Large Language Models in Education: A Systematic Review

Conference Paper · April 2024

DOI: 10.1109/CSTE62025.2024.00031

4 authors, including:

Bingyu Dong
Shaanxi Normal University
2 PUBLICATIONS 8 CITATIONS
SEE PROFILE

Jie Bai
Shaanxi Normal University
4 PUBLICATIONS 21 CITATIONS
SEE PROFILE

Large Language Models in Education: A Systematic Review

Bingyu Dong
Faculty of Education
Shaanxi Normal University
Xi'an, China
dongbingyu@snnu.edu.cn

Jie Bai
Faculty of Education
Shaanxi Normal University
Xi'an, China
baijie@snnu.edu.cn

Tao Xu
School of Software
Northwestern Polytechnical University
Xi'an, China
xutao@nwpu.edu.cn

Yun Zhou*
Faculty of Education
Shaanxi Normal University
Xi'an, China
zhouyun@snnu.edu.cn

2024 6th International Conference on Computer Science and Technologies in Education (CSTE) | 979-8-3503-5180-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CSTE62025.2024.00031

***Abstract*—Large Language Models (LLMs) refer to a type of generative artificial intelligence model that produces responses to natural language input. The purpose of this study is to analyze the current application status of LLMs in the field of education through a systematic review of the literature. Data were sourced from three databases: Web of Science, ERIC, and Google Scholar. The study includes 94 documents, analyzed from both qualitative and quantitative perspectives. The results show that large language models have great potential in the field of education, specifically in generating medical content, serving as an English learning assistant, assisting academic research, and evaluating the quality of tests, etc. However, there are still potential dangers such as hindering the development of critical thinking, creating academic integrity crises, and ethical and moral challenges. These findings showed the current application status of LLMs in education, laying the groundwork to inspire future research.**

***Keywords*—large language model, ChatGPT, artificial intelligence, education, systematic review**

I. Introduction

Large Language Models (LLMs) are a subset of generative Artificial Intelligence (AI) models trained on extensive text data and capable of generating human-like text content based on natural language input [1]. Existing studies indicate that generative AI tools can bring about shifts in educational methodologies [2] and provide educators with developmental opportunities [3]. Students can also use them to support learning activities [4], and this assistance has the potential to enhance learner engagement, satisfaction, and academic performance [5]. Furthermore, students gain opportunities to improve their language skills and foster collaboration [6]. However, the existing research also highlights the limitations of generative AI, including deficiencies in training, a lack of common sense, and challenges in reasoning [7]. The ethical challenges brought by AI are also undeniable [9]. Consequently, cautious use of these tools is recommended, given their potential for misinformation and an increase in inequalities [8]. To extensively explore the status of current applications of LLMs in education and provide insights for future research, this systematic review aims to address the following research questions: (1) What are the publication timeline, quantity, and prevalent terms in applied research on LLMs in education? (2) Which keywords show the highest co-occurrence? What are the primary themes of the investigation? (3) What are the domains, research methodologies, and models employed in the applications of LLMs in education? (4) What is the impact of LLM applications on learning? How are research ethics addressed?

II. Method

This study follows systematic retrieval rules and selects literature for review based on predetermined inclusion and exclusion criteria. The process of literature review was by the recommendations of the PRISMA statement [9] [10].

A. Search Strategy

Three databases were selected, including Web of Science, ERIC, and Google Scholar. The search terms are "large language models" and "education". The search time was from October 2017 to November 2023. The exclusion criteria are as follows: non-English language, unrelated to the theme of education, non-review articles, articles not from journals or conferences, publication dates outside the range of 2017-2023, non-developmental algorithm or technical articles, and articles not available in full text.

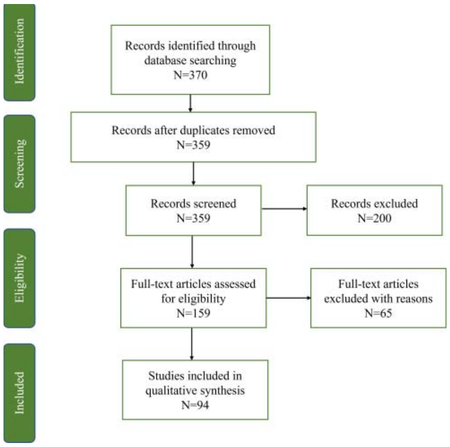
Identification
Records identified through database searching N=370
Screening
Records after duplicates removed N=359
Records screened N=359
Records excluded N=200
Eligibility
Full-text articles assessed for eligibility N=159
Full-text articles excluded with reasons N=65
Included
Studies included in qualitative synthesis N=94

Fig. 1. Flow chart of the study selection process.

B. Selection of Studies

370 documents were initially found in the three databases. After removing 11 duplicate records, 359 documents remained. Their qualifications to participate in this study were further

This work was supported by the Research Projects of the National Natural Science Foundation of China under Grant [62077036].

979-8-3503-5180-4/24/$31.00 ©2024 IEEE
DOI 10.1109/CSTE62025.2024.00031

evaluated through the title and abstract content. After two rounds of screening, a total of 94 documents were included. In the final analysis, the remaining 256 articles were excluded from this study because they did not meet the inclusion and exclusion criteria. Fig. 1 shows the selection process for systematic review projects.

C. Data Analysis

To address the research questions, a combination of quantitative and qualitative research methods was used. Quantitative analysis primarily involves the use of visualized charts and graphs, utilizing tools including Microsoft Excel and VOSviewer.

III. Results

A. Quantitative Analysis

Among the 94 publications in this study, only one was published in 2021, with the remaining articles all published in 2023. As observed in the generated word cloud (Fig. 2), terms such as "ChatGPT", "Education", "Medical", "Language", "Models", "Artificial", "Learning", and "Intelligence" appear with high frequencies. It is evident that the current research on LLMs predominantly focuses on the application of ChatGPT in education, particularly in language learning and medical education. Additionally, keywords such as opportunities and challenges are highlighted, indicating that the development of LLMs in education holds potential for the future.

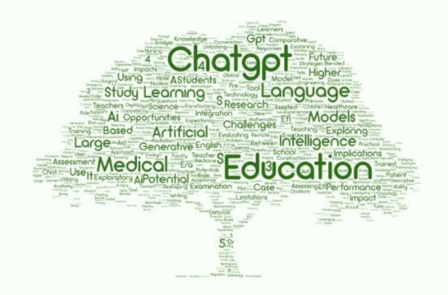

Fig. 2. Word cloud.

Co-authorship, co-occurrence, and citations are three different characteristics of literature review data analysis [11]. Using co-occurrence analysis of keywords under the condition of a minimum of one co-occurring word, a total of 68 keywords were obtained. The results are presented in Fig. 3. The size of the circles represents the frequency of keyword use, and the combination and grouping of circles represent the structure of the research field. Each group is presented in a different color, showing its importance and relevance. Research in this area is divided into four groups: red, yellow, blue, and green, with the red part showing the closest connection.

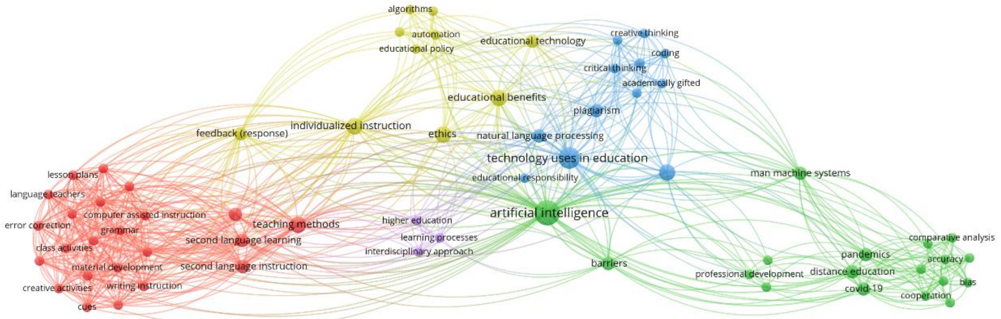

Fig. 3. Co-occurrence(keywords).

B. Content Analysis

More than 90% of large language models used in education use ChatGPT, with a focus on comparative studies between ChatGPT 3.5 and ChatGPT 4. These studies aim to assess the effectiveness of the models. The remaining studies predominantly involve models such as Google Bard, Bing, and other pre-trained models. The application of large language models in education is mainly distributed in the following fields: medical education, English language learning, academic research, and the evaluation of examinations. Additionally, there are studies in physics, mathematics, preschool education, tourism education, architectural biology, and other fields, though these are in a limited number.

1) Medical education

Over half of the articles focus on medical education. The primary research approaches include the following: (1) Dialogue-based evaluation of generation effectiveness: For example, Totelis et al. evaluated ChatGPT's ability to describe anatomical knowledge through conversations. Results indicated satisfactory descriptions but insufficient clinical significance [12]; (2) The investigation of language model abilities: For example, a study examined GPT-4's ability to provide medical advice using a questionnaire-based approach [13]. Other studies involved comparisons between model-generated content and expert opinions [14]. Furthermore, studies assessed how medical students and non-experts evaluate information from ChatGPT compared to standard surgical diagnostic resources [15]. In addition, there are still some studies that use the researchers' personal experiences and theoretical foundations to discuss the potential and challenges of applying the LLMs in education (e.g., [16] [17] [18]).

2) English language learning

Research on English language learning mainly consists of evaluating the potential of ChatGPT-generated content as an English learning aid [19]. Through qualitative research methods such as interviews and grounded theory, Bonner identified four roles that teachers play when using ChatGPT: conversationalist, content provider, assistant, and assessor [20]. Mohamed categorized teachers' perspectives into positive and negative aspects. While teachers acknowledge the tool's ability to furnish students with accurate information, they also express concerns that it might impede the development of students' critical thinking and research skills [21]. In addition, the feedback capabilities of teachers and ChatGPT in English learning are compared. The results indicate that the model provides a greater amount of feedback than the teacher, but there is a tendency for different types of feedback. This study emphasizes collaboration between teachers and LLMs [22].

3) Academic research

Concerning academic research, LLMs can serve as a tool for assisting research, compiling and summarizing information, and functioning as a research assistant [23], further breaking down language learning barriers, improving the paper writing experience, performing data analysis coding and interpretation, and performing complex image analysis [36]. Additionally, scholars have assessed the coherence, accuracy, and relevance of articles generated by ChatGPT [24] in academic writing. However, issues on research ethics, academic integrity, and copyright should be drawn attention and addressed [23] [27].

4) Exam evaluation

Another aspect of the investigation is to evaluate the performance of the models in various examinations. This includes comparing the accuracies of different models and comparing the results generated by models with those produced by real people. For example, the answers generated by ChatGPT-3.5, ChatGPT-4, Bing, and Google Bard, have been compared to explore the capabilities of several models (e.g., [30] [31]). The examination contents include the performance of medical school and law school examinations [25], the China National Medical Licensing Examination [26], written assignments [27], biological tests [28], and MCQ tests [29].

IV. Discussion

A. What Are the Publication Timeline, Quantity, and Prevalent Terms in Applied Research on LLMs in Education?

The main publication year for research on the application of LLMs in education is 2023. Nearly all studies utilizing the GPT series models employed versions 3.5 and above, and GPT-3.5 was released in November 2022. That is the reason the main publication year is predominantly 2023. Even though LLMs have been released within the past few years, a total of 93 documents have been retrieved, indicating that this research direction is quite promising. Through examining the word cloud, we found that prominent keywords include artificial intelligence, ChatGPT, medical, education, language, and LLMs. This observation indicates that the current research is concentrated on these topics.

B. Which Keywords Show the Highest Co-occurrence? What Are the Primary Themes of the Investigation?

Co-occurrence analysis for keywords indicates that terms such as artificial intelligence, teaching methods, second language learning, natural language processing, personalized learning, teaching feedback, educational benefits, and others are used more frequently. In the context of education, these models are primarily applied to second language learning. The aim is to enhance teaching methods, provide instructional feedback, facilitate personalized learning, and generate potential educational benefits.

C. What Are the Domains, RESEARCH Methodologies, and Models Employed in the Applications of LLMs in Education?

The primary fields of application for LLMs in education are concentrated in medical education, English language learning, academic research, and the evaluation of exams. Additionally, there are studies in physics, mathematics, preschool education, tourism education, architectural biology, and other fields, though these are in a limited number. The research methods employed in these studies are diverse, including surveys to explore user perceptions and acceptance of the technology; experimental approaches to investigate its effectiveness in teaching; dialogue-based assessments of its generative capabilities; and qualitative research methods such as interviews or theoretical interpretations to explore individual or collective perspectives and experiences with LLMs.

D. What Is the Impact of LLM Applications on Learning? How Are Research Ethics Addressed?

The existing studies show that LLM applications have a positive impact on performance. Specifically, LLMs can provide feedback to learners in English learning [22], serve as "research assistants" to aid researchers in their studies [23], and support students in active learning and skill development [26] [27]. However, there are concerns that they may hinder the development of students' critical thinking and research skills [21], potentially leading to academic integrity crises [27] [32] [33] and copyright issues [23]. Additionally, since models are trained, errors in information generation may occur [33] [34], and in some cases, they might impact students' moral values, posing challenges to ethics and morality [35] [36]. Therefore, the use of these tools should be cautious, and addressing the aforementioned issues should be a prerequisite for their use.

V. Conclusion

This study conducted a systematic literature review by analyzing the 94 retrieved articles. Results are presented through quantitative and qualitative analyses. The findings indicate that the application of LLMs in education is currently concentrated in specific areas, suggesting that it is still in the developmental stage. Although LLMs show potential in enhancing educational practices, there are still serious issues including academic integrity crises [27], misinformation [33], and the potential for misleading students' moral values [35]. Therefore, standards and ethics should be figured out before employing LLMs massively in education.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62077036 and 62377039, and in part by the National Key R&D Program of China(2022YFC3303600).

References

[1] L. Yan et al., “Practical and ethical challenges of large language models in education: A systematic scoping review,” Br. J. Educ. Technol., vol. n/a, no. n/a, doi: 10.1111/bjet.13370.

[2] F. M. D. Olite, I. del R. M. Suárez, and M. J. V. Ledo, “Chat GPT: origen, evolución, retos e impactos en la educación,” Educ. Médica Super., vol. 37, no. 2, Art. no. 2, May 2023, Accessed: Nov. 10, 2023. [Online]. Available: https://ems.sld.cu/index.php/ems/article/view/3876

[3] M. Halaweh, “ChatGPT in Education: Strategies for Responsible Implementation,” Contemp. Educ. Technol., vol. 15, no. 2, Jan. 2023, [Online]. Available: https://cue.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1385551&site=ehost-live

[4] M. Javaid, A. Haleem, R. P. Singh, S. Khan, and I. H. Khan, “Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system,” BenchCouncil Trans. Benchmarks Stand. Eval., vol. 3, no. 2, p. 100115, Jun. 2023, doi: 10.1016/j.tbench.2023.100115.

[5] M. Firat, “What ChatGPT means for universities: Perceptions of scholars and students,” J. Appl. Learn. Teach., vol. 6, no. 1, Art. no. 1, Apr. 2023, doi: 10.37074/jalt.2023.6.1.22.

[6] F. Fauzi, L. Tuhuteru, F. Sampe, A. M. A. Ausat, and H. R. Hatta, “Analysing the Role of ChatGPT in Improving Student Productivity in Higher Education,” J. Educ., vol. 5, no. 4, Art. no. 4, Apr. 2023, doi: 10.31004/joe.v5i4.2563.

[7] Md. M. Rahman and Y. Watanobe, “ChatGPT for Education and Research: Opportunities, Threats, and Strategies,” Appl. Sci.-BASEL, vol. 13, no. 9, May 2023, doi: 10.3390/app13095783.

[8] J. Qadir, “Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education.” TechRxiv, Dec. 30, 2022. doi: 10.36227/techrxiv.21789434.v1.

[9] M. Montenegro-Rueda, J. Fernández-Cerero, J. M. Fernández-Batanero, and E. López-Meneses, “Impact of the Implementation of ChatGPT in Education: A Systematic Review,” Computers, vol. 12, no. 8, Art. no. 8, Aug. 2023, doi: 10.3390/computers12080153.

[10] M. J. Page et al., “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” J. Clin. Epidemiol., vol. 134, pp. 178–189, Jun. 2021, doi: 10.1016/j.jclinepi.2021.03.001.

[11] M. Pradana, H. Elisa, and S. Syarifuddin, “Discussing ChatGPT in education: A literature review and bibliometric analysis,” Cogent Educ., vol. 10, Aug. 2023, doi: 10.1080/2331186X.2023.2243134.

[12] T. Totlis et al., “The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT,” Surg. Radiol. Anat., vol. 45, no. 10, pp. 1321–1329, Oct. 2023, doi: 10.1007/s00276-023-03229-1.

[13] K. Lower, I. Seth, B. Lim, and N. Seth, “ChatGPT-4: Transforming Medical Education and Addressing Clinical Exposure Challenges in the Post-pandemic Era,” INDIAN J. Orthop., vol. 57, no. 9, pp. 1527–1544, Sep. 2023, doi: 10.1007/s43465-023-00967-7.

[14] M. S. Lebhar, A. Velazquez, S. Goza, and I. C. Hoppe, “Dr. ChatGPT: Utilizing Artificial Intelligence in Surgical Education,” CLEFT PALATE CRANIOFACIAL J., Aug. 2023, doi: 10.1177/10556656231193966.

[15] T. Breeding et al., “The Utilization of ChatGPT in Reshaping Future Medical Education and Learning Perspectives: A Curse or a Blessing?,” Am. Surg., Jun. 2023, doi: 10.1177/00031348231180950.

[16] S. Ahn, “The impending impacts of large language models on medical education.,” Korean J. Med. Educ., vol. 35, no. 1, pp. 103–107, Feb. 2023, doi: 10.3946/kjme.2023.253.

[17] R. Tsang, “Practical Applications of ChatGPT in Undergraduate Medical Education,” J. Med. Educ. Curric. Dev., vol. 10, 2023, doi: 10.1177/23821205231178449.

[18] T. Jowsey, J. Stokes-Parish, R. Singleton, and M. Todorovic, “Medical education empowered by generative artificial intelligence large language models.,” Trends Mol. Med., Sep. 2023, doi: 10.1016/j.molmed.2023.08.012.

[19] J. C. Young and M. Shishido, “Investigating OpenAI’s ChatGPT Potentials in Generating Chatbot’s Dialogue for English as a Foreign Language Learning,” Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, pp. 65–72, Jun. 2023.

[20] E. Bonner, R. Lege, and E. Frazier, “Large Language Model-Based Artificial Intelligence in the Language Classroom: Practical Ideas for Teaching,” Teach. Engl. Technol., vol. 23, no. 1, pp. 23–41, Jan. 2023.

[21] A. M. Mohamed, “Exploring the potential of an AI-based Chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: perceptions of EFL Faculty Members,” Educ. Inf. Technol., Jun. 2023, doi: 10.1007/s10639-023-11917-z.

[22] K. Guo and D. Wang, “To resist it or to embrace it? Examining ChatGPT’s potential to support teacher feedback in EFL writing,” Educ. Inf. Technol., Aug. 2023, doi: 10.1007/s10639-023-12146-0.

[23] A. Pack and J. Maloney, “Using Generative Artificial Intelligence for Language Education Research: Insights from Using OpenAI’s ChatGPT,” TESOL Q., Aug. 2023, doi: 10.1002/tesq.3253.

[24] S. Kikalishvili, “Unlocking the potential of GPT-3 in education: opportunities, limitations, and recommendations for effective integration,” Interact. Learn. Environ., Jun. 2023, doi: 10.1080/10494820.2023.2220401.

[25] T. H. Kung et al., “Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models,” PLOS Digit. Health, vol. 2, no. 2, p. e0000198, Feb. 2023, doi: 10.1371/journal.pdig.0000198.

[26] H. Wang, W. Wu, Z. Dou, L. He, and L. Yang, “Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI,” Int. J. Med. Inf., vol. 177, Sep. 2023, doi: 10.1016/j.ijmedinf.2023.105173.

[27] G. M. Currie, “GPT-4 in Nuclear Medicine Education: Does It Outperform GPT-3.5?,” J. Nucl. Med. Technol., Oct. 2023, doi: 10.2967/jnmt.123.266485.

[28] A. Ignjatovic and L. Stevanovic, “Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education: a descriptive study.,” J. Educ. Eval. Health Prof., vol. 20, pp. 28–28, Oct. 2023, doi: 10.3352/jeehp.2023.20.28.

[29] K. E et al., “Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4.,” BMC Med. Educ., vol. 23, no. 1, pp. 772–772, Oct. 2023, doi: 10.1186/s12909-023-04752-w.

[30] T. Tülübas, M. Demirkol, T. Y. Ozdemir, H. Polat, T. Karakose, and R. Yirci, “An Interview with ChatGPT on Emergency Remote Teaching: A Comparative Analysis Based on Human-AI Collaboration,” Educ. Process Int. J., vol. 12, no. 2, pp. 93–110, Jan. 2023.

[31] J. Roos, A. Kasapovic, T. Jansen, and R. Kaczmarczyk, “Artificial Intelligence in Medical Education: Comparative Analysis of ChatGPT, Bing, and Medical Students in Germany.,” JMIR Med. Educ., vol. 9, pp. e46482–e46482, Sep. 2023, doi: 10.2196/46482.

[32] G. Currie and K. Barry, “ChatGPT in Nuclear Medicine Education.,” J. Nucl. Med. Technol., vol. 51, no. 3, pp. 247–254, Jul. 2023, doi: 10.2967/jnmt.123.265844.

[33] A. Abd-alrazaq et al., “Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions,” JMIR Med. Educ., vol. 9, p. e48291, Jun. 2023, doi: 10.2196/48291.

[34] H. B. Ilgaz and Z. Celik, “The Significance of Artificial Intelligence Platforms in Anatomy Education: An Experience With ChatGPT and Google Bard.,” Cureus, vol. 15, no. 9, pp. e45301–e45301, Sep. 2023, doi: 10.7759/cureus.45301.

[35] M. A. Peters et al., “AI and the future of humanity: ChatGPT-4, philosophy and education - Critical responses,” Educ. Philos. THEORY, May 2023, doi: 10.1080/00131857.2023.2213437.

[36] J G. Borger et al., “Artificial intelligence takes center stage: exploring the capabilities and implications of ChatGPT and other AI-assisted technologies in scientific research and education,” Immunol. CELL Biol., Sep. 2023, doi: 10.1111/imcb.12689.