Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28122

A Systematical Study of LLM-based Application on Education

Xinyu Wang

21ST Century School, Chengdu, China
18981924695@163.com

Abstract. In recent years, Large Language Models (LLMs) have attracted significant attention due to their strong capabilities in various aspects. They can not only deeply understand human language and generate fluent text but also possess a vast reserve of knowledge. With simple prompts, LLMs can generate customized and coherent responses. Such powerful capabilities have brought transformative impacts to various industries, particularly in the field of educational technology, where LLMs can significantly enhance teaching effectiveness. However, the quality of content generated by LLMs has not yet been systematically evaluated, and limitations such as bias and hallucinations have restricted their large-scale application in the education industry. Currently, efforts to utilize LLMs as auxiliary tools in the education sector are continuously advancing. Nevertheless, these studies primarily focus on the performance of LLMs in specific task scenarios and fail to comprehensively analyze the effectiveness of the content they generate across different educational scenarios, as well as the direct impact of LLMs on pedagogy. This leads to a lack of clarity and comprehensiveness in existing conclusions. Therefore, this paper systematically investigates the application of LLMs, with ChatGPT as an example, in education. Firstly, the paper briefly introduces the development background of LLMs, ChatGPT, and pedagogy. Secondly, it provides a detailed analysis and summary of existing cutting-edge research. Thirdly, it conducts an in-depth analysis of the research content of each study from the perspectives of research topics and research methods. Finally, it clarifies the advantages and disadvantages of LLMs and prospects future research directions. The paper points out that, in general, LLMs are beneficial to the development of pedagogy, but improvements are still needed in aspects such as ethical issues and the quality of generated content.

Keywords: LLM, education development, ChatGPT, artificial intelligence, AI agent

1. Introduction

In recent years, Large Language Models (LLMs) represented by GPT and PaLM have developed rapidly and attracted widespread attention from all sectors [1]. Trained on massive amounts of text data, these models can not only deeply understand the semantics of human language and generate fluent text but also possess nearly encyclopedic knowledge reserves. A key feature of LLMs is in-context learning [1], which enables them to generate coherent and high-quality text based on input context and existing information. Thus, they are highly suitable for practical application scenarios

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28122

that require interaction and communication. In addition, LLMs extensively adopt Reinforcement Learning from Human Feedback (RLHF) [1]. Through human feedback on generated results, the model continuously adjusts its behavior to correct errors, optimize outputs, and thereby continuously improve its performance. At present, LLMs have been widely applied in numerous industries such as medicine, agriculture, education, finance, and engineering, supporting various types of generation tasks, including but not limited to multi-turn dialogue, feedback provision, data integration, risk assessment, and text summarization [2]. Specifically, users only need to provide simple instructions or prompts, and LLMs can generate coherent and customized text content. Such powerful capabilities have brought transformative impacts to many industries, especially in the field of educational technology. They can effectively assist educators in their professional development, reduce the burden by replacing repetitive tasks such as homework grading, and provide students with highly personalized learning experiences and real-time feedback, thereby significantly enhancing teaching effectiveness [2].

However, although LLMs perform well in content generation, the quality, reliability, and educational applicability of the content they generate have not been fully evaluated and verified. During the training process, models may absorb social biases and factual errors contained in the data, leading to the generation of content that is no longer objective and rational, or the occurrence of the "hallucination" phenomenon where the generated content seems reasonable but is actually incorrect [2]. This poses significant risks in the practical application of LLMs in the education field. For example, they may provide incorrect reasoning or calculation steps when solving mathematical problems, present fictional factual details or biases when explaining historical events, or generate content lacking rigorous logic and evidence when answering open-ended questions. These situations may cause LLMs to convey incorrect ideas to students, promote the formation of incorrect knowledge and ways of thinking, seriously harm students' growth, and thereby greatly restrict the application and promotion of LLMs in the education industry.

Currently, auxiliary studies specifically focusing on the application of LLMs in the education field are gradually emerging, which deeply explore their integration into pedagogy. For instance, Neumann et al. [3] investigated the effectiveness of a language model-driven chatbot (MoodleBot) as an auxiliary tool for learners in academic courses. These studies mainly focus on the performance of LLMs in specific tasks within educational work but fail to consider the comprehensive performance of the content they generate across different scenarios and groups. Moreover, there is a lack of systematic and multi-angle analysis.

Therefore, taking ChatGPT as an example, this paper conducts a systematic investigation and analysis of the performance of LLMs in pedagogical tasks. Firstly, the paper expounds on the core definitions and development history of LLMs and ChatGPT, and sorts out the relevant context of the integration of education and technology. Then, it screens, classifies, and summarizes existing studies on intelligent pedagogy. Next, from the two dimensions of pedagogical research topics (including empowerment effects, adaptation strategies, capability evaluation, and acceptance attitudes) and pedagogical research methods (including survey research and literature review), it conducts an in-depth analysis of the content, results, advantages, and disadvantages of each cutting-edge study. Finally, based on the comprehensive evaluation and analysis of the investigated studies, it reveals the advantages and limitations of LLMs in pedagogical applications and points out the development directions of future research.

Tests show that LLM systems represented by ChatGPT are generally effective in auxiliary pedagogical tasks. They have unique advantages in enhancing learning experiences, improving

DOI: 10.54254/2755-2721/2025.LD28122

teaching efficiency, and promoting interdisciplinary integration. However, improvements are still needed in aspects such as content bias, content credibility, and evidence of effectiveness.

2. Background

2.1. LLM and ChatGPT

Large Language Models (LLMs) are advanced computational models trained on massive amounts of text data and featuring a large-scale parameter size. The core capability of LLMs is to understand human language and generate text. Their core principle is to predict the next most likely word based on existing text (prompts, keywords) and add each predicted word to the input, ultimately generating coherent and high-quality text [1]. The core technical architecture of LLMs is the Transformer architecture, whose self-attention mechanism helps the model focus on keywords related to the current word (regardless of the distance), thereby handling long-range dependencies [1]. The current core application directions of LLMs include fields such as professional decision support (e.g., medical diagnosis), public service optimization (e.g., education), and creative production (e.g., media and entertainment industry) [2]. ChatGPT is a LLM chatbot developed by the artificial intelligence research company OpenAI, representing a revolutionary technology. ChatGPT has several key advantages: strong natural language generation capabilities, enabling it to generate highly coherent and human-like responses; scalability, allowing it to handle a large number of dialogues simultaneously and generate answers quickly; and customizability, enabling it to provide personalized user experiences. Therefore, ChatGPT is currently widely applied in various fields such as customer service and content creation. However, it also has many unresolved shortcomings, such as potential biases in answers, weak emotional intelligence, limited knowledge base, and lack of empathy [4].

2.2. Research development of education

The main research scope of education includes educational theories, educational practices, educational policies and systems, and cutting-edge educational technologies. Since the 20th century, the concept of "learner-centered" has gradually become a widespread consensus. Personalized learning that focuses on individual differences and needs has become an important goal, and the in-depth integration of technology and education has become a current hotspot. In the early stage (from the late 1950s to the 1960s), computer-assisted instruction gradually emerged, enabling students to actively participate in the learning process and learn at their own pace [5]. Subsequently (in the early 21st century), the rapid popularization of the Internet provided students with a large amount of online learning materials, making it an efficient learning tool [6]. In recent years, artificial intelligence (AI) technology has developed rapidly, becoming a transformative force that reshapes learning experiences and greatly promotes the development of the education field [7]. Despite continuous technological progress, challenges still exist in achieving in-depth personalized learning and providing learning support with natural interaction. The emergence of LLMs has brought numerous opportunities. As a major breakthrough in the field of natural language generation, they have initiated the exploration of intelligent education applications and attracted significant attention [3].

However, on the one hand, existing studies fail to consider the direct impact of LLMs on the development of pedagogy; on the other hand, they do not take into account the effectiveness of AI-assisted education from different topics and perspectives, resulting in one-sided and ambiguous

DOI: 10.54254/2755-2721/2025.LD28122

research results. To fill the relevant research gaps, this paper systematically and multi-anglely studies the application of cutting-edge generative LLM technologies in pedagogical research.

3. Study

3.1. Overview

To systematically analyze and compare the investigated cutting-edge studies, this section analyzes, classifies, and summarizes the selected literature, and summarizes them from the two core dimensions of research methods and research topics to form Table 1. Table 1 includes the specific names of these studies, the research topics, the research methods used, and the main research conclusions. This table aims to provide readers with a clear content summary and serve as the basis for the subsequent in-depth analysis of research topics (3.2) and research methods (3.3).

Table 1. Overview of research methods and topics in cutting-edge literature

Study Name	Category (Method/Topic)	Topic	Method	Result
LLM Agents for Classroom Sim [8]	Survey Research/ Empowerment Effect	(1) Simulation Performance (2) Learning Experience (3) Group Behavior	(1) Multi-agent classroom simulation framework (2) Design novel class roles and a control mechanism	(1) Better learning outcomes (2) More engagement
LLM Chatbot in Higher Education [3]	Survey Research/ Empowerment Effect	(1) Students' acceptance (2) Accuracy of responses (3) Congruence with established courses	(1) RAG-based MoodleBot (2) TAM evaluates user acceptance	LLM-based chatbots can significantly improve the teaching and learning
LLM Agents for Education: Advances and Applications [9]	Literature Review/Capability Evaluation	Overview of LLM agents for education	Task-centric LLM agents for educational contexts	LLM agents have the potential to revolutionize education.
Evaluating Large Language Models: A Comprehensive Survey [10]	Literature Review/Capability Evaluation	Offer a panoramic perspective on the evaluation of LLMs	(1) Integrate insights (2) Expand the scope	LLMs has been astonishing across numerous tasks
Don't Make Your LLM an Evaluation Benchmark Cheater [11]	Survey Research/ Capability Evaluation	Inappropriately using benchmarks and misleadingly interpreting the results	Simulate three extreme leakage issues	Data leakage can largely boost the benchmark results of LLMs
GenAI in Education: Opps, Challenges, Strategies [12]	Survey Research/ Empowerment Effect	(1) The opps and challenges of GenAI's in education (2) The strategies to promote GenAI	Recruit educators to participate in a GenAI training and survey their views online	A positive awareness of opportunities for integrating GenAI
On the Limits of Artificial Intelligence (AI) in Education [13]	Literature Review/Empowerment Effect	(1) Modeling Limits (2) Social Harms (3) Educational Distortion (4) Environmental Costs	Outline a number critical issues and concerns by category	Case for recalibrating current discussions around AI and education.

DOI: 10.54254/2755-2721/2025.LD28122

Table 1. (continued)

A Comprehensive Review on Generative AI for Education [14]	Literature Review/Empowerment Effect	Benefits and challenges of GAI	(1) Scope definition (2) Databases selection and search string design	Implementation of GAI will bring revolution in the field of education.
Unveiling the shadows: Beyond the hype of AI in education [15]	Survey Research/ Empowerment Effect	Potential negative implications of integrating AI in education	(1) Development and validation of a theoretical model (2) Sampling and participation	AI can indeed personalize learning experiences, but it also raises concerns.
ChatGPT Impact in Education [16]	Literature Review/Empowerment Effect	Benefits and challenges of ChatGPT	Kitchenham and Charters' Systematic Literature Review (SLR) approach	Potential benefits and challenges are coexisting of ChatGPT.
GenAI in Education: Perspectives [17]	Literature Review/Empowerment Effect	Benefits and challenges of GenAI	Offer a comprehensive exploration of GenAI from both theoretical and practical perspectives	GenAI tools can significantly improve educational outcomes
Korean Teachers' Views on AI Ed & Training [18]	Survey Research/ Acceptance Attitude	Teachers' perceptions (AI education and AI teacher training programs)	(1) Online survey (2) Quantitative and qualitative data analysis	Teachers hold favorable attitudes toward AI education.
Attitudes towards AI: measurement and associations with personality [19]	Survey Research/ Acceptance Attitude	(1) Attitudes towards AI (2) Associations between AI-related attitudes and personality traits.	Ethics statement, Participants and procedure, online research	(1) Slightly positive attitudes. (2) Significant connections between attitudes and two personality traits.
Responsible HCAI in Education: Review [20]	Literature Review/Adaptation Strategy	(1) Key characteristics of responsible human-centered AI (2) Key stakeholders and their roles	Convergent Qualitative meta-integration approach	Unpack key characteristics of responsible AI and identified essential stakeholders and their responsibilities.
Responsible GenAI for Education [21]	Survey Research/ Adaptation Strategy	Optimising gen AI for educational use cases	Combining engagement through interviews and workshops with a literature review	LearnLM-Tutor outperformed Gemini 1.0.
LLMs in Medical Education Review [22]	Literature Review/Empowerment Effect	LLM applications and impacts in medical education	Systematic search in PubMed, WOS and Embase for articles using LLM keywords	Responsible implementation of LLMs in medical education enhances learning experiences.
LLMs & Game-Based Learning in Education [23]	Literature Review/Adaptation Strategy	(1) Transformative potential and associated challenges of large language models (2) Playful solutions	Theoretical deduction and framework development	Well-designed games indeed address the challenges and promote players holistically.
A Survey on Evaluation of Large Language Models [24]	Literature Review/Capability Evaluation	A comprehensive review of evaluation methods for LLMs	Propose a three-dimensional analysis framework:where, what and how to evaluate	Current LLMs exhibit certain limitations in numerous tasks.

DOI: 10.54254/2755-2721/2025.LD28122

Table 1. (continued)

Student LLM Use in Engineering Education [25]	Survey Research/ Empowerment Effect	(1) LLM-Assisted Essay Quality (2) Detection System Efficacy (3) Students' perceptions	Deductive and inductive approach,combining conceptualization and empirical analysis	(1) Good essays output (2) Detection failure (3) Evolving perspectives
AI-Enhanced Learning Model (Self-Efficacy, Motivation) [26]	Survey Research/ Empowerment Effect	Whether AI capabilities can foster critical thinking awareness by enhancing general self-efficacy and learning motivation.	(1) Formulation of hypothetical model (2) Structural equation modeling (3) Questionnaires (4) Statistical analysis	AI capabilities could indirectly enhance students' critical thinking awareness (not significant).

3.2. LLM-based education research topics

This section analyzes cutting-edge studies from the perspective of task topics, which can be divided into the following four categories: LLM enhancement effects, LLM adaptation strategies, LLM-oriented evaluation, and LLM acceptance attitudes. Each category is summarized as follows:

(1) LLM Enhancement Effects: Studies in this category focus on how LLMs integrate into the education system and deeply explore the transformative impacts of the application and integration of LLMs on learning effects, educators and learners, and teaching processes. A large number of studies have shown that LLMs can significantly improve the teaching and learning processes and final learning outcomes through personalized learning experiences, increased interaction, and other means [3, 8, 12, 14, 16, 17, 22, 25, 26]. A representative study is Zhang et al. [8], which proposed SimClass (a multi-agent classroom simulation teaching framework), introduced real classrooms, invited students to participate, and concluded through feedback from post-class tests that it helps improve students' learning effects (increasing interaction). In particular, Al-Zahrani [15] conducted an in-depth analysis of the potential risks and negative impacts of AI, pointing out that although AI can provide beneficial personalized education, it is accompanied by hidden dangers such as data privacy and security, algorithmic bias, and ethical issues, which require careful handling and the adoption of comprehensive strategies. Studies in this category emphasize the dual nature of LLMs: they can optimize education while requiring measures to address risks.

(2) LLM Adaptation Strategies: This topic focuses on specific methods for integrating LLMs into education, aiming to maximize the utilization of their advantages through the responsible deployment of LLMs, thereby effectively improving educational outcomes. Fu & Weng [20] selected important literature, summarized the key characteristics of responsible artificial intelligence, and revealed the key stakeholders and their responsibilities in the educational environment. Jurenka et al. [21] proposed a comprehensive evaluation protocol, introduced a new generative AI tutoring tool based on Gemini 1.0, and then evaluated its capabilities (outperforming Gemini 1.0 in most measured teaching dimensions). In particular, Huber et al. [23] proposed a new conceptual framework for gamified learning, believing that game-based educational methods are more exploratory and interesting, can fully exploit the potential of LLMs, and listed various entertainment-based educational approaches to demonstrate their effectiveness.

(3) LLM-oriented Capability Evaluation: This topic focuses on the evaluation of various capabilities of LLMs and the assessment of potential hidden dangers to quantify model effects. For example, Guo et al. [10] systematically expounded on the core capabilities of LLMs (covering key aspects such as knowledge and reasoning). In addition, to ensure that LLMs are safe, reliable, and in

DOI: 10.54254/2755-2721/2025.LD28122

line with basic ethical norms, the article also deeply discussed consistency evaluation and safety evaluation, including but not limited to ethical issues, biases, toxicity, and authenticity. In particular, Zhou et al. [11] conducted an empirical study to explore the impact of benchmark data leakage on LLM evaluation, pointing out that it will make test results unreliable and unfair, and calling for the strict avoidance of such behaviors. These studies reveal the complexity of LLM evaluation work, which requires a more rigorous testing framework.

(4) LLM Acceptance Attitudes: This category explores the views and acceptance of AI among educators and students. Lee [18] obtained teachers' opinions on conducting AI education and AI teacher training programs in the educational environment through online questionnaires and interviews. The results showed that teachers hold a positive attitude towards AI education for teaching and future applications. Stein [19] launched a novel psychology-based questionnaire to obtain people's overall attitudes towards AI (not affected by specific scenarios). The article found that groups with high agreeableness and younger groups are more likely to accept AI technology.

3.3. LLM-based education research methods

From the perspective of methods, the cutting-edge studies can be divided into two categories: survey research and literature review. Each method category explains its application, advantages, and the studies it covers:

(1) Survey Research: This method adopts empirical approaches such as experiments, questionnaires, and interviews to collect first-hand evidence for hypothesis verification. Its advantage lies in providing real feedback, but it may be affected by factors such as sample bias. The covered studies include: Zhang et al. [8] invited students to participate in LLM-assisted simulated classrooms and quantified learning outcomes through post-class tests, finding that LLMs increased classroom interaction while improving learning effects; Neumann et al. [3] developed a chatbot MoodleBot based on the RAG architecture and evaluated user acceptance using the Technology Acceptance Model (TAM) questionnaire, with results showing that it was recognized by students as an auxiliary tool; Zhou et al. [11]: conducted an empirical study to test the impact of benchmark leakage on LLM evaluation, simulating three extreme leakage scenarios (test prompts, test sets, and other relevant data), and found that data leakage would make the test results of LLMs unreliable; Ng et al. [12] and Lee [18] conducted questionnaires and interviews with teachers respectively, finding that they held a positive attitude towards the application of AI in education; Al-Zahrani [15] proposed a theoretical model of AI's negative impacts, and then verified its reliability through online questionnaires, pointing out that hidden dangers such as privacy, bias, and transparency are interrelated; Stein [19] launched an online questionnaire designed based on psychology to obtain people's attitudes towards AI, and found that groups with high agreeableness and younger groups held a more inclusive attitude towards AI; Jurenka et al. [21] proposed LLM teaching capability standards through interviews with educators and learners, and the LearnLM-Tutor developed based on Gemini 1.0 performed better in most dimensions; Bernabei et al. [25] asked students to use LLMs to assist in generating personal essays and combined with questionnaires, finding that the text quality was high but LLMs could not identify whether the essays were generated by AI, and students' attitudes towards LLMs changed over time; Jia et al. [26] tested the impact of AI on self-efficacy, learning motivation, and critical thinking through questionnaires, and found that although there was an improvement, the effect was not significant.

(2) Literature Review: This method focuses on systematically reviewing, summarizing, and concluding existing research works, sorting out AI- or LLM-based applications related to pedagogy, and providing route support for subsequent studies. Its advantage lies in covering a relatively

DOI: 10.54254/2755-2721/2025.LD28122

comprehensive range of cutting-edge works and integrating their intelligent characteristics. The main studies included in this section are:

Chu et al. [9] proposed a task-centric taxonomy, classified numerous LLM educational agents, reviewed their applications in various educational fields, and concluded that they have the potential to revolutionize personalized learning, intelligent tutoring, and teaching automation; Guo et al. [10] used a taxonomy to divide LLM evaluation into three categories, systematically expounded on the various capabilities, safety hazards, and potential applications of LLMs, and finally proposed benchmark evaluation to help relevant personnel from all walks of life understand and evaluate the performance of large language models; Selwyn [13] carefully summarized the limitations of AI and its potential negative impacts on various aspects of society through classification and analysis, and called for recalibrating the current discussions around artificial intelligence and education; Mittal et al. [14] selected literature to discuss the pioneering impacts of generative AI on education in terms of competency development and personalized interaction, as well as risks such as copyright and controllability; Bettayeb Anissa et al. [16] used the Systematic Literature Review (SLR) method to explore four core research questions: the benefits and challenges of ChatGPT, its impact on student engagement and learning outcomes, ethical considerations and safeguards, and its impact on educators and teachers; Noroozi et al. [17] reviewed existing studies, discussed various applications, impacts, and challenges of general AI tools including ChatGPT in the education field, and believed that generative AI tools can significantly improve educational outcomes; Fu & Weng [20] adopted an advanced convergent qualitative meta-integration method to integrate existing data, summarized the key characteristics of responsible artificial intelligence and the responsibilities of various stakeholders in the education field; Lucas et al. [22] systematically searched for literature discussing the application of LLMs in the medical field using LLMs and medical education as keywords, pointing out that language models have great potential to change the medical education model, but at the same time, potential negative impacts need to be carefully addressed; Huber et al. [23] pointed out the problem faced by education: it is necessary to seize new opportunities while preventing the neglect of cultivating required skills. Therefore, a gamified education framework was proposed to maximize the use of LLM capabilities; Chang et al. [24] conducted a comprehensive review of LLM evaluation methods using a comparative approach, focusing on three key aspects: what to evaluate, where to evaluate, and how to evaluate, and pointed out the existing defects of LLMs, such as reasoning and robustness tasks.

4. Discussion

4.1. Advantages

Based on a systematic investigation of current cutting-edge literature, we found that LLMs exhibit significant advantages in multiple aspects in their application to pedagogy. Firstly, in terms of learning experience, the customizability of LLMs enables them to generate personalized teaching content according to users' needs and characteristics, and flexibly adjust teaching objectives to meet user needs, greatly improving users' learning experience. For example, Zhang et al. [8] and Neumann et al. [3] constructed LLM-based educational frameworks and dialogue systems to provide students with personalized learning services, significantly enhancing students' engagement and mastery of knowledge. Secondly, in terms of teaching efficiency, LLMs can assist teachers in completing low-complexity and time-consuming tasks such as creating presentation slides, designing lesson plans, and providing student feedback, thereby greatly saving teachers' time. As shown in Ng et al. [12] and Noroozi et al. [17], LLMs can serve as efficient teaching auxiliary tools

DOI: 10.54254/2755-2721/2025.LD28122

to replace tasks such as providing student feedback and grading assignments. Thirdly, LLMs have demonstrated potential for interdisciplinary integration. For instance, in medical education [22] and engineering education [25], LLMs, as learning auxiliary tools, can not only provide support for professional knowledge but also focus on practical scenarios, enhancing students' practical abilities. In particular, the psychology-based questionnaire launched by Stein [19] revealed the relationship between the degree of acceptance of LLMs among different groups and their personality traits, providing an important breakthrough direction for the large-scale application of LLMs in the education industry in the future. Overall, existing studies fully demonstrate the great potential of LLMs in educational assistance, content generation, personalized learning, and other aspects, laying a solid foundation for their further integration into the education system.

4.2. Limitation

Although LLMs show many advantages in pedagogy, they still have a series of significant limitations, which include both common technical issues and specific challenges in educational scenarios. Firstly, in terms of ethics and morality, most studies believe that LLMs have biases, which may lead them to generate content with biases in aspects such as gender, race, and culture, thereby promoting the spread of incorrect perceptions. Risks such as data privacy leakage, scientific integrity issues, and copyright disputes may also pose threats to the safety of society and individuals [10, 14-17, 22]. Secondly, in terms of content credibility, Mittal et al. [14] indicated that LLMs have obvious hallucination problems (Pseudo Imagination). To pursue language fluency, they generate seemingly reasonable but actually false or misleading information without factual basis, which greatly reduces the credibility of the content they generate. Additionally, the relatively weak inductive reasoning ability of LLMs pointed out by Guo et al. [10] and Chang et al. [24], as well as the lack of critical thinking in LLMs proposed by Bettayeb Anissa et al. [16], also reduce the accuracy of their content. Thirdly, the evidence for their effectiveness is still insufficient. Although multiple studies have demonstrated the positive potential of LLMs, most empirical studies are limited to short-term cases with very small sample sizes. For example, the user group in Zhang et al. [8] was only university students with a small number of courses studied, and the research object of Ng et al. [12] focused on educators in Canada with the experiment still in the preliminary stage. This leads to a lack of large-scale and rigorous evidence regarding whether LLMs can improve students' ultimate learning outcomes in a long-term and stable manner, which greatly hinders the confidence of educators and policymakers in applying them on a large scale. In particular, Zhou et al. [11] explained that benchmark data leakage can cause a great deviation between the evaluation results of LLMs and the actual situation, thereby making the test data of LLMs no longer credible. These studies indicate that LLMs still face technical, methodological, and ethical bottlenecks in their application in education.

4.3. Future work

Based on the analysis of the advantages and limitations of existing studies, this paper puts forward the following suggestions for the development of future research directions to promote the deeper and more responsible application of LLMs in pedagogy. Firstly, a standardized and multi-dimensional evaluation framework should be developed to conduct more comprehensive and systematic testing on the capabilities, impacts, risks, and other aspects of LLMs, avoiding data leakage and evaluation bias, so as to ensure the credibility of evaluation results and the accuracy and educational value of the content generated by LLMs. Secondly, focus should be placed on the

DOI: 10.54254/2755-2721/2025.LD28122

technical improvement of LLMs, including capabilities such as emotional intelligence, inductive reasoning, and multi-modal input (e.g., images, language, videos). This will enhance their in-depth understanding of educational contexts and the quality of interaction in complex practical scenarios, while paying attention to learners' emotional states and cultural backgrounds to improve the sense of immersion and inclusiveness in learning. Thirdly, it is crucial to strengthen research on the ethics and fairness of LLMs in education. The boundaries of educational use of LLMs and the attribution of responsibilities should be clarified, and the implementation of the "responsible AI" framework should be promoted, including data de-biasing, privacy protection, and transparency improvement. This ensures that the application of LLMs in education always focuses on promoting the development of education and improving teaching efficiency and quality, rather than posing threats to the safety of individuals and society. In addition, conducting more long-term and large-scale empirical effect studies is also an important direction. Since most current studies are short-term, small-scale, or exploratory in laboratory environments, future research urgently needs to conduct continuous observation and data collection on large-scale samples (e.g., multiple institutions in different regions with different academic backgrounds) over a period spanning several semesters or even academic years. Only through long-term and rigorous empirical evidence can we clarify the potential value and risks of LLMs in educational applications. Finally, AI training for educators should be strengthened. While helping teachers acquire the skills to proficiently use LLMs to assist in generating teaching content and providing teaching feedback, their understanding of LLMs should be improved, thereby enhancing their confidence in using LLMs in educational work and laying a foundation for the integration of education and LLM technology. Through these efforts, LLMs are expected to become an indispensable intelligent auxiliary tool in the education system in the future, truly realizing the vision of "learner-centered" personalized education.

5. Conclusion

This paper systematically reviews the multi-dimensional impacts of LLMs in educational applications. Through the analysis and evaluation of multiple topics such as enhancement effects, adaptation strategies, capability evaluation, and acceptance attitudes, as well as cross-inference using multiple methods including survey research and literature review, it fills the research gaps in "the direct impact of LLMs on the development of education" and "multi-angle validity verification", providing a systematic framework for subsequent studies. Through a comprehensive analysis of cutting-edge literature, the research confirms that LLMs show significant potential in reshaping educational paradigms, optimizing learning experiences, and improving teaching efficiency. At the same time, it also reveals the limitations of LLMs in aspects such as ethics and morality, data bias, and copyright disputes. Based on the above advantages and disadvantages, this study proposes suggestions for future development directions, such as the development of a standardized and multi-dimensional evaluation framework and research on educational ethics.

References

[1] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models [J]. ACM transactions on intelligent systems and technology, 2024, 15(3): 1-45.
[2] Hadi M U, Qureshi R, Shah A, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects [J]. Authorea preprints, 2023, 1(3): 1-26.
[3] Neumann A T, Yin Y, Sowe S, et al. An llm-driven chatbot in higher education for databases and information systems [J]. IEEE Transactions on Education, 2024.

DOI: 10.54254/2755-2721/2025.LD28122

[4] Kalla D, Smith N, Samaah F, et al. Study and analysis of chat GPT and its impact on different fields of study [J]. International journal of innovative science and research technology, 2023, 8(3).
[5] Chambers J A, Sprecher J W. Computer assisted instruction: Current trends and critical issues [J]. Communications of the ACM, 1980, 23(6): 332-342.
[6] Lenhart A, Simon M, Graziano M. The Internet and Education: Findings of the Pew Internet & American Life Project [J]. 2001.
[7] Ayeni O O, Al Hamad N M, Chisom O N, et al. AI in education: A review of personalized learning and educational technology [J]. GSC Advanced Research and Reviews, 2024, 18(2): 261-271.
[8] Zhang Z, Zhang-Li D, Yu J, et al. Simulating classroom education with llm-empowered agents [J]. arXiv preprint arXiv: 2406.19226, 2024.
[9] Chu Z, Wang S, Xie J, et al. Llm agents for education: Advances and applications [J]. arXiv preprint arXiv: 2503.11733, 2025.
[10] Guo Z, Jin R, Liu C, et al. Evaluating large language models: A comprehensive survey [J]. arXiv preprint arXiv: 2310.19736, 2023.
[11] Zhou K, Zhu Y, Chen Z, et al. Don't make your llm an evaluation benchmark cheater [J]. arXiv preprint arXiv: 2311.01964, 2023.
[12] Ng D T K, Chan E K C, Lo C K. Opportunities, challenges and school strategies for integrating generative AI in education [J]. Computers and Education: Artificial Intelligence, 2025: 100373.
[13] Selwyn N. On the limits of artificial intelligence (AI) in education [J]. Nordisk tidsskrift for pedagogikk og kritikk, 2024, 10(1): 3-14.
[14] Mittal U, Sai S, Chamola V. A comprehensive review on generative AI for education [J]. IEEE Access, 2024.
[15] Al-Zahrani A M. Unveiling the shadows: Beyond the hype of AI in education [J]. Heliyon, 2024, 10(9).
[16] Bettayeb A M, Abu Talib M, Sobhe Altayasinah A Z, et al. Exploring the impact of ChatGPT: conversational AI in education [C]//Frontiers in Education. Frontiers Media SA, 2024, 9: 1379796.
[17] Noroozi O, Soleimani S, Farrokhnia M, et al. Generative AI in Education: Pedagogical, Theoretical, and Methodological Perspectives [J]. International Journal of Technology in Education, 2024, 7(3): 373-385.
[18] Lee Y J, Davis R O, Ryu J. Korean in-service teachers' perceptions of implementing artificial intelligence (AI) education for teaching in schools and their AI teacher training programs [J]. International Journal of Information and Education Technology, 2024, 14(2): 214-219.
[19] Stein J P, Messingschlager T, Gnambs T, et al. Attitudes towards AI: measurement and associations with personality [J]. Scientific Reports, 2024, 14(1): 2909.
[20] Fu Y, Weng Z. Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices [J]. Computers and Education: Artificial Intelligence, 2024, 7: 100306.
[21] Jurenka I, Kunesch M, McKee K R, et al. Towards responsible development of generative AI for education: An evaluation-driven approach [J]. arXiv preprint arXiv: 2407.12687, 2024.
[22] Lucas H C, Upperman J S, Robinson J R. A systematic review of large language models and their implications in medical education [J]. Medical education, 2024, 58(11): 1276-1285.
[23] Huber S E, Kiili K, Nebel S, et al. Leveraging the potential of large language models in education through playful and game-based learning [J]. Educational Psychology Review, 2024, 36(1): 25.
[24] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models [J]. ACM transactions on intelligent systems and technology, 2024, 15(3): 1-45.
[25] Bernabei M, Colabianchi S, Falegnami A, et al. Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances [J]. Computers and Education: Artificial Intelligence, 2023, 5: 100172.
[26] Jia X H, Tu J C. Towards a new conceptual model of AI-enhanced learning for college students: The roles of artificial intelligence capabilities, general self-efficacy, learning motivation, and critical thinking awareness [J]. Systems, 2024, 12(3): 74.