



## A Systematic Review of the Limitations and Associated Opportunities of ChatGPT

Ngo Cong-Lem, Ali Soyooof & Diki Tsering

To cite this article: Ngo Cong-Lem, Ali Soyooof & Diki Tsering (08 May 2024): A Systematic Review of the Limitations and Associated Opportunities of ChatGPT, International Journal of Human-Computer Interaction, DOI: [10.1080/10447318.2024.2344142](https://doi.org/10.1080/10447318.2024.2344142)

To link to this article: <https://doi.org/10.1080/10447318.2024.2344142>



© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 08 May 2024.



Submit your article to this journal [↗](#)



Article views: 718






View related articles [↗](#)



View Crossmark data [↗](#)

# A Systematic Review of the Limitations and Associated Opportunities of ChatGPT

Ngo Cong-Lem<sup>a,b</sup> , Ali Soyoof<sup>c</sup> , and Diki Tsering<sup>a</sup> 

<sup>a</sup>BehaviourWorks Australia, Monash University, Clayton, Australia; <sup>b</sup>Faculty of Foreign Languages, Dalat University, Dalat, Vietnam; <sup>c</sup>Faculty of Education, University of Macau, Taipa, Macau SAR, China

## ABSTRACT

This systematic review explores the limitations and opportunities associated with ChatGPT's application across various fields. Following a rigorous screening process of 485 studies identified through searches in Scopus, Web of Science, ERIC, and IEEE Xplore databases, 33 high-quality empirical studies were selected for analysis. The review identifies five key limitations: accuracy and reliability concerns, limitations in critical thinking and problem-solving, multifaceted impacts on learning and development, technical constraints related to input and output, and ethical, legal, and privacy concerns. However, the review also highlights five exciting opportunities: educational support and skill development, workflow enhancement, information retrieval, natural language interaction and assistance, and content creation and ideation. While this review provides valuable insights, it also highlights some gaps. Limited transparency in the studies regarding specific ChatGPT versions used hinders generalizability. Additionally, the extent to which these findings can be transferred to more advanced models like ChatGPT-4 remains unclear. By acknowledging both limitations and opportunities, this review offers a foundation for researchers, developers, and practitioners to consider when exploring the potential and responsible application of ChatGPT and similar evolving AI tools.

## KEYWORDS

ChatGPT; large language model; limitation; opportunity; review

## 1. Introduction

In the rapidly evolving landscape of natural language processing, OpenAI's ChatGPT, introduced in November 2022, has emerged as a groundbreaking artificial intelligence model with versatile applications across various domains (Bubeck et al., 2023). Leveraging cutting-edge deep learning techniques, ChatGPT has demonstrated remarkable capabilities, producing well-written and coherent responses in conversational-style interactions (Hassani & Silva, 2023; Ray, 2023). As the demand for sophisticated language models continues to surge, it is crucial to critically assess the limitations that may hinder language models' optimal performance. Recognizing and understanding these limitations is crucial for researchers, developers, and end-users alike.

Given ChatGPT's recent introduction in 2022, existing research is notably limited in its exploration of the model's applications, primarily focusing on specific domains such as business, education, or academic research (e.g., Lo, 2023; Rahman et al., 2023; Singh & Singh, 2023). Reviews confined to specific fields risk obscuring hidden strengths or shortcomings that may manifest across various applications. Therefore, a comprehensive review is essential to unveil the

broader spectrum of challenges and potentials inherent in ChatGPT's deployment across diverse domains.

This review delves into the wealth of empirical studies conducted on ChatGPT to unveil documented constraints that may encompass accuracy, ethical issues, and other technical limitations. Beyond an examination of limitations, our review endeavors to shed light on the opportunities that arise from understanding and addressing these constraints. By elucidating pathways for improvement, we aim to contribute to the ongoing discussion covering the enhancement of ChatGPT's utility in various domains.

Accordingly, this systematic review endeavors to explore and synthesize the existing empirical literature to address two pivotal research questions:

- What limitations of ChatGPT are documented in the prior empirical literature?
- In light of the limitations identified, what opportunities exist for enhancing the utilization of ChatGPT?

This systematic exploration of ChatGPT's limitations and associated opportunities not only serves as a comprehensive resource for researchers and practitioners but also aims to

**CONTACT** Ngo Cong-Lem  [Cong.ngo@monash.edu](mailto:Cong.ngo@monash.edu)  BehaviourWorks Australia, Monash University, 8 Scenic Boulevard, Clayton, Victoria 3168, Australia; Faculty of Foreign Languages, Dalat University, Dalat, Vietnam

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

foster a deeper understanding of the nuances involved in leveraging state-of-the-art language models. As we navigate the intricate landscape of artificial intelligence, a nuanced understanding of ChatGPT's strengths and weaknesses is paramount for harnessing its full potential and driving innovation in natural language processing.

## 2. A brief overview of research background on ChatGPT

Research has convincingly demonstrated that ChatGPT offers significant advantages and can contribute to various fields of study. Notably, it assists experts across different disciplines in composing reports for various experiments. For instance, Aydın and Karaarslan (2023) found that ChatGPT can be a valuable tool for paraphrasing and academic writing in the healthcare field. Similarly, a study by Kumar (2023) demonstrated that within biomedical sciences, ChatGPT can be employed to produce well-organized and grammatically sound English academic writing. Beyond these specific examples, ChatGPT offers broader advantages for users. It can address complex inquiries by providing comprehensive insights, ranging from general overviews to detailed analyses of intricate phenomena (Tan et al., 2023). Overall, the current body of research suggests that ChatGPT can be a valuable digital resource across diverse fields of study.

The literature highlights education as one of the primary domains where ChatGPT can make significant contributions (e.g., Bitzenbauer, 2023; Poole, 2022; Rudolph et al., 2023; Su & Yang, 2023). This potential, however, necessitates responsible use. Studies by Bitzenbauer (2023) suggest that ChatGPT can enhance critical thinking skills among secondary school students in Germany. In another study, Poole (2022) reported that ChatGPT benefits language teachers by assisting them in designing exercises and lesson plans. Additionally, ChatGPT can empower teachers to create personalized learning experiences and exercises tailored to individual student needs (Su & Yang, 2023). Furthermore, ChatGPT has the potential to revolutionize higher education, particularly in assessment, learning, and teaching methodologies (Rudolph et al., 2023).

While the integration of ChatGPT into the education sector appears promising, Su and Yang (2023) advocate for careful consideration of several factors to maximize its effectiveness. These factors include determining the expected outcome, defining the appropriate level of automation, considering both the ethical and unethical aspects of use, and measuring the efficacy of ChatGPT in achieving the desired learning objectives.

The recommendations outlined by Su and Yang (2023) for the field of education can be broadly applied to various fields of study where experts leverage ChatGPT for different purposes. In other words, it is crucial for experts in different fields to first determine their desired outcomes and then carefully consider the level of automation, ethical implications, and overall effectiveness of ChatGPT within their specific contexts. As an example, General Practitioners (GPs)

writing reports to patients or colleagues could benefit from evaluating the following criteria: (1) identifying the clear purpose of the report, (2) considering the limitations and advantages of using ChatGPT for report writing (including the level of automation and ethical considerations), and (3) determining the appropriate level of human intervention to ensure accuracy, professionalism, and adherence to ethical guidelines. By following this approach, GPs can optimize the use of ChatGPT for report writing while maintaining control and responsibility for the final content.

Despite the promising applications, there is still a limited comprehensive understanding of ChatGPT's limitations and opportunities across fields, based on a synthesis of empirical findings. This systematic review aims to address this gap by critically examining existing empirical studies on ChatGPT.

## 3. Method

The current systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for conducting and reporting systematic reviews, ensuring a transparent and methodologically rigorous approach throughout the review process (Page et al., 2021).

### 3.1. Search strategies

A comprehensive search strategy was implemented to locate primary studies that report empirical evidence on the limitations of ChatGPT. The search was conducted across various databases including Web of Science, Scopus, IEEE Xplore, and ERIC. Keywords utilized in our search comprised "ChatGPT" AND "limitations," along with synonyms such as "weakness," "drawback," "challenge," and "pitfall" (for detailed search strings, refer to [Appendix A](#)).

For establishing inclusion and exclusion criteria, studies were considered eligible if they were published as journal articles, provided empirical findings assessing ChatGPT's performance, and included a discussion on its limitations (details in [Table 1](#)). Publications characterized as conceptual or lacking an empirically supported examination of ChatGPT's performance and limitations were excluded. No date constraints were imposed on the search process. Results were managed using Covidence, a web-based systematic review management tool facilitating deduplication and screening.

The initial screening involved evaluating titles and abstracts collaboratively by all three authors to identify potentially relevant studies. Subsequently, the screening process progressed to a full-text evaluation of studies identified during the initial screening phase. This thorough examination was conducted in duplicate by the authors to ensure rigor and comprehensiveness. Any discrepancies encountered during the initial and full-text screening were resolved through discussion and consensus among the authors or, when necessary, by consulting an additional member of the research team. This systematic approach aimed to enhance

**Table 1.** Inclusion and exclusion criteria.

Criteria	Inclusion	Exclusion
Type of publication	<ul style="list-style-type: none"> <li>Journal article</li> </ul>	<ul style="list-style-type: none"> <li>Reviews</li> <li>Conceptual papers</li> <li>Editorial</li> <li>Conference proceedings</li> <li>Reports</li> <li>Thesis</li> </ul>
Study Focus	<ul style="list-style-type: none"> <li>Reporting empirical findings</li> <li>Conducting evaluation of ChatGPT's performance</li> </ul>	<ul style="list-style-type: none"> <li>Conceptual papers</li> <li>Papers that solely report perception and attitude data without an evaluation of ChatGPT's performance</li> </ul>
Language of publication	English	Other languages
Date range	No date constraint	

the reliability and precision of the study selection process in the systematic review.

### 3.2. Data analysis

A thematic analysis, following the guidelines outlined by Braun and Clarke (2006), was utilized to uncover and categorize the reported limitations of ChatGPT across the diverse studies included. This methodical approach involved thoroughly familiarizing ourselves with each study to gain a deep understanding of ChatGPT's limitations. Subsequently, initial codes were systematically generated to organize key concepts into a structured data extraction table. Importantly, data extraction was conducted in duplicate by two of the authors to ensure precision and reliability in capturing the nuances of ChatGPT's limitations.

As we explored relationships between these codes, initial themes began to emerge, offering a holistic view of recurring patterns that represented more abstract categories of ChatGPT's limitations. The subsequent review and refinement of these themes aimed to ensure clarity and precision in encapsulating the multifaceted challenges identified in the included literature. This qualitative approach, rooted in thematic analysis and bolstered by the dual extraction performed by two authors, provided a rigorous and structured framework for synthesizing the diverse findings across the included studies, contributing to a comprehensive understanding of ChatGPT's limitations.

Data visualizations throughout this review were created using the Matplotlib library in Python (Hunter, 2007).

## 4. Findings

### 4.1. Overview of the included studies

Figure 1 presents the PRISMA flowchart of searching for and screening studies for eligibility in this review. A review of 33 studies identified a diverse range of fields where ChatGPT limitations were investigated (see Figure 2). The most prevalent field of study was health (48.48%), highlighting the growing interest in understanding potential limitations of large language models in critical healthcare applications. This focus on health suggests a cautious approach to ensure responsible use of ChatGPT in this domain. Education (15.15%) emerged as the second most frequent field, indicating concern about potential

shortcomings in educational settings. Engineering (9.09%) and other fields like psychology, chemistry, and physics (all around 3%) were also represented, showcasing a broader exploration of limitations across various disciplines. This distribution underscores the widespread interest in evaluating ChatGPT's limitations across diverse application areas, with a particular emphasis on ensuring its safe and effective use in healthcare and education.

In terms of the version of ChatGPT employed in the studies reviewed, 23 studies (69.7%) did not report the specific version of ChatGPT used. The remaining studies provided version information, with ChatGPT-3 (15.15%) appearing most frequently, followed by versions 3.5 (12.12%) and a single study exploring a combination of 3.5 and 4 (3.03%). To investigate the potential influence of version differences on the limitations identified in this review, we conducted a sub-analysis of studies examining versions 3 and 3.5. This analysis revealed no significant discrepancies from the overall findings on limitations and opportunities discussed below. Due to the limited presence of research on ChatGPT-4 (only one study identified), a similar sub-analysis for this version was not feasible. The limitations associated with the scarcity of research on ChatGPT-4 and the generalizability of the review's conclusions on ChatGPT limitations will be addressed later in our discussion of the limitations of the review.

### 4.2. RQ1: What limitations of ChatGPT are documented in prior empirical literature?

Our analysis of 33 studies identified limitations associated with ChatGPT. The most prevalent limitation concerned accuracy and reliability, with these issues found in 47.06% of the total instances identified across all studies. This highlights ChatGPT's potential to generate misleading or incorrect information. Limitations in critical and problem-solving thinking were present in 22.06% of instances, suggesting shortcomings in handling complex scenarios that require independent analysis. Ethical considerations, including potential biases and discriminatory outputs due to training data, were observed in 13.24% of instances, raising concerns about ethical, legal, and privacy issues. Furthermore, limitations in understanding context and suitability for in-depth exploration of specialized topics were found in 11.76% of ChatGPT interactions, potentially leading to adverse effects on users' learning and development. Finally, 10.29% of

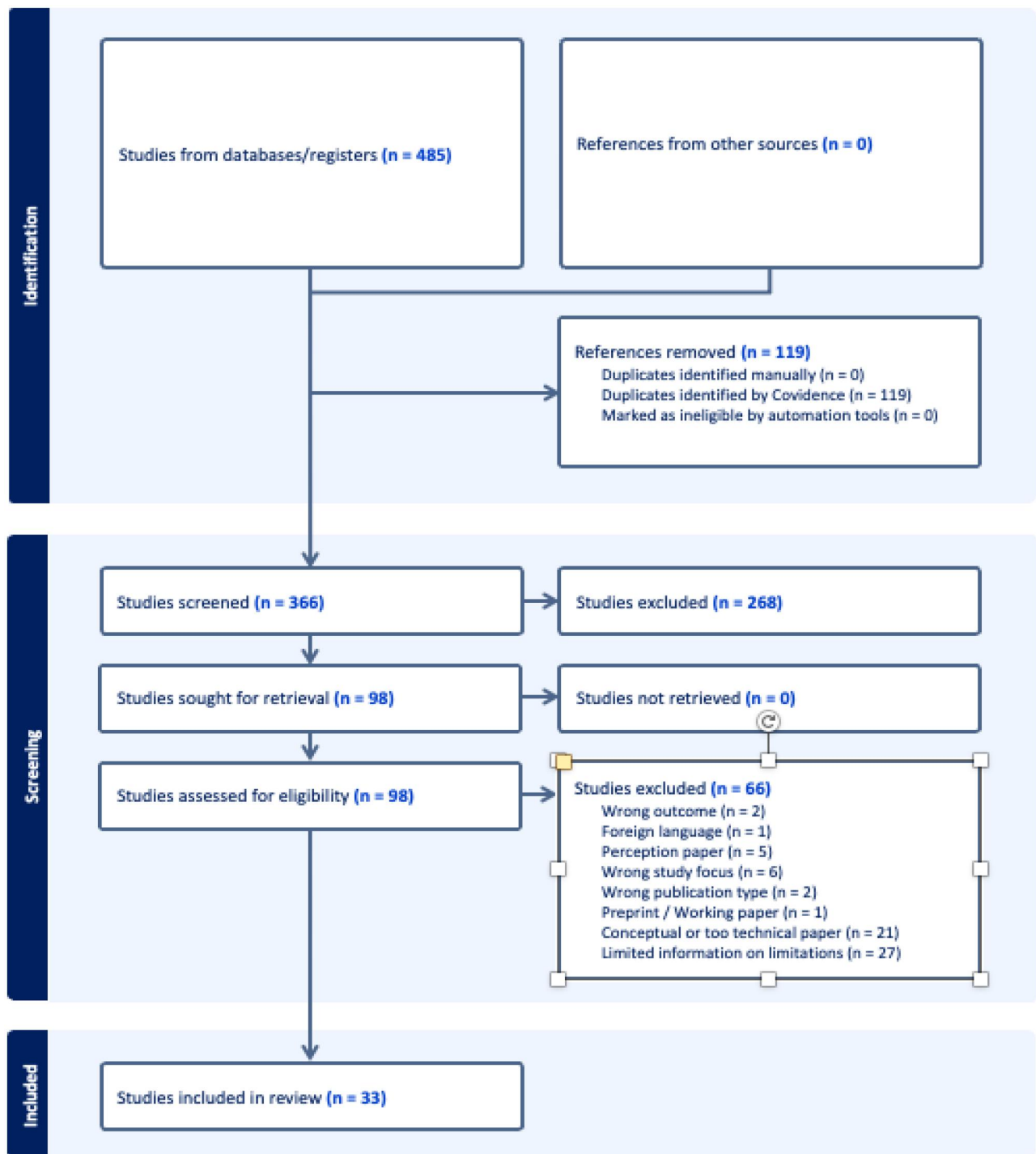


Figure 1. Flowchart of stages of the review.

instances suggested limitations in handling diverse inputs and outputs, potentially hindering the usability of ChatGPT for complex tasks. These findings underscore the need for continued development and responsible use of large language models like ChatGPT (Table 2 and Figure 3).

#### 4.2.1. Accuracy and reliability concerns

ChatGPT faces a significant limitation concerning its accuracy and reliability, particularly evident in evaluations within the health science domain (Ali, 2023; Ariyaratne et al., 2023; Clark, 2023). Ali's (2023) examination revealed significant factual

inaccuracies, with ChatGPT providing unreliable and poorly informed responses, especially on contentious health issues. Wagner and Ertl-Wagner's findings (2023) further underscored this concern, indicating up to 33% of responses by ChatGPT's to radiology questions were inaccurate, highlighting a substantial deficiency in accuracy within the medical domain.

Au Yeung et al. (2023) tasked ChatGPT with predicting medical diagnoses based on clinical histories. While the AI provided overall high-quality responses in terms of relevance (83%), it missed crucial diagnoses in 60% of its outputs. This deficiency poses a significant risk, particularly in



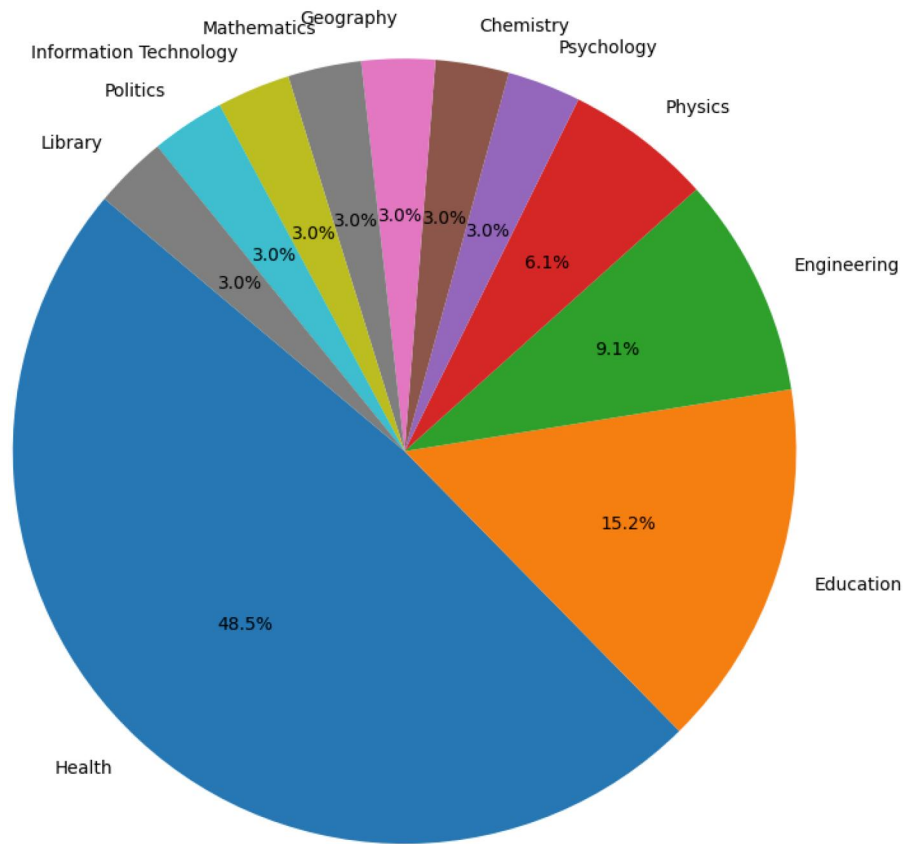


Figure 2. Distribution of the fields of research evaluating ChatGPT's limitations.

Table 2. Summary of the limitations of ChatGPT.

Types of ChatGPT limitations	Frequency (percentage)
Accuracy and reliability	32 (47.06%)
Critical and problem-solving thinking	15 (22.06%)
Ethical, legal, and privacy concerns	9 (13.24%)
Potential adverse effect on users' learning and development	8 (11.76%)
Input and output technical constraints	7 (10.29%)

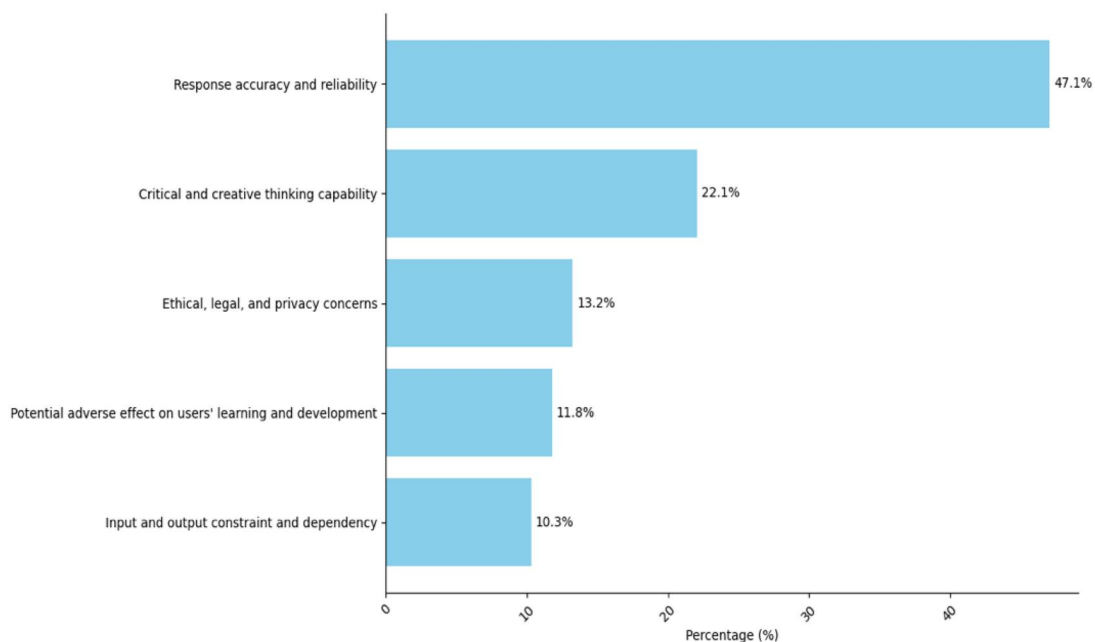


Figure 3. Limitations of ChatGPT ( $n = 71$  instances of limitations).

healthcare contexts, where ChatGPT is likely to generate misleading outputs, potentially perpetuating harmful health beliefs or reinforcing biases.

Fergus et al.'s (2023) study in the pharmaceutical program domain of chemistry found inconsistencies in ChatGPT's responses to test questions. Each answer contained a different error attributed to technical randomness. Similarly, Hoch et al.'s (2023) medical quiz study revealed significant domain-specific variations in ChatGPT's performance. It achieved a 72% accuracy rate for allergology, a field studying hypersensitive reactions of the immune system, questions, and the rest of the responses were inaccurate (Hoch et al., 2023).

Beyond health science, accuracy concerns persist. Clark's (2023) evaluation in a chemistry test resulted in a concerning accuracy rate: only 44% of responses were correct, falling below the average score of participants. This inaccuracy extended to medical assessments, with ChatGPT falling short of the required score in the GP test. Reliability issues were noted by Clark (2023), Duong and Solomon (2023), and Seth et al. (2023), highlighting inconsistencies in ChatGPT's answers to identical questions and criticizing its suitability as a source of sample answers for examinations. Lai (2023) explored the AI chatbot's potential use in addressing inquiries of library service users and found that it performed poorly on advanced research questions, complex inquiries, and queries involving locally specific information.

Seth et al. (2023) further exposed a troubling aspect of ChatGPT's behavior—the generation of fake references, labeled as hallucination of references. Similar findings on hallucinations of the chatbot were reported in the study by McIntosh et al. (2024). Wagner and Ertl-Wagner (2023) discovered that 63.8% of ChatGPT's references in response to radiology questions were fabricated, accentuating broader reliability concerns. Hoch's extensive study involving a medical board certification test revealed domain-specific performance challenges, with ChatGPT's accuracy varying significantly by domain.

In summary, the themes of accuracy and reliability emerge as prominent limitations in ChatGPT. These limitations encompass technical inaccuracies, inconsistencies in responses, and domain-specific performance challenges.

#### **4.2.2. Limitations in critical thinking and problem-solving**

A second limitation concerns ChatGPT's capability for accomplishing critical thinking, problem-solving, and mathematical tasks (Cascella et al., 2023; Clark, 2023; Giannos & Delardas, 2023). Cascella et al.'s (2023) evaluation, involving the composition of a medical note, highlighted deficiencies in addressing causal relations among health conditions, indicating inadequacy in complex reasoning. Clark (2023) emphasized the model's proficiency in addressing general questions over problem-solving or skill-specific queries, while Duong and Solomon's (2023) study revealed ChatGPT's preference for memory-based questions rather than critical thinking tasks. Sanmarchi et al. (2023) assessed ChatGPT's ability to design studies and suggest plastic surgery options, revealing limitations in constructing conceptual frameworks and narrative structures. Seth et al.'s (2023) examination of ChatGPT's responses to plastic surgery

questions highlighted inadequacies in addressing specialized topics, particularly critical thinking skills for complex issues like thumb arthritis.

In educational contexts, Giannos and Delardas (2023) reported ChatGPT's subpar performance on critical thinking and mathematical questions, with more incorrect than correct responses. Parsons and Curry's (2024) evaluation echoed these concerns. They assessed ChatGPT's capability in completing a graduate instructional design assignment for a 12th-grade media literacy course. The chatbot primarily provided superficial information and demonstrated a limited capacity to customize its responses or justify them with details. Rahman and Watanobe's (2023) scrutiny of ChatGPT's mathematical capabilities found dissatisfaction in generating codes and correcting errors, exposing weaknesses in basic mathematical tasks. Kortemeyer (2023) further found that ChatGPT narrowly passed the introductory course in Physics and exhibited "many of the preconceptions and errors of a beginning learner" (p. 1).

The problem-solving capability of ChatGPT for coding practices was also questioned in the study by Shoufan (2023), where the chatbot showed inconsistent responses and struggled to complete the given codes, even the ones it generated itself. Collectively, these findings underscore ChatGPT's limited capacity for critical thinking and problem-solving across diverse domains.

#### **4.2.3. Multifaceted impact on learning and development**

This section explores the multifaceted impact ChatGPT's responses might have on users' learning and development. Concerns include learners' potential overreliance on the tool leading to declines in critical thinking skills. Additionally, the risk of bias and incomplete information in ChatGPT's responses is another consideration. Finally, the potential psychological effects on vulnerable individuals seeking interaction and decision-making support from the AI tool warrant consideration.

In the realm of education, Alafnan et al. (2023) highlighted the positive impact of ChatGPT in providing reliable input to answer test questions. However, they cautioned against overreliance and irresponsible use of the AI tool, emphasizing the potential consequences of "human unintelligence and unlearning" if not used judiciously (p. 60). Clark (2023) echoed concerns about overreliance on ChatGPT, suggesting that excessive dependence could result in passivity and a decline in critical thinking skills among learners. Notably, the challenge of detecting logical fallacies in ChatGPT is a particular concern. The model's ability to provide seemingly logical explanations, even when flawed, may mislead users who lack specific expertise in the subject matter.

Giannos and Delardas (2023) assessed ChatGPT's capability for education and test preparation, concluding that while the AI chatbot is adept at providing tutoring support for general problem solving and reading comprehension, its limitations in scientific and mathematical knowledge and skills render it an unreliable independent tool for supporting students. They also underscored the potential for misuse, highlighting concerns about cheating and gaining unfair advantages during standardized admission tests.

Ibrahim et al. (2023) raised the issue of potential bias in ChatGPT's responses, asserting that the model might be influenced by the dataset used for training, aligning more closely with the political and philosophical values of Western and more developed countries. Sallam et al. (2023) echoed these concerns, particularly in medical education, where biased, outdated, and incomplete content in ChatGPT's responses could pose risks to learners. They noted potential adverse consequences, including discouraging critical thinking and communication skills among medical students.

Additionally, Stojanov (2023) warned of the psychological impact on vulnerable individuals, such as those grieving or extremely shy, who may turn to ChatGPT for solace and interaction. Stojanov also highlighted the risk of individuals relying on the AI tool for crucial life decisions, potentially weakening their personal agency and responsibility. These varied concerns collectively emphasize the need for a cautious and informed approach to the integration of ChatGPT in educational settings.

#### 4.2.4. Technical constraints related to input and output

The effectiveness of ChatGPT is further contingent upon technical constraints of its input and output. This limitation of ChatGPT-3 and ChatGPT-3.5 poses challenges, particularly in disciplines like mathematics and chemistry, where communication often involves signs and symbols. Fergus et al. (2023) conducted examinations in the field of chemistry, revealing instances where ChatGPT struggled, particularly in tasks requiring the drawing of structures between reactants and products.

Furthermore, the efficacy of ChatGPT is influenced by the type of questions posed to it. Notably, the chatbot exhibited a significantly higher performance when responding to single-choice questions compared to multiple-choice questions (Hoch et al., 2023). In an extensive study encompassing 2,576 questions, Hoch et al. (2023) observed a 63% accuracy rate for single-choice questions, in contrast to a 34% accuracy rate for multiple-choice questions.

The phrasing of prompts for ChatGPT responses is also a pivotal factor affecting the chatbot's performance. Sallam et al. (2023) acknowledged that the formulation of prompts, coupled with the word limit imposed on ChatGPT's output, could influence the amount of information generated, subsequently impacting the clarity and effectiveness of the responses. Similarly, Stojanov (2023) reported that ChatGPT's inherent word limit in its output may result in responses containing incomplete information, posing challenges to comprehension.

#### 4.2.5. Ethical, legal and privacy concerns

Previous studies have addressed academic integrity, legal, privacy, and ethical concerns associated with the use of ChatGPT (Au Yeung et al., 2023; Alafnan et al., 2023; Ibrahim et al., 2023; Sallam et al., 2023; Sanmarchi et al., 2023). Academic integrity emerges as a prominent concern, particularly in light of the challenges posed by most plagiarism detection software in identifying content generated by ChatGPT. Fergus et al. (2023) conducted an examination using Turnitin to assess

plagiarism in ChatGPT's output, concluding that the Turnitin report failed to raise any alerts necessitating further investigation into academic integrity (p. 1674). This inability to detect generated content raises concerns about the potential misuse of ChatGPT and its impact on academic honesty.

Furthermore, educators face challenges in distinguishing between students' original work and content generated by ChatGPT, making assessment of individual abilities more complex. Alafnan et al. (2023) argued that the high accuracy and reliability of ChatGPT's responses may impede instructors' ability to differentiate between independently working students and those heavily reliant on automation. This, in turn, can compromise the evaluation of learning outcomes, causing a significant challenge in assessing students' performance. The implications of ChatGPT on academic integrity, underscored by these studies, highlighting the need for careful consideration and regulation in its educational use.

Other legal and ethical issues, including privacy and copyright infringements, were also raised in the literature. The answers generated by ChatGPT raise privacy concerns that may lead to further legal ramifications (Au Yeung et al., 2023; Ibrahim et al., 2023; Sallam et al., 2023; Sanmarchi et al., 2023). Notably, the potential biases in ChatGPT's responses, possibly leaning towards specific political parties or perspectives, raise red flags regarding the validity of its content (Au Yeung et al., 2023). Sallam et al. (2023) specifically assessed responses to health and public education prompts, revealing concerns about plagiarism, copyright issues, academic dishonesty, and the absence of personal and emotional interactions, which are essential for communication skills in healthcare education.

### 4.3. RQ2: In light of the limitations identified, what opportunities exist for enhancing the utilization of ChatGPT?

Table 3 presents a list of opportunities for ChatGPT identified in this review, offering actionable insights for capitalizing on its strengths and capabilities.

#### 4.3.1. Educational support and skill development

ChatGPT's impact on education is multifaceted. It provides educational content, aids in learning processes, and contributes to essential skills development. Scholars have discussed various ways ChatGPT can support this domain, including creating course materials, designing lesson plans and assessments, providing feedback, explaining complex knowledge, and personalizing the learning experience (Clark, 2023; Rahman & Watanobe, 2023). Day (2023) suggests using

**Table 3.** Opportunities of ChatGPT identified in the included studies ( $n = 44$  instances of opportunities).

The opportunities of ChatGPT	Frequency	Percentage (%)
Human like interaction and assistance	6	13.64
Education support and skill development	16	36.36
Task automation and workflow enhancement	11	25
Content creation and ideation	4	9.09
Information retrieval and application	7	15.91



ChatGPT to develop writing course materials. Drawing on Vygotsky's sociocultural theory, Stojanov (2023) discusses how ChatGPT could serve as a knowledgeable learning peer, aiding knowledge exploration. Similarly, Rahman et al. (2023) discuss benefits for learners, educators, and researchers. Learners can employ ChatGPT as a learning assistant for exploring complex concepts, problem-solving, and receiving personalized guidance. Educators can leverage ChatGPT for lesson planning, generating customized resources and activities, answering student questions, and assisting with assessment. Researchers can improve their work by using ChatGPT to check and improve writing, request literature summaries, or suggest research ideas.

#### **4.3.2. ChatGPT as a workflow enhancer**

Beyond education, ChatGPT's ability to automate tasks and enhance professional workflows optimizes operational efficiency and resource utilization. In the construction industry, Prieto et al. (2023) tested ChatGPT's application in creating a coherent and logical construction project schedule. Participants found it satisfactory and indicated its potential for automating preliminary and time-consuming tasks. Similarly, Sanmarchi et al. (2023) suggests ChatGPT as a valuable tool for designing research studies and following international guidelines, for both experienced and less experienced researchers.

#### **4.3.3. Information retrieval powerhouse**

ChatGPT's prowess in retrieving and applying information across various domains empowers users with informed decision-making and problem-solving. Alafnan et al. (2023) discovered that ChatGPT has the potential to function as a valuable platform for students seeking information on diverse topics. They asserted that ChatGPT's capabilities could potentially replace traditional search engines by offering students accurate and reliable information. Duong and Solomon (2023) compared ChatGPT's ability to respond to genetics questions against human performance, revealing that the chatbot approached human-level proficiency. Stojanov (2023) discusses how ChatGPT played a crucial role in providing valuable content, aiding in the ongoing pursuit of learning and exploration of new knowledge.

#### **4.3.4. Natural language interaction and assistance**

ChatGPT's ability to engage users in natural conversations and provide human-like assistance positions it as a valuable virtual companion. Lahat et al. (2023) explored using ChatGPT to answer 110 real-life medical questions from patients, finding it relatively useful and satisfactory, albeit with moderate effectiveness. Other scholars interacted with the chatbot for tasks such as creating a construction project (Prieto et al., 2023) or discussing a plastic surgery topic (Seth et al., 2023). Prieto et al. (2023) highlighted that the conversation-based chatbot is advantageous compared to other single-prompted AI tools as it allows users to modify project aspects as needed.

#### **4.3.5. Content creation and ideation**

Finally, ChatGPT facilitates creative content generation, text transformation, and ideation processes, making it a versatile tool for content creators and innovators. Ariyaratne et al. (2023) discussed using ChatGPT for research, suggesting that "the format of articles generated by ChatGPT can be used as a draft template to write an expanded version of the article" (p. 4). Similarly, ChatGPT can be used to enhance research processes by assisting researchers in generating hypotheses, exploring literature, and translating research findings into a more understandable language (Cascella et al., 2023). In education, ChatGPT can be used to create course materials, such as for writing courses (Day, 2023). Regarding ideation capability, Clark (2023) demonstrated that ChatGPT could be used to support problem conceptualization in chemistry education. A similar conclusion is reached in engineering education as Nikolic et al. (2023) indicated that ChatGPT can support students by aiding in the generation of project ideas, providing information, assisting with project structure, delivering summaries, and offering feedback on ethical considerations and workplace health and safety risks associated with their projects. The text transforming function is another advantageous feature of this generative AI tool. Prieto et al. (2023) indicated the use of ChatGPT is useful for transforming research writing into more readily understandable language (Figure 4).

## **5. Discussion**

This systematic review identified five key limitations associated with ChatGPT's application across diverse fields. Accuracy and reliability emerged as a primary concern, particularly in critical domains like healthcare (Fergus et al., 2023). Additionally, limitations were found in ChatGPT's ability to perform complex cognitive tasks such as critical thinking and problem-solving (Clark, 2023). Studies identified potential negative effects on learners' development due to overreliance on the tool, potentially hindering the development of critical thinking skills (Alafnan et al., 2023; Sallam et al., 2023). Finally, ethical considerations surrounding academic integrity, privacy, and copyright infringement emerged as limitations requiring careful attention when deploying ChatGPT in educational and professional settings (Ibrahim et al., 2023; Puthenpura et al., 2023).

The analysis of included studies also revealed five key themes highlighting potential opportunities presented by ChatGPT. One area of potential lies in information retrieval, where research suggests it can be a valuable tool for finding information across various subjects. Another promising area is natural interaction and support, with ChatGPT's ability to hold natural conversations making it a potential candidate as a virtual companion or assistant in fields like medicine (Lahat et al., 2023) and creative endeavors (Seth et al., 2023). Studies also indicate that ChatGPT may automate tasks and improve workflow efficiency (Prieto et al., 2023; Sanmarchi et al., 2023). Within the educational domain, research explores its potential for personalized learning experiences, creating course materials, and supporting

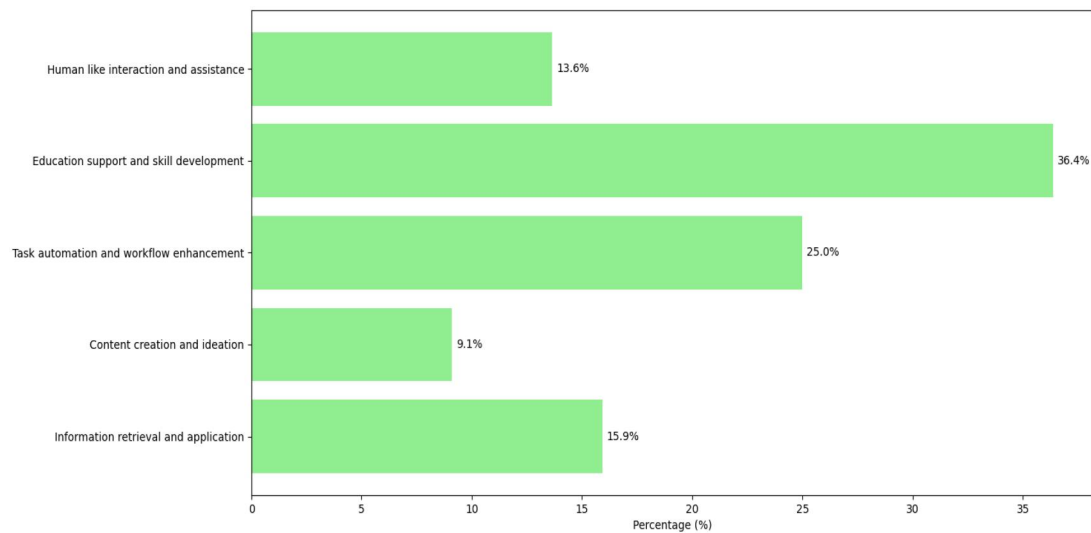


Figure 4. Opportunities for ChatGPT application.

students (Day, 2023; Rahman et al., 2023). Finally, ChatGPT shows promise in creative text generation and supporting brainstorming processes, highlighting its potential as a tool for content creation, research, and generating new ideas (Ariyaratne et al., 2023; Cascella et al., 2023).

The limitations and associated opportunities of ChatGPT identified in this review align with findings from previous reviews on its affordances and limitations e.g., (Aydin & Karaarslan, 2023; Ray, 2023; Sok & Heng, 2024). Aydin and Karaarslan (2023) highlight similar concerns in their review, including ChatGPT's potential bias towards certain political views, its ability to deliver misleading information with equal confidence, and limitations in critical thinking and creativity. Similarly, the opportunities identified in this study resonate with the findings on opportunities in a review by Sok and Heng (2024). They suggest that ChatGPT has the potential to enhance the field of higher education by stimulating innovative assessment methods, improving research writing and design, and boosting productivity. However, the current review, to the authors' best knowledge, is the first systematic review following the PRISMA approach that specifically targets the limitations of ChatGPT, thereby providing more transparent and robust evidence on the limitations and related opportunities of ChatGPT

This review advocates for a cyclical collaborative approach among researchers, practitioners, and developers as essential for the sustainable development of ChatGPT. Grounded in the understanding that ChatGPT presents a double-edged sword, with both opportunities and challenges, expert supervision is crucial (Alafnan et al., 2023; Amin et al., 2023; Au Yeung et al., 2023). To maximize its potential, the proposed model in Figure 5 outlines a three-stage cyclical process. In stage one, researchers can explore methods to optimize ChatGPT's benefits and minimize limitations within specific fields. Stage two involves practitioners applying these research findings in real-world settings. Finally, developers can refine ChatGPT's technical capabilities based on both theoretical advancements and practical feedback from practitioners. This cyclical process

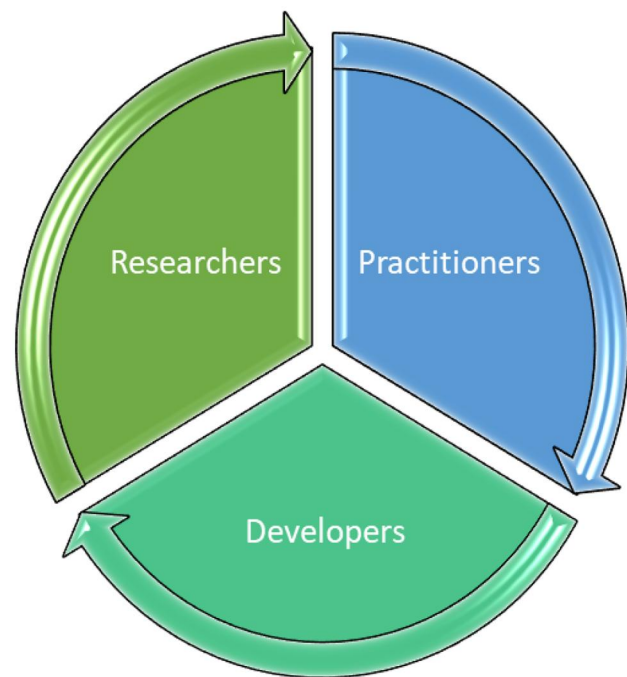


Figure 5. The cyclical evolution of ChatGPT by researchers, practitioners, and developers.

necessitates all parties to remain updated on the latest developments and collaborate to ensure ChatGPT's continued evolution.

The current review is subject to several limitations. First, while this systematic review identified a substantial pool of 485 studies on ChatGPT limitations through searches in Scopus, Web of Science, ERIC, and IEEE databases, it should be acknowledged that it may have overlooked potential studies not indexed within these major databases. Second, while this review offers a comprehensive analysis of ChatGPT's limitations and opportunities, it focuses on the reported findings within the included studies and does not assess their methodological quality. This limits the ability to definitively determine the generalizability of the findings or identify potential biases within the research. Third, the

generalizability of the identified limitations to more advanced iterations of ChatGPT, such as version 4, remains unclear. While the studies included explored limitations of earlier versions (3 and 3.5), it is uncertain if these limitations persist or change in the latest iteration. Additionally, a significant portion (69.7%) of the reviewed studies did not report the specific ChatGPT version used. This lack of transparency hinders our ability to definitively assess how limitations might vary across different versions.

However, while ChatGPT-4 reportedly leverages larger datasets, potentially leading to enhanced performance and incorporating plugin functionalities, previous scholars indicated that many limitations identified in ChatGPT-3.5 are still applicable to it. While advancements have been made, OpenAI (2023) acknowledges ChatGPT-4 still exhibits limitations from earlier versions, including hallucinations, unreliability, and a limited context window, and lacks the ability to learn from experience. Supporting this, Suchman et al. (2023) found no demonstrable advantage for ChatGPT-4 in a medical test, even showing a performance deficit compared to the free version (ChatGPT-3.5) on gastroenterology self-assessment tests.

Regarding future research directions, it is crucial for researchers to explicitly report the specific version of ChatGPT used in their studies to enhance the generalizability and reliability of research findings in the future. This facilitates comparisons across studies and allows for a more nuanced understanding of how limitations evolve across ChatGPT versions. Next, developing best practices for educators, assessment methods that leverage ChatGPT's strengths, and research on its impact on learning outcomes are essential next steps. Finally, integration with specific domains presents a promising avenue for future research. Investigating the potential of integrating ChatGPT with specialized tools in various contexts, along with domain-specific training methods and the associated ethical considerations, is recommended.

## 6. Conclusion

This systematic review examined limitations and opportunities associated with ChatGPT's application across various fields. By analyzing 33 carefully screened empirical studies, it offers a comprehensive picture of ChatGPT's capabilities. The review identified five key limitations: accuracy concerns in critical domains like healthcare, limitations in complex cognitive tasks, potential negative impacts on learners' development due to overreliance, and ethical considerations surrounding privacy, copyright, and academic integrity. However, the review also highlights five opportunities. ChatGPT has the potential to be a valuable tool for users seeking information across various domains. Its ability to engage in natural conversations positions it as a potential virtual companion or assistant. The review also found promise in its ability to automate tasks and enhance workflows, leading to improved efficiency. Within education, ChatGPT presents opportunities for personalized learning experiences, course material creation, and student support. Finally, the

review suggests promise in the ability of ChatGPT to generate creative text formats and support ideation processes, highlighting its potential as a tool for content creation, research, and brainstorming. By acknowledging both limitations and opportunities, this review offers valuable insights for researchers, developers, and users to consider when exploring the potential and responsible application of ChatGPT.

## Disclosure statement

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

## ORCID

Ngo Cong-Lem  <http://orcid.org/0000-0002-5257-8264>

Ali Soyoof  <http://orcid.org/0000-0002-8037-5632>

Diki Tsering  <http://orcid.org/0000-0003-0157-4009>

## References

- Alafnan, M. A., Dishari, S., Jovic, M., & Lomidze, K. (2023). ChatGPT as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses. *Journal of Artificial Intelligence and Technology*, 3(2), 60–68. <https://doi.org/10.37965/jait.2023.0184>
- Ali, M. J. (2023). ChatGPT and lacrimal drainage disorders: Performance and scope of improvement. *Ophthalmic Plastic and Reconstructive Surgery*, 39(5), 515–514. <https://doi.org/10.1097/IOP.0000000000002418>
- Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intelligent Systems*, 38(2), 15–23. <https://doi.org/10.1109/MIS.2023.3254179>
- Ariyaratne, S., Iyengar, K. P., Nischal, N., Chitti Babu, N., & Botchu, R. (2023). A comparison of ChatGPT-generated articles with human-written articles. *Skeletal Radiology*, 52(9), 1755–1758. <https://doi.org/10.1007/s00256-023-04340-5>
- Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R. J., & Teo, J. T. (2023). AI chatbots not yet ready for clinical use. *Frontiers in Digital Health*, 5, 1161098. <https://doi.org/10.3389/fdgh.2023.1161098>
- Aydin, Ö., & Karaarslan, E. (2023). Is ChatGPT leading generative AI? What is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems*, 11(3), 118–134. <https://doi.org/10.21541/apjess.1293702>
- Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, 15(3), ep430. <https://doi.org/10.30935/cedtech/13176>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. <https://doi.org/10.48550/ARXIV.2303.12712>
- Cadamuro, J., Cabitza, F., Debeljak, Z., De Bruyne, S., Frans, G., Perez, S. M., Ozdemir, H., Tolios, A., Carobene, A., & Padoan, A. (2023). Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group. *Clinical Chemistry and Laboratory Medicine*, 61(7), 1158–1166. <https://doi.org/10.1515/cclm-2023-0355>



- Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1), 33. <https://doi.org/10.1007/s10916-023-01925-4>
- Clark, T. M. (2023). Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *Journal of Chemical Education*, 100(5), 1905–1916. <https://doi.org/10.1021/acs.jchemed.3c00027>
- Day, T. (2023). A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT. *The Professional Geographer*, 75(6), 1024–1027. <https://doi.org/10.1080/00330124.2023.2190373>
- Duong, D., & Solomon, B. D. (2023). Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics*, 32(4), 466–468. <https://doi.org/10.1038/s41431-023-01396-8>
- Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education*, 100(4), 1672–1675. <https://doi.org/10.1021/acs.jchemed.3c00087>
- Giannos, P., & Delardas, O. (2023). Performance of ChatGPT on UK standardized admission tests: Insights from the BMAT, TMUA, LNAT, and TSA examinations. *JMIR Medical Education*, 9, e47737. <https://doi.org/10.2196/47737>
- Gregorcic, B., & Pendrill, A. M. (2023). ChatGPT and the frustrated Socrates. *Physics Education*, 58(3), 035021. <https://doi.org/10.1088/1361-6552/acc299>
- Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: How AI-assisted conversational interfaces are revolutionizing the field. *Big Data and Cognitive Computing*, 7(2), 62. <https://doi.org/10.3390/bdcc7020062>
- Hoch, C. C., Wollenberg, B., Lüers, J. C., Knoedler, S., Knoedler, L., Frank, K., Cotofana, S., & Alfertshofer, M. (2023). ChatGPT's quiz skills in different otolaryngology subspecialties: An analysis of 2576 single-choice and multiple-choice board certification preparation questions. *European Archives of Oto-Rhino-Laryngology*, 280(9), 4271–4278. <https://doi.org/10.1007/s00405-023-08051-4>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Ibrahim, H., Asim, R., Zaffar, F., Rahwan, T., & Zaki, Y. (2023). Rethinking homework in the age of artificial intelligence. *IEEE Intelligent Systems*, 38(2), 24–27. <https://doi.org/10.1109/MIS.2023.3255599>
- Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, 19(1), 1–18. <https://doi.org/10.1103/PhysRevPhysEducRes.19.010132>
- Kumar, H. A. (2023). Analysis of chatgpt tool to assess the potential of its utility for academic writing in biomedical domain. *Biology, Engineering, Medicine and Science Reports*, 9(1), 24–30. <https://doi.org/10.5530/bems.9.1.5>
- Lahat, A., Shachar, E., Avidan, B., Glicksberg, B., & Klang, E. (2023). Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: Are we there yet? *Diagnostics*, 13(11), 1950. <https://doi.org/10.3390/diagnostics13111950>
- Lai, K. (2023). How well does ChatGPT handle reference inquiries? An analysis based on question types and question complexities. *College & Research Libraries*, 84(6), 974–995. <https://doi.org/10.5860/crl.84.6.974>
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. N. (2024). A culturally sensitive test to evaluate nuanced GPT hallucination. *IEEE Transactions on Artificial Intelligence*, 1–13. <https://doi.org/10.1109/TAI.2023.3332837>
- Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., Lyden, S., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*, 48(4), 559–614. <https://doi.org/10.1080/03043797.2023.2213169>
- OpenAI. (2023). *GPT-4 technical report*. <https://cdn.openai.com/papers/gpt-4.pdf>
- Page, M. J., McKenzie, J. E., 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., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, 88(2021), 105906. <https://doi.org/10.1016/j.ijsu.2021.105906>
- Parsons, B., & Curry, J. H. (2024). Can ChatGPT pass graduate-level instructional design assignments? Potential implications of artificial intelligence in education and a call to action. *TechTrends*, 68(1), 67–78. <https://doi.org/10.1007/s11528-023-00912-3>
- Poole, F. (2022). *Using Chatgpt to design language material and exercises*. <https://fltmag.com/chatgpt-design-material-exercises/>
- Prieto, S. A., Mengiste, E. T., & García de Soto, B. (2023). Investigating the use of ChatGPT for the scheduling of construction projects. *Buildings*, 13(4), 857. <https://doi.org/10.3390/buildings13040857>
- Puthenpura, V., Nadkarni, S., DiLuna, M., Hieftje, K., & Marks, A. (2023). Personality changes and staring spells in a 12-year-old child: A case report incorporating ChatGPT, a natural language processing tool driven by artificial intelligence (AI). *Cureus*, 15(3), e36408. <https://doi.org/10.7759/cureus.36408>
- Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9), 5783. <https://doi.org/10.3390/app13095783>
- Rahman, M., Terano, H. J. R., Rahman, N., Salamzadeh, A., & Rahaman, S. (2023). Chatgpt and academic research: A review and recommendations based on practical examples. *Journal of Education, Management and Development Studies*, 3(1), 1–12. <https://doi.org/10.52631/jemds.v3i1.175>
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
- Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, 12(3), 148. <https://doi.org/10.3390/socsci12030148>
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, 6(1), 342–363. <https://doi.org/10.37074/jalt.2023.6.1.9>
- Sallam, M., Salim, N., Barakat, M., & Al-Tammemi, A. (2023). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J*, 3(1), e103. <https://doi.org/10.52225/narra.v3i1.103>
- Sanmarchi, F., Bucci, A., Nuzzolese, A. G., Carullo, G., Toscano, F., Nante, N., & Golinelli, D. (2023). A step-by-step researcher's guide to the use of an AI-based transformer in epidemiology: An exploratory analysis of ChatGPT using the STROBE checklist for observational studies. *Journal of Public Health*, 1–36. <https://doi.org/10.1007/s10389-023-01936-y>
- Segal, S., & Khanna, A. K. (2023). Anesthetic management of a patient with juvenile hyaline fibromatosis: A case report written with the assistance of the large language model chatgpt. *Cureus*, 15(3), e35946. <https://doi.org/10.7759/cureus.35946>
- Seth, I., Sinkjær Kenney, P., Bulloch, G., Hunter-Smith, D. J., Bo Thomsen, J., & Rozen, W. M. (2023). Artificial or augmented authorship? A conversation with a Chatbot on base of thumb arthritis. *Plastic and Reconstructive Surgery. Global Open*, 11(5), e4999. <https://doi.org/10.1097/GOX.0000000000004999>
- Shoufan, A. (2023). Can students without prior knowledge use ChatGPT to answer test questions? An empirical study. *ACM Transactions on Computing Education*, 23(4), 1–29. <https://doi.org/10.1145/3628162>

- Singh, H., & Singh, A. (2023). Chatgpt: Systematic review, applications, and agenda for multidisciplinary research. *Journal of Chinese Economic and Business Studies*, 21(2), 193–212. <https://doi.org/10.1080/14765284.2023.2210482>
- Sok, S., & Heng, K. (2024). Opportunities, challenges, and strategies for using ChatGPT in higher education: A literature review. *Journal of Digital Educational Technology*, 4(1), ep2401. <https://doi.org/10.30935/jdet/14027>
- Stojanov, A. (2023). Learning with ChatGPT 3.5 as a more knowledgeable other: An autoethnographic study. *International Journal of Educational Technology in Higher Education*, 20(1), 35. <https://doi.org/10.1186/s41239-023-00404-7>
- Su, J., & Yang, W. (2023). Unlocking the power of chatgpt: A framework for applying generative ai in education. *ECNU Review of Education*, 6(3), 355–366. <https://doi.org/10.1177/20965311231168423>
- Suchman, K., Garg, S., & Trindade, A. J. (2023). Chat generative pre-trained transformer fails the multiple-choice American College of Gastroenterology Self-Assessment Test. *The American Journal of Gastroenterology*, 118(12), 2280–2282. <https://doi.org/10.14309/ajg.0000000000002320>
- Tan, T. F., Thirunavukarasu, A. J., Campbell, J. P., Keane, P. A., Pasquale, L. R., Abramoff, M. D., Kalpathy-Cramer, J., Lum, F., Kim, J. E., Baxter, S. L., & Ting, D. S. W. (2023). Generative artificial intelligence through chatgpt and other large language models in ophthalmology: Clinical applications and challenges. *Ophthalmology Science*, 3(4), 100394. <https://doi.org/10.1016/j.xops.2023.100394>
- Thirunavukarasu, A. J., Hassan, R., Mahmood, S., Sanghera, R., Barzangi, K., El Mukashfi, M., & Shah, S. (2023). Trialling a large language model (ChatGPT) in General practice with the applied knowledge test: Observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education*, 9, e46599. <https://doi.org/10.2196/46599>
- Wagner, M. W., & Ertl-Wagner, B. B. (2023). Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Canadian Association of Radiologists Journal*, 75(1), 69–73. <https://doi.org/10.1177/08465371231171125>

## About the authors

**Ngo Cong-Lem** is a Research Fellow at BehaviourWorks Australia, Monash University and a lecturer at the Faculty of Foreign Languages, Dalat University. His research interests involve educational technologies, evidence synthesis and translation, second language studies, and continuing professional learning.

**Ali Soyoo** is a research fellow at University of Macau. His field of interest is Computer Assisted Language Learning (CALL), digital

games, second language vocabulary learning, and Informal Digital Language Learning of English (IDLE).

**Diki Tsering** is a research officer at Monash Sustainable Development Institute. Her main interest lies in applying systematic review principles to deliver high-quality evidence reviews that translate research knowledge into practice and make positive contributions to society.

## Appendix A

### Databases and Search Strings

Database: **Scopus**

Date of Search: 26 June 2023

Yield: 169

Search string: TITLE-ABS-KEY (chatgpt AND (limitation\* OR challenge\* OR drawback\* OR problem\* OR challenge\* OR issue\* OR concern\* OR risk\* OR disadvantage\* OR flaw\* OR weakness\* OR shortcoming\* OR pitfall\* OR downside\* OR bias\* OR error\* OR ethic\*)) AND (LIMIT-TO (DOCTYPE, "ar"))

Database: **Web of Science**

Date of Search: 26 June 2023

Yield: 150

Search string: TS=(ChatGPT AND (limitation\* OR challenge\* OR drawback\* OR problem\* OR challenge\* OR issue\* OR concern\* OR risk\* OR disadvantage\* OR flaw\* OR weakness\* OR shortcoming\* OR pitfall\* OR downside\* OR bias\* OR error\* OR ethic\*))

Database: **ERIC**

Search Date: 09 March 2024

Yield: 108

Search string: ChatGPT AND (limitation\* OR challenge\* OR drawback\* OR problem\* OR issue\* OR concern\* OR risk\* OR disadvantage\* OR flaw\* OR weakness\* OR shortcoming\* OR pitfall\* OR downside\* OR bias\* OR error\* OR ethic\*)

Database: **IEEE Xplore**

Search Date: 09 March 2024

Yield: 58 (Filters applied: 'Journals' and 'Early Access Articles')

Search string: ("ChatGPT" AND (limitation\* OR challenge\* OR drawback\* OR problem\* OR issue\* OR concern\* OR bias\* OR risk\* OR disadvantage\* OR flaw OR flaws OR weakness OR weaknesses OR shortcoming OR shortcomings OR pitfall OR pitfalls OR downside OR downsides OR error OR errors OR ethic OR ethics))

## Appendix B

### A Summary of the Included Studies

Study, Country, ChatGPT version	Aims	Method	Limitations	Opportunities
Alafnan et al. (2023) NA Kuwait	To investigate opportunities and challenges in using ChatGPT for students and instructions of communication and writing courses	ChatGPT was asked to generate responses to 30 theory-based questions (x 5 times for each question). Its responses were checked by Turnitin software and experts.	<ul style="list-style-type: none"> <li>potential adverse effect on students' learning and development if used inappropriately and unethically</li> </ul>	<ul style="list-style-type: none"> <li>technology-enhanced teaching and input for learning and discussion</li> </ul>
Ali (2023) NA India	To test ChatGPT's knowledge and opinion on a controversial health topic	21 prompts were inputted into the ChatGPT for responses, which were evaluated by experts (i.e., surgeons)	<ul style="list-style-type: none"> <li>several factual inaccuracies;</li> <li>generic response to controversial topics lacking evidence support;</li> <li>below average level of accuracy (40%)</li> </ul>	N/A

(continued)



Continued.

Study, Country, ChatGPT version	Aims	Method	Limitations	Opportunities
Amin et al. (2023) NA Germany	To evaluate ChatGPT's capacity to perform text classification	ChatGPT was asked to predict personalities, sentiment analysis and suicide tendency based on prompts crafted from three relevant datasets on these topics	<ul style="list-style-type: none"> <li>• lower performance compared to a specialized language model (RoBERTa-base)</li> </ul>	<ul style="list-style-type: none"> <li>• decent performance compared to other text classification models (i.e., Word2Vec and bag-of-words baseline)</li> <li>• robustness against noisy data</li> <li>• no training by users needed for ChatGPT</li> </ul>
Ariyaratne et al. (2023) ChatGPT-3 UK	To compare the article writing of ChatGPT and humans	ChatGPT was asked to write about a radiology topic, which was then assessed on a 5 point scale from being bad and inaccurate to being excellent and accurate by radiologists	<ul style="list-style-type: none"> <li>• 4 out of 5 articles written by ChatGPT being factually inaccurate</li> <li>• providing risky medical suggestions</li> <li>• fictitious references</li> </ul>	N/A
Au Yeung et al. (2023) NA UK	To test ChatGPT's capacity to predict medical diagnoses	ChatGPT was given clinical vignettes and asked to predict diagnoses.	<ul style="list-style-type: none"> <li>• One or more critical diagnoses were missing in 60% of responses of ChatGPT</li> <li>• general prediction of diseases only</li> <li>• potential bias in clinical diagnosis against Black people</li> <li>• "takes the truth of prompts at face-value", which influences the accuracy of its response</li> </ul>	N/A
Cadamuro et al. (2023) NA Austria, Italy, Croatia	To test ChatGPT's capability to interpret laboratory test results	ChatGPT was asked to interpret 10 simulated laboratory reports, drafted as optimized prompts. Its output was evaluated by experts in terms of relevance, accuracy, helpfulness and safety	<ul style="list-style-type: none"> <li>• superficial interpretations, most of which lack coherence</li> <li>• more suitable for test-by-test interpretation</li> </ul>	<ul style="list-style-type: none"> <li>• able to recognise all laboratory tests</li> </ul>
Cascella et al. (2023) NA Italy	To test ChatGPT's use in healthcare context	ChatGPT was provided with some input and then asked to: - compose a medical note for a patient admitted to an emergency - write a research conclusion based on some information about the research method and finding - write an abstract based on csv (comma-separated values) formatted data	<ul style="list-style-type: none"> <li>• lack capability in interpreting or explaining causal relations among components/ conditions</li> <li>• no performance of statistical analysis</li> <li>• often not aware of limitations unless requested</li> </ul>	<ul style="list-style-type: none"> <li>• aiding in the research process by generating hypothesis, exploring literature, extracting important information</li> <li>• communicating research findings in a clear and understandable manner</li> </ul>
Clark (2023) NA USA	To examine the capability of ChatGPT in answering a chemistry test	ChatGPT was used to answer closed (multiple choice) and opened ended questions for a chemistry test	<ul style="list-style-type: none"> <li>• inadequate in problem solving or answering questions requiring specific skills</li> <li>• only able to achieve 44% score in the chemistry test, i.e., well below the class's average score at 69%</li> <li>• providing seemingly logical but flawed explanations</li> <li>• not well-equipped for generating sample responses for exam purpose</li> </ul>	<ul style="list-style-type: none"> <li>• potential use to create assignments for students to analyze and improve its response</li> </ul>
Day (2023) NA Canada	To investigate the accuracy of references generated by ChatGPT	ChatGPT was asked to answer questions on various topics commonly of interest to geographers	<ul style="list-style-type: none"> <li>• References generated through a predictive process rather than facts</li> <li>• subject matter knowledge is required to detect incorrect information, a skill students need to develop</li> </ul>	<ul style="list-style-type: none"> <li>• a supporting tool for teaching writing</li> </ul>

(continued)

Continued.

Study, Country, ChatGPT version	Aims	Method	Limitations	Opportunities
Duong and Solomon (2023) NA USA	To assess ChatGPT performance in answering questions related to biomedical field	The responses from ChatGPT to 85 multiple-choice questions on human genetics were contrasted with human responses.	<ul style="list-style-type: none"> <li>not particularly adept at answering critical thinking and calculation-based questions but more suitable for memorisation-based questions</li> <li>inconsistency in answers and explanations where one might select the wrong answer but then provide a correct explanation</li> <li>not suitable for clinical or high-stake uses</li> </ul>	<ul style="list-style-type: none"> <li>rapid and accurate responses to genetic questions</li> <li>potential use to support healthcare professionals in treatment and diagnosis and patients in having accessible medical information</li> </ul>
Fergus et al. (2023) NA UK	To evaluate ChatGPT's response to year end exam assessments	ChatGPT was asked to answer exam questions from two modules of a pharmaceutical program	<ul style="list-style-type: none"> <li>failing to pass year end exams with the total grade on module 1 and 2 being 34.1% and 18.3%, respectively</li> <li>unable to respond to non-text questions</li> <li>ChatGPT-generated texts not being detected by Turnitin</li> </ul>	<ul style="list-style-type: none"> <li>able to provide well-articulated answers to text-based questions</li> <li>catalyst for discussion on academic integrity and assessment design</li> </ul>
Giannos and Delardas (2023) ChatGPT-3.5 UK	To test ChatGPT's performance on several uni admission tests	ChatGPT was asked to respond to 509 multiple-choice questions on various topics. Its responses were evaluated against various skills such as critical thinking, logical thinking, math, problem solving, and reading comprehension.	<ul style="list-style-type: none"> <li>limited scientific and mathematical skills</li> <li>poor performance on critical thinking and reasoning skills</li> <li>providing more incorrect than correct responses</li> </ul>	<ul style="list-style-type: none"> <li>well-written responses</li> <li>a catalyst for redesigning educational assessment</li> </ul>
Gregorcic and Pendrill (2023) NA Sweden	To test effectiveness of ChatGPT in answering basic physics questions	ChatGPT was asked to answer a physics question "A teddy bear is thrown into the air. What is its acceleration in the highest point?"	<ul style="list-style-type: none"> <li>inaccurate responses with contradictions</li> <li>not yet adequate to be a cheating tool for physics student or as a physics tutor</li> </ul>	<ul style="list-style-type: none"> <li>potential use for generating lesson materials</li> </ul>
Hoch et al. (2023) NA Germany	To test ChatGPT's performance on a board certification exam	ChatGPT was asked to answer 2576 single-choice and multiple-choice board certification preparation questions	<ul style="list-style-type: none"> <li>limited performance depending on the test/question format and specific domain of knowledge; more accurate in allergology (72% correct responses) compared to otolaryngology (i.e., 71% answers being incorrect)</li> <li>better performance in answering open ended questions rather than multiple choice questions</li> </ul>	<ul style="list-style-type: none"> <li>a supplementary tool for otolaryngology board certification preparation</li> </ul>
Ibrahim et al. (2023) NA United Arab Emirates	To evaluate potential risk of plagiarism of ChatGPT	ChatGPT was asked to answer questions from two introductory and two advanced tertiary level courses	<ul style="list-style-type: none"> <li>failing to reach the passing grade in the advanced course questions</li> </ul>	<ul style="list-style-type: none"> <li>excellent grade on questions from the introductory courses</li> </ul>
Kortemeyer (2023) NA Switzerland	To assess whether ChatGPT could successfully complete introductory physics courses	ChatGPT's ability to handle calculus-based physics content was assessed by administering representative assessments from a real course. The model's responses were then graded using the same criteria applied to student work.	<ul style="list-style-type: none"> <li>demonstrating beginner-like errors</li> <li>Presenting facts and fiction with similar confidence</li> <li>Probabilistic nature leading to inconsistent results</li> <li>core issues remaining for more newer versions</li> </ul>	The necessity to develop epistemologies when ChatGPT assumes the role of subject matter experts.
Lahat et al. (2023) NA United Arab Emirates	To test ChatGPT's performance in answer patients' real-life questions	ChatGPT was asked to answer 110 real-life questions from the patients	<ul style="list-style-type: none"> <li>moderately accurate and reliable only</li> <li>quality of responses depending on question input</li> </ul>	<ul style="list-style-type: none"> <li>a useful source of reference information</li> </ul>

(continued)

Continued.

Study, Country, ChatGPT version	Aims	Method	Limitations	Opportunities
Lai (2023) ChatGPT 3.5 Canada	To evaluate ChatGPT's proficiency in managing various question types and difficulty levels related to references in library services.	ChatGPT was assigned to answer questions about references received by McGill University's library services. Its responses were subsequently assessed using rubrics that considered completeness, accuracy, and the provision of additional references if the user's inquiry was not fully addressed.	<ul style="list-style-type: none"> <li>Struggling with factual accuracy in its response</li> <li>Difficulty handling advanced questions</li> <li>Failing to answer questions requiring nuance, additional resources and referrals</li> </ul>	Leveraging ChatGPT as a tool for crafting neutral-tone letters and professional responses.
McIntosh et al. (2024) ChatGPT 3.5 and 4.0 Australia & New Zealand	to assess how different GPT models respond to a Culturally Sensitive Test aimed at detecting hallucinations across diverse cultural and linguistic contexts	Different versions of ChatGPT underwent a Culturally Sensitive Test comprising 70 questions spanning real-world contexts. Model responses were scored as 0 for hallucinated and 1 for non-hallucinated answers.	<ul style="list-style-type: none"> <li>Hallucinations</li> <li>Ethical concerns</li> <li>Inconsistent performance</li> </ul>	NA
Nikolic et al. (2023) NA Australia	To examine ChatGPT's response to assessment prompts	ChatGPT was asked to respond to engineering assessment prompts from 10 subjects across 7 Australian universities	<ul style="list-style-type: none"> <li>ChatGPT's response having word limit, often being generic, lacking specific details, fabricating answers and inaccurate calculations</li> <li>requiring pre-training with background information, which can be time consuming</li> </ul>	<ul style="list-style-type: none"> <li>passable responses from ChatGPT with minimised changes to authentic assessment input</li> </ul>
Parsons and Curry (2024) ChatGPT-3 USA	To explore ChatGPT's capability in fulfilling graduate-level instructional design assignments.	This research employed a needs, task, and learner analysis to evaluate ChatGPT's capacity to generate instructional materials for a 12th-grade media literacy module. Expert evaluation and grading rubrics were then used to benchmark the quality of the bot's outputs.	<ul style="list-style-type: none"> <li>Struggling to adapting their response to specific context</li> <li>Only including knowledge prior to September 2021</li> <li>Providing generic and superficial responses when asked to contextualise its responses</li> <li>Responses depending on the complexity and format of questions</li> </ul>	Input for teacher and curriculum specialist training in integrating AI capabilities.
Prieto et al. (2023) ChatGPT-3.5 USA, United Arab Emirates	To test ChatGPT's performance to create a construction project schedule	ChatGPT was asked to generate a construction schedule for a simple project	<ul style="list-style-type: none"> <li>generic responses and fabricating (incorrect) answers</li> <li>quality of response largely depending on the input/ prompt</li> </ul>	<ul style="list-style-type: none"> <li>able to generate coherent schedule to fulfill the task requirements</li> <li>potential for automating preliminary and time-consuming tasks</li> </ul>
Puthenpura et al. (2023) NA USA	To explore the benefit of ChatGPT in assisting writing up a case report	Carefully drafted prompts about a case (i.e., based on case presentation, diagnostic test results and treatment results) was inputted into ChatGPT for response.	<ul style="list-style-type: none"> <li>ChatGPT's response containing incomplete information, which is difficult to interpret without subject matter knowledge</li> <li>reference hallucination</li> <li>plagiarism concerns</li> </ul>	<ul style="list-style-type: none"> <li>an assisting tool in streamlining the writing process</li> </ul>
Rahman and Watanobe (2023) NA Bangladesh, Japan	To evaluate ChatGPT's performance in assisting students in learning coding skills	ChatGPT was asked to generate codes based on clear or partially clear information as well as correct errors in codes	<ul style="list-style-type: none"> <li>poor mathematical skills: failing calculation (counting numbers) that elementary children can do</li> <li>codes generated by ChatGPT may including errors that require human check</li> <li>concerns for students' potential overreliance on ChatGPT</li> </ul>	<ul style="list-style-type: none"> <li>nearly precise responses to technical queries across a diverse array of subjects.</li> </ul>
Rozado (2023) NA New Zealand	To evaluate potential political bias in ChatGPT's response	ChatGPT was asked to answer 15 different political orientation tests	<ul style="list-style-type: none"> <li>ChatGPT consistently demonstrating bias toward left-wing political viewpoints (14/15 tests)</li> </ul>	N/A

(continued)

Continued.

Study, Country, ChatGPT version	Aims	Method	Limitations	Opportunities
Sallam et al. (2023) NA Jordan	To examine pros and cons of ChatGPT in health and public health education	ChatGPT was asked to respond to prompts about medical, dental and pharmacy topics, each with 5 prompts; its responses were then assessed by experts in terms of conciseness, accuracy and clarity	<ul style="list-style-type: none"> <li>inaccurate and biased content</li> <li>privacy concerns</li> <li>students' potential deterioration in critical thinking due to overreliance on ChatGPT</li> </ul>	<ul style="list-style-type: none"> <li>enhancing personalized learning, clinical reasoning skills and understanding of intricate medical concepts</li> </ul>
Sanmarchi et al. (2023) ChatGPT-3 Italy	To examine the potential of ChatGPT in designing and conducting epidemiological study	ChatGPT was asked to suggest study questions and design based on an existing paper and then evaluate its response in light of coherence and relevance (by 3 senior researchers)	<ul style="list-style-type: none"> <li>ethical and legal consequences due to inaccurate data</li> <li>reproducibility issue with ChatGPT due to its inconsistent response</li> <li>inadequate to design conceptual and structure of the study/paper</li> </ul>	<ul style="list-style-type: none"> <li>a valuable support for researchers in setting up an epidemiological study with its response being most effective in method, data analysis and offering recommendations</li> </ul>
Segal and Khanna (2023) NA USA	To investigate both the capabilities and constraints of ChatGPT in aiding the composition of a case report	Asked to compose a text about a case based on crafted prompts with relevant medical information	<ul style="list-style-type: none"> <li>providing erroneous description of the genetics and overestimating the condition of the disease</li> <li>hallucinating references</li> </ul>	<ul style="list-style-type: none"> <li>can be used to generate a rough draft for research and writing purposes</li> </ul>
Seth et al. (2023) ChatGPT-3 Australia, Denmark	To test the value of ChatGPT's input in medical field (i.e. thumb arthritis), particularly for research writing	ChatGPT was asked to answer 5 questions about plastic surgery regarding thumb arthritis	<ul style="list-style-type: none"> <li>superficial information; not creative and cannot be used to generate plastic surgery solutions</li> <li>hallucinating references</li> <li>able to provide accurate and relevant information (albeit superficial)</li> </ul>	N/A
Shoufan (2023) NA United Arab Emirates	to assess ChatGPT's effectiveness in aiding students with no prior knowledge in answering assessment questions	Computer engineering students (experiment group: $n = 41-56$ ) used ChatGPT to answer previous test questions before learning about the related topics. Their scores were compared with those of previous-term students (control group: $n = 24-61$ ) who answered the same questions in a quiz or exam setting.	<ul style="list-style-type: none"> <li>Struggling with tasks involving code completion, image analysis, and consistency</li> <li>Performance varying depending on the type and format of questions</li> <li>Providing potentially misleading and incomplete responses</li> </ul>	Awareness of ChatGPT limitations shapes educational practices, prompting adjustments to assessment tasks.
Stojanov (2023) ChatGPT-3.5 New Zealand	To report on experiences of using and limitations related to ChatGPT	An ethnographic study by the author where he learnt how to use ChatGPT, had conversation with it and reflected on his experiences	<ul style="list-style-type: none"> <li>responses with superficial and potentially contradictory information</li> </ul>	<ul style="list-style-type: none"> <li>providing good general knowledge of technical topics in a prompt and efficient manner</li> <li>-serving as a learning aid or "a more knowledgeable other" (p. 1)</li> </ul>
Thirunavukarasu et al. (2023) NA UK	To evaluate strength and weaknesses of ChatGPT in general practitioner setting	ChatGPT was asked to answer questions of Applied Knowledge Test (AKT) (i.e., a medical test)	<ul style="list-style-type: none"> <li>unable to achieve sufficient scores to pass the test (60.17% vs 70.42% required to pass)</li> <li>performance quality inconsistent with the difficulty levels of the test questions</li> </ul>	<ul style="list-style-type: none"> <li>achieving a level of proficiency comparable to that of a human expert</li> <li>could be used to automate tasks or as an assistant in clinical settings</li> </ul>
Wagner and Ertl-Wagner (2023) ChatGPT-3 Canada	To test ChatGPT's accuracy and reliability in answering radiologist questions	ChatGPT was asked to answer 88 questions and evaluated by radiologists including the authenticity of its answer	<ul style="list-style-type: none"> <li>providing inaccurate responses (or responses with errors): 33%</li> <li>hallucinating references: 63.8% self-created references</li> </ul>	N/A

Note. NA = Not applicable (i.e., the study did not provide sufficient information to determine the version of ChatGPT used).