ChatGPT to develop practical abilities and problem-solving skills, such as programming skills and writing skills. Table 12 presents these two categories along with their specific learning objectives and corresponding effects.

Overall, ChatGPT interventions are generally more effective for knowledge acquisition, with most targets showing significantly improved performance. Five out of seven studies reported significantly higher performance with ChatGPT integration. These positive outcomes were observed in medical terminology, programming knowledge, dental knowledge, grammar knowledge and music knowledge. Two studies found no significant differences in surgery knowledge and pharmacy knowledge.

In skills building, writing skills showed predominantly positive results, with four out of five studies reporting significantly higher performance, while one study found no significant difference. However, programming skill development showed no significant differences in both studies. The results for problem-solving

TABLE 12 | Outcomes by learning objectives.

Category	Objectives	Articles	Effects
Knowledge Acquisition	Surgery knowledge	Araji and Brooks (2024)	No (1)
	Medical terminology	Hsu (2024)	Higher (1)
	Programming knowledge	Huesca et al. (2024)	Higher (1)
	Dental knowledge	Kavadella et al. (2024)	Higher (1)
	Pharmacy knowledge	Svendsen et al. (2024)	No (1)
	Grammar knowledge	Kucuk (2024)	Higher (1)
	Music knowledge	Zhou and Kim (2024)	Higher (1)
Skills building	Writing skill	Escalante et al. (2023); Boudouaia et al. (2024); Song and Song (2023); Tsai et al. (2024); Wiboolyasarin et al. (2024)	Higher (4); No (1)
	Digital storytelling skill	Avello et al. (2024)	No (1)
	Problem-solving skill	Dasari et al. (2024); Wahba et al. (2024)	Lower (1); Higher (1)
	Programming skill	Kosar et al. (2024); Sun et al. (2024)	No (2)
	Creativity	Niloy et al. (2024)	Lower (1)
	Content creation	Li (2023) Li et al. (2024)	Higher (1); Mixed (1)
Knowledge acquisition & skills building	Theoretical knowledge and clinical skills	Wu et al. (2025)	Mixed (1)

skills were mixed, with one study indicating significantly higher performance and another reporting significantly lower performance. Similarly, in creative and content creation tasks, the findings were inconsistent. One study found significantly lower performance in creativity, whereas content creation studies produced varied results—one study reported significantly higher performance, while the other presented a mix of outcomes. Wu et al. (2025) was the only study that combines both knowledge acquisition and skills building as learning objectives, and it reported mixed results.

3.3.4 | Outcomes by Intervention Duration

The outcomes by intervention duration are demonstrated in Figure 5. Among the six studies examining short-term interventions, which lasted within a single day, only one showed higher performance, although four reported no difference and one showed lower performance. For medium-term interventions, spanning more than 1 day but up to 3 months, the results were largely positive. Out of 12 studies, 10 demonstrated high performance, while only one reported no effect and another showed mixed results. Long-term interventions, lasting more than 3 months, displayed a more varied outcome. None of the three studies reported higher performance, with one showing no difference, one reporting lower performance and another yielding mixed results.

Overall, the findings suggest that ChatGPT-based interventions have a generally positive impact on academic outcomes, particularly in knowledge acquisition and writing skill development. However, their effectiveness varies across disciplines, learning objectives and intervention durations.

4 | Discussion

The purpose of this review was to examine the articles related to the integration of ChatGPT into educational interventions in order to summarise the attempts that researchers have made to inform future applications. The 21 selected articles were coded and analysed with reference to three key dimensions, which correspond to the predefined research questions: research design (RQ1), intervention characteristics (RQ2) and learning outcomes (RQ3).

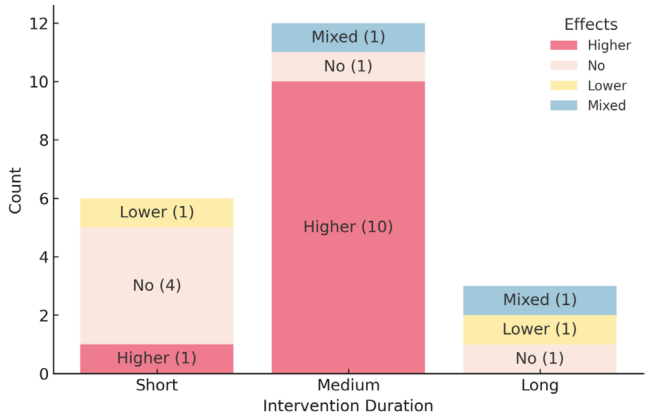

FIGURE 5 | Outcomes by duration.

This review has five main findings: (1) True experimental designs were widely used, ensuring higher internal validity. (2) Medium sample sizes and pretest–posttest designs are most common. (3) Interventions focus mainly on language and STEM fields, with task assistance being the dominant intervention type, followed by general learning support. (4) ChatGPT was generally effective for knowledge acquisition, but its impact on skills development varied. (5) Medium-term interventions showed the best results, while short-term effects were limited and long-term outcomes were mixed. In the following section, we discuss these findings and their implications for future research and educational practice.

4.1 | Characteristics of Experimental Design

In Ansari et al. (2024)'s review, it is pointed out that it is necessary to conduct strong randomised controlled trials to examine the positive/negative effects or lack of effects of ChatGPT use on teacher instruction and student development. This coincides with the results of our review, and some researchers have already conducted randomised controlled trials. In the 21 included studies, 12 experimental designs used a true experiment study design, which means randomised assignment of participants and control of variables. The remaining 9 studies used a quasi-experimental design without random assignment. This may be due to the limitations of random assignment in educational settings, considerations of educational equity, ethical issues and operational feasibility. Although true experiments are considered the most rigorous, it is not always practical in real-world scenarios to use random assignment.

Several studies in this review had sample sizes that fell short of Creswell's recommended minimum of 15 participants per group (Creswell 2015, 146). Specifically, three of the included studies did not meet this criterion. Most of the studies had a total sample size of no more than 100 individuals. The smaller sample sizes of these studies may affect the robustness and generalizability of the results, limiting their inferential validity to larger audiences. However, considering that ChatGPT is an emerging technology that has not been launched for a long time, current research mainly focuses on exploring its potential and application in education. Therefore, small sample experimental research is reasonable and acceptable in this context. At the same time, several studies also mentioned the lack of diversity in their sample (Hsu 2024; Kavadella et al. 2024; Niloy et al. 2024; Wiboolyasarin et al. 2024). This limitation may be due to the fact that most of the studies were limited to a single department or a few classes at a single university, thus limiting the ability to collect a more diverse and representative sample. This limitation may also affect the generalizability of the studies' findings because they may not adequately reflect differences in responses across populations (e.g., age, gender, cultural background or discipline). As a result, conclusions drawn from these studies may be more applicable to specific subgroups than to the broader student population. Ansari et al. (2024) also mentioned in their review that there is a need to investigate the existing practices

of teachers and students in ChatGPT use with a large sample. Therefore, future studies need to include larger and more diverse samples.

Furthermore, the majority of the studies employed a pretest–posttest design. This approach was effective in demonstrating the direct impact of the intervention on student academic performance, especially when other variables were controlled for. Six other studies tested only after the intervention, and while these studies also reported some outcomes, the lack of pretest data did not allow for a complete measure of the enhancement of academic performance by ChatGPT.

4.2 | Intervention and Implementation

4.2.1 | Discipline Focus

This review found that the disciplinary focus of the included studies was primarily in the areas of language and literature, as well as STEM. There is also a considerable amount of research in the area of health education. The reasons for this may be that language and literature may benefit from ChatGPT's language processing capabilities, while STEM fields may value its problem-solving and code-generating abilities. Interestingly, the arts, represented here by a single study in music, appear to be underexplored. Another study on the teaching of 'creative writing' reported negative results of ChatGPT (Niloy et al. 2024). This suggests opportunities for future research to investigate how ChatGPT might be applied in creative disciplines. While our review focused on applications of general-purpose ChatGPT across disciplines, it is worth noting that some researchers are exploring more specialised AI solutions for specific educational contexts. For instance, Hakim et al. (2024) developed a custom chatbot tailored specifically for occupational health and safety (OHS) education. This suggests that discipline-specific AI solutions may offer certain advantages over general-purpose models like ChatGPT, especially in specialised fields.

4.2.2 | Task Assistance and General Learning Support

In the 'task assistance' type of intervention, structured tasks often give students clear instructions on how and when to use the tool. This type of intervention is generally of a shorter duration, which allows for more precise control over the variables, as the instructor can monitor and guide the students' use of ChatGPT during the completion of the tasks. The shorter duration of these interventions makes them easier to implement within the constraints of a typical academic schedule and allows for more rapid assessment of the tool's impact on specific learning outcomes. Task assistance occurs in two main scenarios: first, discipline-specific problem solving, such as in writing, programming, or data analysis tasks, students can use ChatGPT to generate first drafts, optimise code, or obtain relevant information to complete tasks faster and more efficiently. The second is instant feedback and improvement, such as grammatical corrections, logical adjustments, or structural optimization. This not only speeds up the efficiency of task completion, but also improves the quality of students' work.

However, due to its emphasis on the process of task completion, it may cause students to rely excessively on ChatGPT's advice, reducing opportunities for critical thinking and problem solving. In particular, the tasks in the included studies are mainly text-based tasks, while other forms of tasks are relatively rare. Since ChatGPT is primarily a text-based tool, it naturally aligns with tasks that involve writing, editing, or organising written content. However, students might simply copy or lightly modify the content generated by ChatGPT without fully engaging with or internalising the material. This can lead to a surface-level understanding, where students may not critically analyse or reflect on the content provided by ChatGPT. Addressing these limitations requires exploring how ChatGPT can be effectively applied across different disciplines and assignment types, particularly by expanding its use beyond text-based tasks. A good example of this is Kosar et al. (2024), who designed an 'assignment defence' part to reduce plagiarism and promote deeper learning. In this process, students must answer conceptual and analytical questions about their code, demonstrate modifications and complete missing parts. This interactive defence not only ensures a genuine understanding of programming concepts but also fosters critical thinking. Given the increasing ease of plagiarism with ChatGPT, such measures help encourage active engagement with the material rather than passive reliance on AI-generated content. Furthermore, the findings indicate that the main way of application in the field of language education is to use ChatGPT to give students feedback on their writing, and there seems to be less application in other language skills. Interventions in listening or speaking could be explored in the future using ChatGPT's new voice dialogue feature. Students could also use ChatGPT in conversational practice, debates, or simulations, which would require them to process and respond to information in real-time, promoting deeper cognitive engagement.

The other type of intervention, general learning support, involves a relatively longer duration and has fewer restrictions on interaction mode (not limited to the classroom). This approach provides students with greater flexibility and the freedom to explore and use the tool according to their preferences, to promote a personalised learning experience. But at the same time, there are some potential issues worth discussing. One key issue is the control of variables, especially when the experimental and control groups are not restricted to using the AI only in the classroom. In such cases, it becomes difficult to ensure that students in the control group do not independently access or use ChatGPT outside of the classroom, which could confound the results and undermine the validity of the study's findings. Another important consideration is the variability in how often and for how long students use ChatGPT. In a flexible, student-driven learning environment, it can be difficult to ensure that all students are using the tool sufficiently to produce measurable effects. Inconsistencies in frequency and duration of use can lead to varying levels of exposure to the intervention, which in turn can affect intervention outcomes and reduce the ability to draw firm conclusions about intervention effects.

Introducing new methods to monitor the frequency and content of students' use of ChatGPT may better control variables, like in Sun et al. (2024), who documented students' programming behaviour

by monitoring platform logs and recording computer screen video. Such a monitoring tool would ensure that the usage behaviours of both the experimental and control groups were accurately recorded, thus avoiding confounding of results caused by the control group's exposure to ChatGPT outside of the classroom. In addition, by monitoring the data, the researcher was able to better understand the impact of different frequencies and content of use on learning outcomes, thus ensuring consistency in the effects of the intervention and providing a more reliable basis for in-depth analyses of the use of ChatGPT in personalised learning.

4.2.3 | Roles of ChatGPT in Interventions

Five main roles played by ChatGPT in the interventions were also analysed: content generator, information retriever, feedback and evaluator, learning companion and personalised learning guide. ChatGPT's diverse roles in education demonstrate its flexibility in supporting instructional tasks. From content generation to personalised instruction, ChatGPT's application scenarios are broad and deep. These roles complement traditional teaching methods, especially in the areas of feedback and assessment, and ChatGPT shows the potential to replace or augment the role of the human teacher.

4.2.4 | Interaction Modes

Three different modes of interactions showed how and where students interact with ChatGPT. The in-class mode provides structured instruction but limits student autonomy; the out-of-class model enhances student flexibility but faces issues of academic integrity and dependency; the blended mode combines the strengths of both in and out of the class and provides a more balanced learning experience but is complex to implement and monitor; and the indirect mode, while providing objective feedback, lacks direct student interaction with ChatGPT.

The choice of interaction mode should be based on different disciplines and learning objectives. The in-class mode may be more suitable for short-term tasks, while the blended and out-of-class modes may be more effective for long-term learning. At the same time, it is important to optimise these modes by addressing their limitations. For instance, to mitigate academic integrity concerns in the out-of-class mode, AI detection tools, oral presentations, and formative assessments can be incorporated to reduce plagiarism and over-reliance. In the in-class mode, integrating group discussions, peer evaluations and teacher-guided activities can enhance student engagement and autonomy.

4.3 | Outcomes

This review examined ChatGPT's impact on academic performance by comparing experimental and control groups. Results show that 52.4% of studies reported significantly higher performance, 28.6% found no difference, 9.5% showed lower performance, and 9.5% had mixed results. These findings indicate that while ChatGPT has the potential to enhance academic performance, its impact is not uniform across disciplines, learning objectives and intervention duration.

4.3.1 | Outcomes Across Different Disciplines

In language and literature education, ChatGPT interventions proved particularly beneficial for structured tasks such as EFL (English as a Foreign Language) writing and grammar learning, but less effective for creative writing, suggesting that AI tools may be better at supporting rule-based language acquisition than creative expression. The underperformance in creative writing is likely due to AI's data-driven approach, which may priorities convention over originality, potentially limiting students' creative development. This highlights the need to consider the alignment between task type and AI capabilities when designing interventions.

Results in STEM were more mixed. While three out of seven studies reported significantly higher performance, some studies found no significant differences, particularly in self-directed learning for programming. The lower performance reported by Dasari et al. (2024) in a mathematics study suggests that reliance on ChatGPT without adequate instructional support may hinder deep conceptual understanding. In addition, the mixed results reported in STEM teacher training suggest that while ChatGPT may facilitate task-based learning, its benefits may not extend to all aspects of education.

In health education, the mixed results between theoretical knowledge gains and practical skill development reveal a critical limitation of current AI tools: they excel at delivering information, but struggle to facilitate the hands-on learning experiences essential for clinical skills. As noted by Araji and Brooks (2024), many participants pointed out that certain concepts were difficult to grasp because ChatGPT, being text-based, cannot generate images during their learning. The importance of concept maps and graphics in medical education is well established—they simplify concepts, promote active learning and enhance critical thinking. This text-only limitation represents a significant constraint for disciplines like medicine where visual learning is crucial for comprehension and retention. The theory-practice gap suggests that ChatGPT may best serve as one component within a broader pedagogical framework that includes visual aids and hands-on practice for skill development.

The only study in the field of music showed that ChatGPT has a significant improvement in music knowledge acquisition, but more application effects in this field remain to be explored.

The different outcomes across disciplines likely reflect varying alignment between ChatGPT's capabilities and disciplinary epistemologies. Subjects with well-defined rules and structured content knowledge appear more compatible with AI-assisted instruction than those requiring tacit knowledge, creative synthesis, visual conceptualisation, or physical skill development. These findings suggest that effective integration of ChatGPT in education requires careful consideration of disciplinary contexts, learning objectives and instructional design. Rather than viewing AI as a universal educational solution, educators should

strategically implement ChatGPT as part of a blended approach, particularly for knowledge acquisition and structured learning tasks, while maintaining traditional methods for creative and practical skill development.

4.3.2 | Outcomes and Learning Objectives

When examining learning objectives, ChatGPT interventions appear particularly effective for knowledge acquisition, with five out of seven studies reporting significantly higher performance. This suggests that ChatGPT's language modelling capabilities are well-suited for facilitating theoretical understanding across diverse subjects like medical terminology, programming knowledge and English grammar. The ability of AI to present information in a conversational, responsive way can create a more engaging learning experience than traditional methods. This can potentially improve knowledge retention. The disparity between programming knowledge and programming skill outcomes deserves special attention. While ChatGPT effectively conveys programming concepts, it appears less capable of facilitating skill development. This knowledge-skill gap likely stems from the fundamental difference between understanding theoretical principles and applying them in practice. Programming skill development requires hands-on experience, trial-and-error learning and debugging practice that may not be adequately supported through text-based AI interaction alone. Similarly, Wu et al. (2025)'s finding of improved knowledge but unchanged clinical skills reinforces this distinction between theoretical understanding and practical application.

4.3.3 | Outcomes and Intervention Duration

For duration, medium-term use (1 day to 3 months) showed mostly good results, while both short-term and long-term use were less effective. Short-term interventions may not give students enough time to get used to ChatGPT and adjust their learning strategies. On the other hand, long-term use raises concerns about staying engaged and becoming too dependent. Students may see early benefits, but over time, the impact may decrease as the novelty wears off or they rely too much on AI instead of building their own skills. These findings suggest that intervention duration alone is not the only factor; how learners engage with ChatGPT over time and the instructional strategies in place also play a critical role in determining outcomes.

In interpreting the inconsistent or limited effects reported in some studies, it is crucial to consider factors such as duration, task design, disciplinary context and the level of instructional scaffolding—all of which may influence the effectiveness of ChatGPT-based interventions.

4.4 | Future Directions

Based on this review's findings, several research priorities emerge. Methodologically, future research should adopt larger and more diverse samples and prioritise true experimental designs with randomised assignments. Additionally, incorporating pre-test and post-test methodologies will allow for a more precise assessment of ChatGPT's impact on learning outcomes to ensure stronger causal inferences and broader generalisability.

The current application scope of ChatGPT reveals disciplinary imbalances. While its integration in language learning and STEM education is well documented, its potential in creative disciplines remains unexplored. Future research should investigate its role in supporting education in fields such as artistic expression and design pedagogy, and explore customised AI chatbots tailored for specific disciplines to enhance subject specific learning.

Current interventions present both implementation challenges and opportunities for improvement. Task assistance interventions predominantly focus on text-based tasks, raising concerns about student dependency and diminished critical thinking. Future research should explore non-textual tasks such as conversational practice, debates, and simulations, and integrate scaffolding techniques that gradually reduce AI assistance to foster independent learning. Developing pedagogical frameworks that position ChatGPT as a catalyst for analytical thinking rather than a direct problem solver is also essential. In general learning support interventions, variations in students' usage frequency and engagement patterns present challenges for experimental control. Future research should integrate monitoring tools to track and record interaction data, capturing details such as usage duration, response patterns and engagement depth. This would provide a more precise understanding of how different interaction behaviours influence learning outcomes.

While ChatGPT excels at knowledge acquisition, its effectiveness in skills development (e.g., programming, clinical skills, creativity) is inconsistent. Future research should explore hybrid learning models that integrate ChatGPT with practical tasks, project-based learning and hands-on training to bridge the gap between knowledge and application. Although current findings indicate that medium-term interventions (1 day–3 months) yield the best outcomes, the underlying reasons behind this pattern remain unclear. Further research is needed to examine whether factors such as engagement levels, adaptation to AI-assisted learning, cognitive load and potential reliance on AI contribute to the observed differences in learning effectiveness across intervention durations.

5 | Conclusion

This systematic review, in accordance with the PRISMA guidelines, summarised the current state of the art in the integration of ChatGPT in higher education. After a thorough literature search and screening process, 21 studies were included in this review. The articles were coded according to research design, intervention and outcomes, offering a clear picture of how ChatGPT is currently being researched and used in higher education settings, as well as comparing the differences in outcomes between ChatGPT-based interventions and traditional methods.

The findings indicate that true experimental designs were widely used, with most studies employing medium sample sizes and

pretest-posttest designs. ChatGPT interventions were primarily applied in language and STEM education, with task assistance being the dominant approach. Five main roles of ChatGPT were discovered, and four modes of interaction were identified. While ChatGPT proved effective for knowledge acquisition, its impact on skills development varied across disciplines. Medium-term interventions (1 day–3 months) yielded the most positive outcomes, whereas short-term effects were limited, and long-term interventions produced mixed results.

This study has several limitations. First, we only included papers that directly reported on academic performance, without examining other dimensions that ChatGPT might influence, such as motivation, attitude, or technology acceptance. This focus may have overlooked other important aspects of how ChatGPT impacts the overall educational experience. Second, at the time of writing this review, ChatGPT had been available for less than 2 years, meaning the included studies represent early-stage explorations of its use in higher education. There may be more in-depth studies in the future that deserve further analysis. Third, given the accelerating pace of AI innovation, ChatGPT's features and capabilities are subject to continuous evolution. Most of the papers included used ChatGPT3.5, while the optimised ChatGPT4 and ChatGPT4o, or other Generative AI tools (e.g., Claude, Bart) may have more powerful features. It would be beneficial for future research to consider the newer versions of this technology or other tools to evaluate their impact and potential advantages in educational settings. This will ensure that the findings remain relevant in an ever-changing technological landscape.

While this review outlines current trends and evidence, further research is required to enhance methodological rigour for more reliable results, explore ChatGPT's role in underrepresented disciplines for wider application, and refine intervention strategies to improve learning outcomes. As AI continues to evolve, its integration should be critically assessed to maximise learning benefits while addressing potential challenges.

Author Contributions

Yiwen Jin: conceptualization, writing – original draft, methodology, investigation, formal analysis. **Lies Sercu:** supervision, conceptualization, methodology, writing – review and editing.

Acknowledgements

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

Adeshola, I., and A. P. Adepoju. 2023. "The Opportunities and Challenges of ChatGPT in Education." *Interactive Learning Environments* 32: 1–14. https://doi-org.kuleuven.e-bronnen.be/10.1080/10494820.2023.2253858.

Alnaqbi, N. M., and W. Fouda. 2023. "Exploring the Role of ChatGPT and Social Media in Enhancing Student Evaluation of Teaching Styles in Higher Education Using Neutrosophic Sets." *International Journal of Neutrosophic Science* 20, no. 4: 181–190. https://doi.org/10.54216/IJNS.200414.

Ansari, A. N., S. Ahmad, and S. M. Bhutta. 2024. "Mapping the Global Evidence Around the Use of ChatGPT in Higher Education: A Systematic Scoping Review." *Education and Information Technologies* 29, no. 9: 11281–11321. https://doi.org/10.1007/s10639-023-12223-4.

Araji, T., and A. D. Brooks. 2024. "Evaluating the Role of ChatGPT as a Study Aid in Medical Education in Surgery." *Journal of Surgical Education* 81, no. 5: 753–757. https://doi.org/10.1016/j.jsurg.2024.01.014.

Avello, R., T. Gajderowicz, and V. G. Gómez-Rodríguez. 2024. "Is ChatGPT Helpful for Graduate Students in Acquiring Knowledge About Digital Storytelling and Reducing Their Cognitive Load? An Experiment." *Revista de Educación a Distancia (RED)* 24, no. 78: 604621. https://doi.org/10.6018/red.604621.

Baig, M. I., and E. Yadegaridehkordi. 2024. "ChatGPT in the Higher Education: A Systematic Literature Review and Research Challenges." *International Journal of Educational Research* 127: 102411. https://doi.org/10.1016/j.ijer.2024.102411.

Bingham, A. J., and P. Witkowsky. 2022. "Deductive and Inductive Approaches to Qualitative Data Analysis." In *Analyzing and Interpreting Qualitative Data: After the Interview*, edited by C. Vanover, P. Mihas, and J. Saldaña, 133–146. SAGE Publications.

Boudouaia, A., S. Mouas, and B. Kouider. 2024. "A Study on ChatGPT-4 as an Innovative Approach to Enhancing English as a Foreign Language Writing Learning." *Journal of Educational Computing Research* 62, no. 6: 1289–1317. https://doi.org/10.1177/07356331241247465.

Chen, X., Z. Hu, and C. Wang. 2024. "Empowering Education Development Through AIGC: A Systematic Literature Review." *Education and Information Technologies* 29: 17485–17537. https://doi.org/10.1007/s10639-024-12549-7.

Creswell, J. W. 2015. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. pearson.

Dasari, D., A. Hendriyanto, S. Sahara, et al. 2024. "ChatGPT in Didactical Tetrahedron, Does It Make an Exception? A Case Study in Mathematics Teaching and Learning." *Frontiers in Education* 8: 1295413. https://doi.org/10.3389/feduc.2023.1295413.

Dempere, J., K. Modugu, A. Hesham, and L. K. Ramasamy. 2023. "The Impact of ChatGPT on Higher Education." *Frontiers in Education* 8: 1206936. https://doi.org/10.3389/feduc.2023.1206936.

Escalante, J., A. Pack, and A. Barrett. 2023. "AI-Generated Feedback on Writing: Insights Into Efficacy and ENL Student Preference." *International Journal of Educational Technology in Higher Education* 20, no. 1: 57. https://doi.org/10.1186/s41239-023-00425-2.

Hakim, V. G. A., N. A. Paiman, and M. H. S. Rahman. 2024. "Genie-On-Demand: A Custom AI Chatbot for Enhancing Learning Performance, Self-Efficacy, and Technology Acceptance in Occupational Health and Safety for Engineering Education." *Computer Applications in Engineering Education* 32, no. 6: e22800.

Heck, T., C. Keller, and M. Rittberger. 2024. "Coverage and Similarity of Bibliographic Databases to Find Most Relevant Literature for Systematic Reviews in Education." *International Journal on Digital Libraries* 25: 365–376. https://doi.org/10.1007/s00799-023-00364-3.

Hong, Q. N., P. Pluye, S. Fàbregues, et al. 2018. "Mixed Methods Appraisal Tool (MMAT)." version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada.

Hsu, M.-H. 2024. "Mastering Medical Terminology With ChatGPT and Termbot." *Health Education Journal* 83, no. 4: 352–358. https://doi.org/10.1177/00178969231197371.

Huesca, G., Y. Martínez-Treviño, J. M. Molina-Espinosa, et al. 2024. "Effectiveness of Using ChatGPT as a Tool to Strengthen Benefits of the Flipped Learning Strategy." *Education in Science* 14, no. 6: 660. https://doi.org/10.3390/educsci14060660.

Kasneci, E., K. Seßler, S. Küchemann, et al. 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.

Kavadella, A., M. A. D. Da Silva, E. G. Kaklamanos, V. Stamatopoulos, and K. Giannakopoulos. 2024. "Evaluation of Chatgpt's Real-Life Implementation in Undergraduate Dental Education: Mixed Methods Study." *JMIR Medical Education* 10, no. 1: e51344. https://doi.org/10.2196/51344.

Kosar, T., D. Ostojić, Y. D. Liu, and M. Mernik. 2024. "Computer Science Education in ChatGPT Era: Experiences From an Experiment in a Programming Course for Novice Programmers." *Mathematics* 12, no. 5: 629. https://doi.org/10.3390/math12050629.

Kucuk, T. 2024. "ChatGPT Integrated Grammar Teaching and Learning in EFL Classes: A Study on Tishk International University Students in Erbil, Iraq." *Arab World English Journal (AWEJ)* 1, no. 1: 100–111. https://ssrn.com/abstract=4814669. https://doi.org/10.24093/awej/ChatGPT.6.

Labadze, L., M. Grigolia, and L. Machaidze. 2023. "Role of AI Chatbots in Education: Systematic Literature Review." *International Journal of Educational Technology in Higher Education* 20: 56. https://doi.org/10.1186/s41239-023-00426-1.

Li, H. 2023. "Effects of a ChatGPT-Based Flipped Learning Guiding Approach on Learners' Courseware Project Performances and Perceptions." *Australasian Journal of Educational Technology* 39, no. 5: 40–58. https://doi.org/10.14742/ajet.8923.

Li, T., Y. Ji, and Z. Zhan. 2024. "Expert or Machine? Comparing the Effect of Pairing Student Teacher With In-Service Teacher and ChatGPT on Their Critical Thinking, Learning Performance, and Cognitive Load in an Integrated-STEM Course." *Asia Pacific Journal of Education* 44, no. 1: 45–60. https://doi.org/10.1080/02188791.2024.2305163.

Lo, C. K. 2023. "What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature." *Education in Science* 13, no. 4: 410. https://doi.org/10.3390/educsci13040410.

Lu, Q., Y. Yao, L. Xiao, M. Yuan, J. Wang, and X. Zhu. 2024. "Can ChatGPT Effectively Complement Teacher Assessment of Undergraduate Students' Academic Writing?" *Assessment & Evaluation in Higher Education* 49, no. 5: 616–633. https://doi.org/10.1080/02602938.2024.2301722.

Meyer, J. G., R. J. Urbanowicz, P. C. N. Martin, et al. 2023. "ChatGPT and Large Language Models in Academia: Opportunities and Challenges." *Biodata Mining* 16: 20 https://doi-org.kuleuven.e-bronnen.be/10.1186/s13040-023-00339-9.

Montenegro-Rueda, M., J. Fernández-Cerero, J. M. Fernández-Batanero, and E. López-Meneses. 2023. "Impact of the Implementation of ChatGPT in Education: A Systematic Review." *Compute* 12: 153. https://doi.org/10.3390/computers12080153.

Niloy, A. C., S. Akter, N. Sultana, J. Sultana, and S. I. U. Rahman. 2024. "Is Chatgpt a Menace for Creative Writing Ability? An Experiment." *Journal of Computer Assisted Learning* 40, no. 2: 919–930. https://doi.org/10.1111/jcal.12929.

Page, M. J., J. E. McKenzie, P. M. Bossuyt, et al. 2021. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." *BMJ (Clinical Research Ed.)* 372: n71. https://doi.org/10.1136/bmj.n71.

Shang, S., and S. Geng. 2024. "Empowering Learners With AI-Generated Content for Programming Learning and Computational Thinking: The Lens of Extended Effective Use Theory." *Journal of Computer Assisted Learning* 40, no. 4: 1941–1958. https://doi.org/10.1111/jcal.12996.

Song, C., and Y. Song. 2023. "Enhancing Academic Writing Skills and Motivation: Assessing the Efficacy of ChatGPT in AI-Assisted Language Learning for EFL Students." *Frontiers in Psychology* 14: 1260843. https://doi.org/10.3389/fpsyg.2023.1260843.

Sun, D., A. Boudouaia, C. Zhu, and Y. Li. 2024. "Would ChatGPT-Facilitated Programming Mode Impact College Students' Programming Behaviors, Performances, and Perceptions? An Empirical Study." *International Journal of Educational Technology in Higher Education* 21, no. 1: 14–22. https://doi.org/10.1186/s41239-024-00446-5.

Svendsen, K., M. Askar, D. Umer, and K. H. Halvorsen. 2024. "Short-Term Learning Effect of ChatGPT on Pharmacy Students' Learning." *Exploratory Research in Clinical and Social Pharmacy* 15: 100478. https://doi.org/10.1016/j.rcsop.2024.100478.

Tsai, C. Y., Y. T. Lin, and I. K. Brown. 2024. "Impacts of ChatGPT-Assisted Writing for EFL English Majors: Feasibility and Challenges." *Education and Information Technologies* 29: 22427–22445. https://doi.org/10.1007/s10639-024-12722-y.

Urban, M., F. Děchtěrenko, J. Lukavský, et al. 2024. "ChatGPT Improves Creative Problem-Solving Performance in University Students: An Experimental Study." *Computers & Education* 215: 105031. https://doi.org/10.1016/j.compedu.2024.105031.

Wahba, F., A. O. Ajlouni, and M. A. Abumosa. 2024. "The Impact of ChatGPT-Based Learning Statistics on Undergraduates' Statistical Reasoning and Attitudes Toward Statistics." *Eurasia Journal of Mathematics, Science and Technology Education* 20, no. 7: em2468. https://doi.org/10.29333/ejmste/14726.

Wiboolyasarin, W., K. Wiboolyasarin, K. Suwanwihok, N. Jinowat, and R. Muenjanchoey. 2024. "Synergizing Collaborative Writing and AI Feedback: An Investigation Into Enhancing L2 Writing Proficiency in Wiki-Based Environments." *Computers and Education: Artificial Intelligence* 6: 100228. https://doi.org/10.1016/j.caeai.2024.100228.

Wu, C., L. Chen, M. Han, Z. Li, N. Yang, and C. Yu. 2025. "Application of ChatGPT-Based Blended Medical Teaching in Clinical Education of Hepatobiliary Surgery." *Medical Teacher* 47, no. 3: 1–5. https://doi.org/10.1080/0142159X.2024.2339412.

Zhai, C., S. Wibowo, and L. D. Li. 2024. "The Effects of Over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review." *Smart Learning Environments* 11: 28. https://doi.org/10.1186/s40561-024-00316-7.

Zhang, P., and G. Tur. 2024. "A Systematic Review of ChatGPT Use in K-12 Education." *European Journal of Education* 59, no. 2: e12599. https://doi.org/10.1111/ejed.12599.

Zhou, W., and Y. J. Kim. 2024. "Innovative Music Education: An Empirical Assessment of ChatGPT-4's Impact on Student Learning Experiences." *Education and Information Technologies* 29: 20855–20881. https://doi.org/10.1007/s10639-024-12705-z.

Appendix A

Quality Appraisal

TABLE A1 | Mixed Methods Appraisal Tool (MMAT)—quantitative randomised controlled trials.

Author (year)	Questions[a]					Overall appraisal
	1	2	3	4	5	
Araji and Brooks (2024)	CT	Y	Y	N	Y	Include
Avello et al. (2024)	Y	Y	Y	N	Y	Include
Boudouaia et al. (2024)	Y	Y	Y	N	Y	Include
Hsu (2024)	Y	Y	Y	N	Y	Include
Huesca et al. (2024)	Y	Y	Y	N	Y	Include
Kavadella et al. (2024)	Y	CT	Y	Y	Y	Include
Kosar et al. (2024)	Y	Y	Y	N	Y	Include
Kucuk (2024)	Y	Y	Y	N	Y	Include
Song and Song (2023)	Y	Y	Y	N	Y	Include
Svendsen et al. (2024)	Y	Y	Y	N	Y	Include
Tsai et al. (2024)	Y	Y	Y	Y	Y	Include
Wu et al. (2025)	Y	Y	Y	Y	Y	Include

Abbreviations: CT, can't tell; N, no; Y, yes.
[a]1. Is randomisation appropriately performed? 2. Are the groups comparable at baseline? 3. Are there complete outcome data? 4. Are outcome assessors blinded to the intervention provided? 5. Did the participants adhere to the assigned intervention?

TABLE A2 | Mixed Methods Appraisal Tool (MMAT)—quantitative nonrandomized studies.

Author (year)	Questions[a]					Overall appraisal
	1	2	3	4	5	
Dasari et al. (2024)	CT	Y	Y	CT	Y	Include
Escalante et al. (2023)	CT	Y	Y	CT	Y	Include
Li (2023)	CT	Y	Y	CT	Y	Include
Li et al. (2024)	CT	Y	Y	CT	Y	Include
Niloy et al. (2024)	Y	Y	Y	CT	Y	Include
Sun et al. (2024)	Y	Y	Y	CT	Y	Include
Wahba et al. (2024)	Y	Y	Y	CT	Y	Include
Wiboolyasarin et al. (2024)	Y	Y	Y	Y	Y	Include
Zhou and Kim (2024)	Y	Y	Y	Y	Y	Include

Abbreviations: CT, can't tell; N, no; Y, yes.
[a]1. Are the participants representative of the target population? 2. Are measurements appropriate regarding both the outcome and intervention (or exposure)? 3. Are there complete outcome data? 4. Are the confounders accounted for in the design and analysis? 5. During the study period, is the intervention administered (or exposure occurred) as intended?

Appendix B

Summary of Included Studies

	Authors	Experiment design	Sample size	Duration	Discipline	Types of intervention	Interaction mode	Assessment timing	Outcome
1	Araji and Brooks (2024)	True experiment	19	One class session	Health education	Task assistance	In class	Pretest & posttest	No significant difference
2	Avello et al. (2024)	True experiment	41	One class session	Language and literature	Task assistance	In class	Pretest & posttest	No significant difference
3	Boudouaia et al. (2024)	True experiment	76	10 weeks	Language and literature	Task assistance	Blended	Pretest & posttest	Significantly higher
4	Dasari et al. (2024)	Quasi-experiment	29	1 semester	STEM	General learning support	Blended	Posttest only	Significantly lower
5	Escalante et al. (2023)	Quasi-experiment	91	6 weeks	Language and literature	Task assistance	Indirect	Pretest & posttest	No significant difference
6	Hsu (2024)	True experiment	60	2 months	Health Education	General learning support	Out of class	Pretest & posttest	Significantly higher
7	Huesca et al. (2024)	True experiment	365	10 weeks	STEM	General learning support	In class	Pretest & posttest	Significantly higher
8	Kavadella et al. (2024)	True experiment	77	1 month	Health education	Task assistance	Blended	Posttest only	Significantly higher
9	Kosar et al. (2024)	True experiment	182	1 semester	STEM	Task assistance	Blended	Posttest only	No significant difference
10	Kucuk (2024)	True experiment	60	7 weeks	Language and literature	General learning support	Blended	Pretest & posttest	Significantly higher
11	Li (2023)	Quasi-experiment	81	8 weeks	STEM	General learning support	Blended	Pretest & posttest	Significantly higher
12	Li et al. (2024)	Quasi-experiment	23	2 months	STEM	Task assistance	Blended	Posttest only	Mixed
13	Niloy et al. (2024)	Quasi-experiment	600	1 day	Language and literature	Task assistance	In class	Pretest & posttest	Significantly lower
14	Song and Song (2023)	True experiment	50	12 weeks	Language and literature	Task assistance	Blended	Pretest & posttest	Significantly higher
15	Sun et al. (2024)	Quasi-experiment	82	One class session	STEM	Task assistance	In class	Posttest only	Neutral
16	Svendsen et al. (2024)	True experiment	31	One class session	Health education	General learning support	In class	Pretest & posttest	Significantly higher
17	Tsai et al. (2024)	True experiment	52	One class session	Language and literature	Task assistance	In class	Pretest & posttest	Significantly higher
18	Wahba et al. (2024)	Quasi-experiment	56	3 weeks	STEM	Task assistance	Blended	Pretest & posttest	Significantly higher
19	Wiboolyasarin et al. (2024)	Quasi-experiment	39	10 weeks	Language and literature	Task assistance	In class	Pretest & posttest	Significantly higher
20	Wu et al. (2025)	True experiment	61	1 semester	Health education	General learning support	Blended	Posttest only	Mixed
21	Zhou and Kim (2024)	Quasi-experiment	74	10 weeks	Music	General learning support	Blended	Pretest & posttest	Significantly higher