Education and Information Technologies
https://doi.org/10.1007/s10639-024-12898-3

SI: GENERATIVE ARTIFICIAL INTELLIGENCE

Impact of assignment completion assisted by Large Language Model-based chatbot on middle school students' learning

Yumeng Zhu[1] · Caifeng Zhu[2] · Tao Wu[3] · Shulei Wang[4] · Yiyun Zhou[4] · Jingyuan Chen[1] · Fei Wu[3] · Yan Li[1]

Received: 3 December 2023 / Accepted: 3 July 2024

Abstract
With the prevalence of Large Language Model-based chatbots, middle school students are increasingly likely to engage with these tools to complete their assignments, raising concerns about its potential to harm students' learning motivation and learning outcomes. However, we know little about its real impact. Through quasi-experiment research with 127 Chinese middle school students, we examined the impact of completing assignments with a Large Language Model-based chatbot, wisdomBot, on middle school students' assignment performance, learning outcomes, learning motivation, learning satisfaction, and learning experiences; we also summarized teachers' reflections on learning design. Compared to control groups, the Large Language Model chatbot-assisted group demonstrated significantly higher assignment submission rates, word counts, and scores in assignment performance. However, they gained significantly lower scores on materials refinement and knowledge tests. No significant differences have been observed in learning motivation, satisfaction, enjoyment, and students' ability to migrate their knowledge. The majority of students expressed satisfaction and willingness to continue using the tool. We also identified three key gaps in learning designs, including providing scaffolds for the potential prompts, suggesting group collaboration mode, and relinquishing the authoritarian of the teacher. Our findings provide insights regarding with Large Language Model-based chatbots we could better design assignment assessment tools, facilitate students' autonomous learning, provide emotional support, propose guidelines and instructions about applying Large Language Model-based chatbots in K-12, as well as design specialized educational Large Language Model-based chatbots.

Keywords Large language model · Chatbot · K-12 · Middle school · Assessment

Extended author information available on the last page of the article

Published online: 19 July 2024

1 Introduction

Large language model-based chatbots (LLM-based chatbots) are widely recognized for their advanced capabilities in natural language understanding and human-like text generation, making them a popular choice for artificial intelligence services.They have the potential to revolutionize various facets of human daily life, including education (Meyer et al., 2023; Tamkin et al., 2021; Rillig et al., 2023). LLM-based chatbots have amassed a burgeoning user base, among which students substitute a substantial group (Ofcom, 2023; Welding, 2023). For example, according to Ofcom's 2023 report on UK students, 79% of online teenagers aged 13–17 and 40% of children aged 7–12 apply generative AI tools and services, including LLM-based chatbots (Ofcom, 2023). Students' rapid adoption and potential reliance on LLM-based chatbots have raised both expectations and concerns about the applications of LLM-based chatbots in education (Kasneci, 2023; Strzelecki, 2023). One significant concern surrounding LLM-based chatbots is their potential to generate high-quality content, thereby aiding students in completing assignments without honing their knowledge and skills (Perkins, 2023).

Adding to this, students demonstrate a significant tendency to apply AI tools, including LLM-based robots, in their assignments. According to the BestColleges survey of 1,000 current US undergraduate and graduate students, 50% of those who have utilized AI tools acknowledged their use to aid in completing assignments or exams (Kaplan-Rakowski et al., 2023). Students outsourcing all assignments to LLM-based chatbots may harm their critical thinking, creativity, learning motivation, learning outcomes, and personal development (Yu, 2023). This makes these tools challenging for students in the early stages of acquiring scientific skills and domain knowledge, as they do not understand their underlying mechanisms, potential weaknesses, and enough domain knowledge to identify mistakes (Jungherr, 2023). On the other hand, the difficulty in distinguishing between content generated by LLM-based chatbots, such as ChatGPT, and authentic student writing increases teachers' difficulties in assessing students' learning performances and outcomes. We know little about how different methods of completing assignments, whether through outsourcing to an LLM-based chatbot or not, influence students' learning. Existing research focuses on evaluating the application of LLM-based chatbots in higher educational contexts, such as assessing teachers' user experience (Zhai, 2022), and students' acceptance and use of these tools (Strzelecki, 2023). However, limited research has investigated the chatbots' impact in the context of K-12 education (Alemdag, 2023; Wollny et al., 2021).

We find it important to understand the impact of completing assignments with LLM-based chatbots on students. Furthermore, there is a lack of empirical case studies exploring how to optimize assignment design in response to the educational changes introduced by LLM-based bots. To the best of our knowledge, there has been no research specifically focused on this topic. Thus, the main purpose of this study is to investigate the impact on middle school students of using LLM-based chatbots to complete their assignments. Our research questions are as follows:

RQ1 What is the impact on middle school students' assignment performance, learning test scores, learning motivation, and learning satisfaction when they cooperate with an LLM-based chatbot to complete assignments?

RQ2 Are students satisfied with doing assignments with the LLM-based chatbot and willing to use it for further learning?

RQ3 What is the experience of designing LLM chatbot-assisted assignments for middle school students?

2 Background

2.1 LLM-based chatbot and its educational application

Large Language Models (LLM), pre-trained language models of significant size(e.g., model size or data size), were first coined to distinguish from other language models due to their emergent abilities on downstream tasks, including natural language understanding, human-like text generation across various topics, interaction abilities with users in different contexts (Zhao et al., 2023). Among the applications of LLM, the chatbot has emerged as a popular format. A chatbot is "A computer program designed to simulate conversation with a human user, usually over the internet; esp. one used to provide information or assistance to the user as part of an automated service (Oxford English Dictionary, 2023)". With rich knowledge captured from large-scale pretraining data, LLM models have the potential to serve as domain experts to improve chatbot performance (Zhao et al., 2023).

LLM-based chatbots have gained an increasing user base and are anticipated to bring revolutionary impact to human daily lives. One of the phenomenal LLM-based chatbot products is ChatGPT, which has gained billions of users (Mahajan, 2023). Among all LLM-based chatbot users, students become a substantial group. Take the UK as an example, K-12 UK students are more likely to apply generative artificial intelligence (AI) than adults, and over 40% of students aged 7–17 use generative AI tools or services, including LLM-based chatbots (Ofcom, 2023). In a recent survey by BestColleges, where 1,000 current US undergraduate and graduate students participated, it was found that over 43% of respondents have hands-on experience with AI tools, including ChatGPT (Welding, 2023). This prevalence among students not only attests to the popularity of LLM-based chatbots but also emphasizes their potential to revolutionize students' learning. As the user base of LLM-based chatbots continues to grow, it becomes imperative to explore effective ways to integrate these tools into students' learning process, ensuring they enhance students' learning experiences and outcomes while upholding ethical standards.

Indeed, learning with chatbots has already been regarded as one of the most essential and practical approaches to improving students' learning outcomes, learning enjoyment, learning experiences (Ait et al., 2023; Fryer et al., 2019), and self-efficacy (Chang et al., 2022) by helping teachers tailor educational content, promoting quick access to educational information, multitasking, etc. For example, Guo et al. (2023) integrated CNN-based argumentative chatbots into 44 undergraduate students' in-class debates for three 90-minute sessions. The results showed that students assisted by chatbot not only reported relatively higher enjoyment, but also performed significantly better in argumentation generation, organization, sufficiency, and elaboration. LLM-based chatbots are expected to have a wider educational application and benefit to improve student's learning experiences and outcomes, as LLM enables chatbots to generate cohesive and informative human-like responses to students (Kasneci, 2023). Existing research focuses on evaluating the application of LLM-based chatbots in higher educational contexts, including assessing teachers' user experience (Zhai, 2022), students' motivation (Jishnu et al., 2023), instructors' and students' perception (Shoufan, 2023), and attitudes (Strzelecki, 2023) of these tools. However, limited research has investigated the chatbots' impact in the context of K-12, longitude, and non-language learning (Alemdag, 2023).

2.2 LLM-based chatbots' impact on students' assignments

Assignments are essential components for secondary school students which can promote students' engagement and help them apply the knowledge learned in class (Buijs & Admiraal, 2013; Grogan, 2017). Inquiry-based learning (IBL) is the process of acquiring knowledge and skills by seeking information and emphasizing the active participation of learners in discovering new information (Adarkwah et al., 2023). To encourage inquiry-based learning and active learning, many assignments require students to select a question, retrieve information (either online or textbook), integrate this new information with their knowledge learned in class, and organize them to answer the question (Pedaste et al., 2015). By engaging in inquiry-driven assignments, students not only deepen their understanding of ICT concepts but also hone their ability to navigate digital landscapes, critically assess information, and apply their findings to solve complex problems (Adarkwah et al., 2023).

LLM-based chatbots have the potential to outperform search engines for students' information retrieval in inquiry-driven assignments, especially on those that demand content synthesis (AlAfnan et al., 2023; Lee, 2023). By prompting them with application-based queries, LLM-based chatbots can generate logical answers for students to answer theory-based questions and generate ideas for application-based questions (Lee, 2023). LLM-based chatbots can also create individualized feedback and facilitate students' self-regulation and knowledge construction in learning (Gan et al., 2023). Both benefits can help to improve students' learning motivation and engagement (Wu et al., 2024). Nevertheless, empirical evidence is essential to substantiate these claims of LLM-based chatbots' benefits on middle school students' inquiry-based assignments.

In addition to LLM-based chatbots' potential advantages for students' assignments, their application in education has brought significant ethical and practical challenges (Adiguzel et al., 2023; Kasneci et al., 2023; Karthikeyan, 2023). One of them is that students may outsource all assignments to it (Karthikeyan, 2023). Students demonstrate a significant potential for applying AI tools, including LLM-based robots, in their assignments. According to the BestColleges survey of 1,000 current

US undergraduate and graduate students, 50% of those who have utilized AI tools acknowledged their use to aid in completing assignments or exams (Welding, 2023). Moreover, some students did not perceive AI-aided plagiarism as fundamentally distinct from plagiarism that occurred in the past (Kaplan-Rakowski et al., 2023).

Teachers are concerned about students' outsourcing behavior. However, teachers have already reflected that it was difficult to discern whether students complete written homework assignments independently or not (Liu et al., 2023). Take essay writing as an example, in an experiment where 69 high school teachers and 140 students read pairs of essays—one by a student and one by ChatGPT—and identified which was generated by the chatbot, the results showed that accuracy was only 70% for teachers, and 62% for students; well-written essays were especially hard to differentiate from the ChatGPT texts (Waltzer et al., 2023). Even programs could not detect AI-generated content (Moritz, 2023). The difficulty in distinguishing between content generated by AI, such as ChatGPT, and authentic student writing is a challenging theme in educational research. In other words, it is hard for teachers to assess students' original learning outcomes if they apply tools like ChatGPT to help them finish assignments. We know little about how different methods of completing assignments, whether through outsourcing to an LLM-based chatbot or not, influence students' learning outcomes.

Over-reliance on LLM-based robots may harm students' critical thinking and active learning motivation, and may further affect students' learning outcomes and personal development (Yu, 2023). Considering this, educational institutes have published regulations according to the usage of Generative AI technologies. For example, The University of Oxford clarified that unauthorized use of AI tools (e.g. ChatGPT) in assessed works is "a serious disciplinary offense"(University of Oxford). Guidelines have also been developed to scaffold students better use AI tools. For example, the University of North Carolina at Chapel Hill developed a set of guidelines for instructional applications of generative AI (The University of North Carolina at Chapel Hill, 2023). To the best of our knowledge, there has been no research specifically dedicated to examining the real impact of completing assignments with LLM-based chatbots on students. Furthermore, there is a lack of empirical case studies exploring how to optimize assignment design in response to the educational changes introduced by LLM-based bots.

3 Study design

Given our focus on the impact of LLM-based chatbot-assisted assignments on middle school students' learning, we choose a chatbot suitable for our research context, design learning content according to Compulsory Education Information and Communication Technology (ICT) Curriculum Standards, design assignments, and conduct quasi-experimental research. Participants are randomized to either the experimental group (completing assignments with LLM-based chatbot) or the control group (completing assignments with search engine) by the teacher. To evaluate students' assignment performances, we self-developed a scoring rubric through thematic coding. To examine the students' learning outcomes, we designed a pretest and a posttest.

3.1 LLM-based chatbot selection

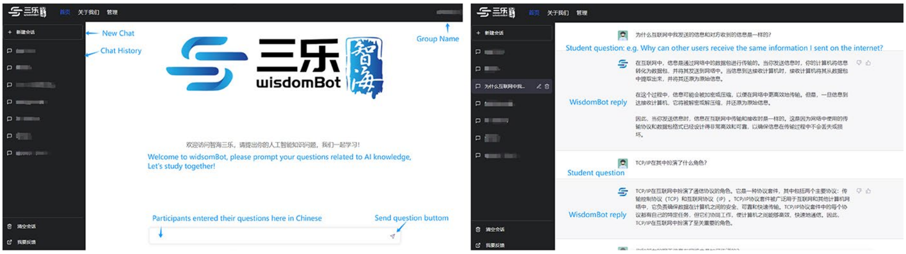
Due to our potential research participants (Chinese is the native language of participants), we investigated through existing LLM-based chatbots and chose Zhihai-Sanle (wisdomBot). WisdomBot is a closed-source LLM-based chatbot developed for Chinese learners to learn Artificial Intelligence and related Information and Communications Technology knowledge. It is jointly developed by Zhejiang University, Higher Education Publishing House, Aliyun and Huayuan Computing. WisdomBot utilizes the AliCloud Tongyi Qianwen (7B) general model as its foundation model. It has been further enhanced through pre-training and fine-tuning using a diverse corpus comprising high-quality textbooks, academic papers, and graduation theses in the field of AI.

To ensure the usability of wisdomBot, we compared its response time, i.e. the time spent on the first word or token generated by the language model, with ChatGPT 3.5-turbo. According to 5 requests, wisdomBot outperformed ChatGPT 3.5-turbo. The average response time of wisdomBot is 1.142 ± 0.007 s, while that of ChatGPT 3.5-turbo is 1.402 ± 0.013 s. As previous studies have proven the usability of ChatGPT 3.5-turbo in middle school students' learning (Li et al., 2023; Meyer et al., 2023; Zha et al., 2024), wisdomBot's response time is suitable for our context. We did not specifically care about hallucination because there is also fake information and/or news online. Indeed, we encouraged and expected students to audit the generated context.

Therefore, we believe that wisdomBot is suitable for our context. Additionally, we are authorized to attain or withdraw participants' log data from wisdomBot. The user interface of wisdomBot is shown in Fig. 1.

3.2 Learning content

In China, the compulsory education system is guided by Compulsory Education Curriculum Standards (Standards), published by the Chinese Ministry of Education to ensure a consistent educational experience across the country (Li & Xue, 2021). These Standards include discipline core qualities (key competencies), learning content requirements, academic requirements, and teaching suggestions (Wei, 2022). All



欢迎访问智海三乐，请提出你的人工智能知识问题，我们一起学习！

Welcome to widsomBot, please prompt your questions related to AI knowledge, Let's study together!

为什么互联网中我发送的信息和对方收到的信息是一样的?

Student question: e.g. Why can other users receive the same information I sent on the internet?

在互联网中，信息是通过网络中的数据包进行传输的。当你发送信息时，你的计算机将信息转化为数据包，并将其发送到网络中。当信息到达接收计算机时，接收计算机将其从数据包中提取出来，并将其还原为原始信息。

在这个过程中，信息可能会被加密或压缩，以便在网络中更高效地传输。但是，一旦信息到达接收计算机，它将被解密或解压缩，并还原为原始信息。

因此，当你发送信息时，信息在互联网中传输和接收时是一样的。这是因为网络中使用的传输协议和数据包格式已经设计得非常高效和可靠，以确保信息在传输过程中不会丢失或损坏。

TCP/IP在其中扮演了什么角色?

TCP/IP在互联网中扮演了通信协议的角色。它是一种协议套件，其中包括两个主要协议：传输控制协议（TCP）和互联网协议（IP）。TCP/IP协议套件被广泛用于互联网和其他计算机网络中，它负责确保数据在计算机之间的安全、可靠和快速传输。TCP/IP协议套件中的每个协议都有自己的特定任务，但它们协同工作，使计算机之间能够高效、快速地通信。因此，TCP/IP在互联网中扮演了至关重要的角色。

Fig. 1 User interfaces of wisdomBot

schools are required to follow the standards and can adjust and additions according to their actual situation and characteristics.

According to the latest Standards, all Chinese middle school students need to learn Internet Applications and Innovation (Chinese Ministry of Education, 2022). We incorporated the standard learning requirements into two modules. For the first module, students will learn about the definition of the Internet, the history of Internet development, the Internet application and innovation, and the Internet's influence on different industries. For the second module, the students will learn about internet connection, internet services, Internet Protocol, TCP/IP, IP addressing, and DHCP. In each lesson, the teacher provides the same slides for the experimental group and the control group.

3.3 Assignment design

Digital learning and innovation is one of the key competencies required by the Standards for the ICT curriculum. Specifically, students are expected to: (1) Effectively search for learning resources according to their learning needs, explore new methods and modes of IT-supported learning, and improve the quality of learning with the help of IT. (2) Choose appropriate digital devices to support learning in the learning process, change the way of learning, and have the ability to utilize information technology for independent and cooperative learning.

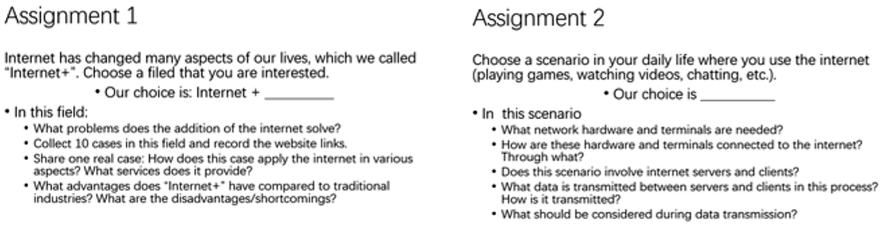
The recommended learning activities, as outlined by the Chinese Ministry of Education (2022), emphasize active and inquiry-based learning strategies facilitated by digital technologies, including search engines. Given the potential of LLM-based chatbots to surpass traditional search engines and emerge as prominent digital tools in the future, integrating them into students' inquiry-based information retrieval assignments aligns well with the ICT curriculum objectives.

Three assignments are designed, including pilot assignment, assignment 1, and assignment 2. In the pilot assignment, we first taught participants how to use wisdomBot (Baidu, the largest Chinese search engine for the control group) to prompt and get information. We then highlighted prerequisites and ethical considerations associated with using LLM tools (e.g., recognizing that LLM is not a person, acknowledging the potential generation of false information, and encouraging critical evaluation of content produced by LLM). In the end, participants collaborated in groups, working with wisdomBot for 10 min on the topic they liked.

Each of Assignments 1 and 2 contains 4–5 questions and lasts for 20–30 min (as shown in Fig. 2). For each assignment, students are required to search for additional information and integrate it with the knowledge learned in each module. Students in the experimental group need to finish them with wisdomBot; while students in the control group need to finish them with the search engine.

3.4 Assignment scoring rubric

To analyze students' assignments, we applied a thematic coding process to develop a scoring rubric (dimensions are shown in Table 1) and assessed students' work using this rubric. For the first assignment, we divided assignment sheets into two equally



Assignment 1

Internet has changed many aspects of our lives, which we called "Internet+". Choose a filed that you are interested.

- Our choice is: Internet + ________

- In this field:
  - What problems does the addition of the internet solve?
  - Collect 10 cases in this field and record the website links.
  - Share one real case: How does this case apply the internet in various aspects? What services does it provide?
  - What advantages does "Internet+" have compared to traditional industries? What are the disadvantages/shortcomings?

Assignment 2

Choose a scenario in your daily life where you use the internet (playing games, watching videos, chatting, etc.).

- Our choice is ________

- In this scenario
  - What network hardware and terminals are needed?
  - How are these hardware and terminals connected to the internet? Through what?
  - Does this scenario involve internet servers and clients?
  - What data is transmitted between servers and clients in this process? How is it transmitted?
  - What should be considered during data transmission?

Fig. 2 Translated Assignment 1 and Assignment 2

Table 1 Dimensions of assignments scoring rubric we developed through thematic coding

Dimension	Definition
Completion	Assess the extent to which the assignment is fully completed.
Creativity	Evaluate the level of creativity demonstrated in the work.
Accuracy	Examine the correctness of the content or information presented.
Material Refinement	Assess the degree of secondary processing or refinement of materials(from wisdomBot or search engine)
Depth	Assess the extent to which the content demonstrates thorough understanding.
Logic	Evaluate the logical flow and coherence of the work.
Page Aesthetics	Consider the visual attractiveness and formatting skills in creating the document.

Note The first two authors scored 1–5 for each dimension, full matrix is appended in Appendix

sized sets. The first two authors independently analyzed the first set of assignment sheets to derive an initial set of scoring rubric dimensions. They then met to consolidate and reconcile these dimensions into one rubric. Subsequently, they applied the rubric to score the second set of assignment sheets, achieving a Cohen's kappa of 0.76, indicating a high level of agreement between scorings. The first author completed scoring the first set of assignment sheets. For the second assignment, we divided assignment sheets into two equally sized sets, and the first two authors applied the rubric to score the first set of assignment sheets. The Cohen's kappa was 0.78. The first author then finished scoring the second set of assignment sheets.

3.5 Instrument design

The instruments used in this study included subjective tests and self-reported questionnaires. In China, ICT courses do not have standardized quizzes and final exams. Teachers usually create their assessments to evaluate students' learning outcomes and assign grades accordingly. Thus, in our research, the first two authors (one is a doctoral student majoring in educational technology, another is an experienced ICT teacher) self-developed the pretest and posttest. All tests are completed independently by students, without the use of LLM, search engines, or collaboration with peers.

The pretest includes 15 concepts that students may have heard about in their daily lives and learned previously. Students were asked to tick the terms/concepts they had heard about. The posttest includes two parts: (1) knowledge test, 11 True/False questions about learned concepts; (2) migration test, 10 True/False questions regarding internet connection in the context of watching videos on cell phones. We designed a knowledge test to investigate whether completing assignments with wisdomBot can help students remember and understand concepts better; we also designed a migration test to test if completing assignments with wisdomBot can help children to better apply the knowledge they learned to the new context.

The questionnaire investigated students' demographic information and learning motivations (including intrinsic goal orientation, assignment value, control of learning belief, and self-efficacy for learning performance). We adopted items from the well-established Motivated Strategies for Learning Questionnaire (MSLQ). As students do not take a test in ICT courses, they are not likely to have test anxiety and Extrinsic Goal Orientation (in fact, we measured them in the pretest and students are confused about these two dimensions). We also dropped the items related to the examination and sub-scale of learning strategies as we focus on the impact of LLM on students' learning motivation. Answers were scored on a five-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). Additionally, in the post questionnaire, we added two dimensions to reflect their feedback, i.e. students' enjoyment(score from 1 to 10) and satisfaction(score from 1 to 10) with the courses, and their LLM continue usage intention. The continued usage intention scale is adopted and adapted from Lin (2011).

4 Study method

4.1 Ethical statement

This study belongs to a larger program approved by the Ethics Committee of the Department of Psychology department in the University. Informed consent was obtained before our research.

4.2 Participants and recruitment

We chose middle school students (children aged 11–14) as our participants for several reasons. First, children in this age group are capable of interacting with and prompting LLM-based chatbots for learning purposes (Jauhiainen & Guerra, 2023). Middle school students found tools like ChatGPT highly effective and are willing to apply ChatGPT to facilitate their learning and broaden their interests (Li et al., 2023). Secondly, Chinese students aged from 11 onward tend to engage increasingly in online activities and digital technologies, suggesting that many of them are in the initial stages of interacting with LLM-based chatbots. (Youth League, 2022). This provides us with an opportunity to guide participants in responsibly using LLM-based chatbots right from the start.

Participants were recruited through convenience sampling. We sent emails to local middle schools. These emails provided details about the research study and outlined the study design and prerequisite(students need to assess laptops once a module) for participation. One middle school offered us the opportunity to collaborate with a teacher for our research in seventh-grade ICT courses. Four classes were involved with approximately 36 students in each class. Classes were randomized to either the experimental group or the control group by the teacher. No financial incentives were paid to participants. All participants have the right to quit the survey or withdraw their data before November 2023. In the end, 127 participants were involved in our research (average age = 12.37, s.d. = 0.51), including 61 participants in the control group, and 66 participants in the experimental group.

4.3 Experimental procedure

Our experiment lasted 9 weeks. Table 2 shows the timeline and learning contents for both groups. Considering the limited research on applying LLM in classrooms and the potential ethical concerns associated with its use, we planned to co-design learning materials and assignments with the teacher meticulously throughout the research. To mitigate the potential impact of teachers' variability on the outcome, the assignments of all four classes were allocated by the same teacher (the second author) and supervised by the same teachers (the first two authors) throughout the study.

Before the first week, the first and the third authors interviewed the teacher (the second author) about her requirements of integrating a wisdomBot into her teaching, students' accessibility to computers, and learning designs, formats, and assessments commonly applied in ICT classes in previous years. Then, the first two authors designed the pretest and pilot assignment.

In the first week, we obtained informed consent and they finished a pretest including two sections: (1) a prior knowledge test; and (2) a questionnaire about their background information, learning motivation, and self-efficacy.

In week 2, classes were randomly divided into two groups: experimental group (2 classes of students completing assignments with LLM) and control group (2 classes of students completing assignments with search engine). Participants worked in

Table 2 Timeline and learning contents for both groups for 9 weeks

	Experimental Group	Control Group	Teacher
Week 1	Informed consent; pre-test		Co-design pilot assignment based on results
Week 2	Ethics instruction; pilot assignment		Reflect and Co-design Assignment 1
Week 3–4	Module 1 teaching: Getting to Know the Internet		Teaching
Week 5	Assignment 1 with wisdomBot	Assignment 1 with search engine	Reflect and Co-design Assignment 2
Week 6–7	Module 2 teaching: Exploring the Internet		Teaching
Week 8	Assignment 2 with wisdomBot	Assignment 2 with search engine	Reflect and summarise
Week 9	Post-test		

groups (2–3 people per group) to complete a pilot assignment in class. After the pilot assignment, the first two authors supervised and observed students' performance and interaction with LLM and search engines Based on the experience of the pilot assignment, we designed the learning materials for Module 1 and Assignment 1.

In the following weeks, students attended approximately 90-minute lessons taught by the second author; then they were allocated into groups to finish the assignment with wisdomBot or a Search engine. Based on students' performance and feedback in weeks 3–5, we co-designed learning materials for Module 2 and Assignment 2. In the last week, participants took the post-test.

When participants interacted with the WisdomBot, the first two authors supervised, supported, observed every class, and took notes on students' performances. Participants were encouraged to reflect and give feedback on their learning experiences throughout the course. After each assignment, the first two authors reflected on and discussed students' learning process, scored students' assignments, and the first author summarized potential learning strategies.

5 Data collection and analysis

5.1 Data collection

Pre-post tests and questionnaire data were delivered to each student and collected through online questionnaire platforms. Students completed assignment sheets on their computers, submitting them to their teachers in class. Students' conversations with wisdomBot were stored in the database of wisdomBot and exported after each assignment. To ensure participants' privacy, students utilized assigned indices instead of their actual names or any other identifying information for all measurements and LLM accounts. Additionally, after each assignment, teachers meticulously documented reflections on the design of assignments, observations of students' behaviors, and recommendations for further assignment design. The recorded material was then transcribed and summarized into themes, according to the action research method, to gain understandings of assignment design strategies and scaffolding strategies of LLM chatbot-assisted assignments for middle school students. After reflecting on the transcriptions, the first two authors completed the field notes and then designed for the next assignment. Noticeably, the design for children's interaction with LLM-based chatbots remained flexible to guarantee students' learning experiences in the classroom and in completing assignments.

5.2 Data analysis

Pre-post tests are scored by the first two authors based on the correct answers. In each test, one point was awarded for every correct answer, while incorrect responses received a zero point. All test scores have been standardized to a maximum of 10 points for further analysis and visualization.

For the self-reported instruments, we applied SPSS and Python to clean, describe, and analyze data. All the Cronbach's alpha values for each construct, measured in both the pre- and post-questionnaires, are above 0.8. The normality was checked in advance using the Shapiro-Wilk test and KS test. All the variables did not follow a normal distribution instead of Intrinsic goal orientation for experimental group post-test scores. Thus, non-parametric tests are applied for statistical analysis. Both parametric and non-parametric tests led to the same conclusions. For the sake of clarity, only non-parametric analyses are presented. To analyze students' assignment sheets, we calculated the submission rate and average word count for each assignment in both the experimental and control groups. We also assessed students' work using our self-developed scoring rubric (mentioned in Sect. 3.4). The first two authors compared students' assignments with wisdomBot's answers to score the dimension of material refinement and analyzed students' conversations with it to triangulate our field notes in class.

For teachers' reflection, action research was employed to summarize the experience of designing and scaffolding LLM chatbot-assisted assignments. Action research is a systematic intervention process initiated by insiders who conduct research about their professional actions and implement changes subsequently (Özer et al., 2020). This method is widely used in school-based pedagogy, emphasizing teachers' active participation in the research process and how they apply their knowledge to improve teaching practices, aiming to better comprehend teaching practices and develop teaching strategies (Henthorn et al., 2024). Action research includes observing, reflecting, acting, evaluating, and modifying and the findings are usually organized as themes to summarize the experience (Bilgic & Dogusoy, 2023). Aligned to previous research, the first two authors, who were also the designers and supervisors of the assignments, acted as insiders and provided observations and reflections aimed at answering "What is the experience of designing LLM chatbot-assisted assignments for middle school students?".

6 Results

6.1 Students' changes within groups

We applied the Wilcoxon Signed-Rank Test to examine students' pre and post-learning motivation (as shown in Table 3). No statistically significant improvements have been observed in each group. We calculated the mean of each construct and found that the control group's average intrinsic goal orientation and task value have slightly increased, the control of learning behaviors and self-efficacy of learning and performance has slightly decreased; while the experimental group's average intrinsic goal orientation, the control of learning behaviors, and self-efficacy of learning and performance have increased, task value has slightly decreased.

Table 3 Descriptive analysis and wilcoxon signed-rank test within groups

	Control group (*n*=61)						Experimental group (*n*=66)					
	Pretest		Posttest		Stas test		Pretest		Posttest		Stas test	
	M	SD	M	SD	Z	*p*	M	SD	M	SD	Z	*p*
IGO	3.75	0.76	3.67	1.08	–0.052	0.959	3.91	0.81	3.94	0.86	–0.040	0.968
TV	3.86	0.65	3.78	1.00	–0.005	0.996	3.87	0.77	3.97	0.70	–1.355	0.175
CLB	3.72	0.67	3.75	0.90	–0.365	0.715	3.66	0.68	3.70	0.74	–0.442	0.658
SLP	3.71	0.70	3.72	0.95	–0.318	0.750	3.73	0.86	3.77	0.87	–0.039	0.969

Note Stats Test represents Wilcoxon Signed-Rank Test. IGO represents intrinsic goal orientation, TV represents task value, CLB represents the control of learning behaviors, and SLP represents self-efficacy of learning and performance. *$p<.05$; **$p<.01$; ***$p<.001$

6.2 Students' differences between groups

6.2.1 Assignment assessment

The experimental group's submission rate (87.50%) and average Chinese word count (1,130) were higher than those of the control group (62.50%, 805). Figure 3 shows the distribution of students' average scores for assignment 1 and assignment 2 according to the scoring rubric. According to Mann Whitney test, significant differences were found between the control group and experimental group in word count (Z = -3.78, $p \leq .001$), completion (Z = -2.70, $p = .007 < .05$), accuracy (Z = -2.60, $p = .009 < .05$), material refinement (Z = -2.67, $p = .007 < .05$), depth (Z = -4.65, $p < .001$), and logic (Z = -4.99, $p < .001$). No significant difference has been found in page aesthetics (Z = -0.88, $p = .381 > .05$) and creativity (Z = -1.43, $p = .154 > .05$).

6.2.2 Learning test, learning enjoyment, and learning satisfaction

We applied the Mann-Whitney U test to examine the differences in subjective tests between the experimental group and the control group. As shown in Table 4; Fig. 4, Students did not exhibit significant differences in the pre-knowledge test (Z = -0.381, $p = .703$). However, the control group scored significantly higher than the experimental group in post-knowledge test (Z=-3.52, $p \leq$ 0.001). The experimental group has lower average scores in the post-knowledge test and migration test and higher average scores in perceived enjoyment and satisfaction. There were no significant differences between the control group and experimental group in the post-migration test (Z = -1.104, p = .270), course enjoyment (Z = -0.78, $p = .436$), and course classification (Z = -0.52, $p = .602$).

6.2.3 Learning motivation

As shown in Table 5, there were no statistical differences in learning motivation (including IGO, TC, CLB, and SLP) between the control group and the experimental group. However, when we analyzed the variations in each dimension between the pre-test and post-test, we found the experimental group might have a greater tendency to increase their IGO, CLB, and SLP (as shown in Fig. 5).

6.3 Students' satisfaction and their intention to continue using WisdomBot

Overall, students were satisfied with wisdomBot and showed continued enthusiasm for the use of it. Among 66 students in the experimental group, 43 students expressed that they were satisfied with WisdomBot, 20 students remained neutral, and only 3 students reported dissatisfaction. More than two-thirds of students and the teacher are willing to continue to use wisdomBot. Specifically, 44 students expressed a positive willingness to continue using WisdomBot, 18 students remained neutral, and only 4 students expressed a preference not to use WisdomBot in the future.

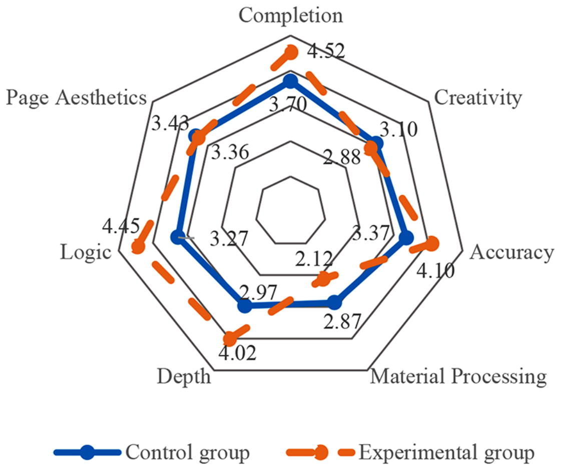


Fig. 3 Average assignment scores distribution (range 1–5 for each dimension) of experimental group and control group according to the scoring rubric

Table 4 Descriptive analysis and mann-whitney U test results of learning tests, learning enjoyment, and learning satisfaction between groups

	Control group (*n*=61)		Experimental group (*n*=66)		Stats test	
	M	SD	M	SD	Z	*p*-value
Pre KT	7.27	1.69	7.37	1.66	–0.38	0.703
Post KT	5.45	1.92	4.17	2.14	–3.520	<0.001***
Post MT	5.04	1.68	4.71	1.26	–1.104	0.270
Enjoyment	7.61	2.70	8.06	2.10	–0.78	0.436
Satisfaction	7.77	2.62	7.98	2.60	–0.52	0.602

Note Stats Test represents Mann Whitney U test. KT represents the Knowledge Test, MT represents the Migration Test. $*p<.05$; $**p<.01$; $***p<.001$

6.4 Learning design reflection during the research

We compiled three themes from the teachers' reflections throughout the research, including providing scaffolds for the potential prompts, suggesting group collaboration modes, and relinquishing the authority of the teacher. In the following session, we first present the teachers' reflection, then briefly describe where this reflection comes from and how we design assignments and interventions to fix some of them, along with the effect. Suggestions for future learning design will be discussed in the Sect. 8. In general, we reflected that teachers should facilitate students' autonomous learning when they interact with wisdomBot.

6.4.1 Providing scaffolds for the potential prompts

Students in the control group typed keywords in search engines, scanned the titles of websites, clicked on those related, read through them, and copied useful pieces into assignments. However, students in the experiment spent a lot of time prompting wisdomBot, copied without reading the generated content, even if the teachers have repeatedly emphasized the potential hallucination of wisdomBot.

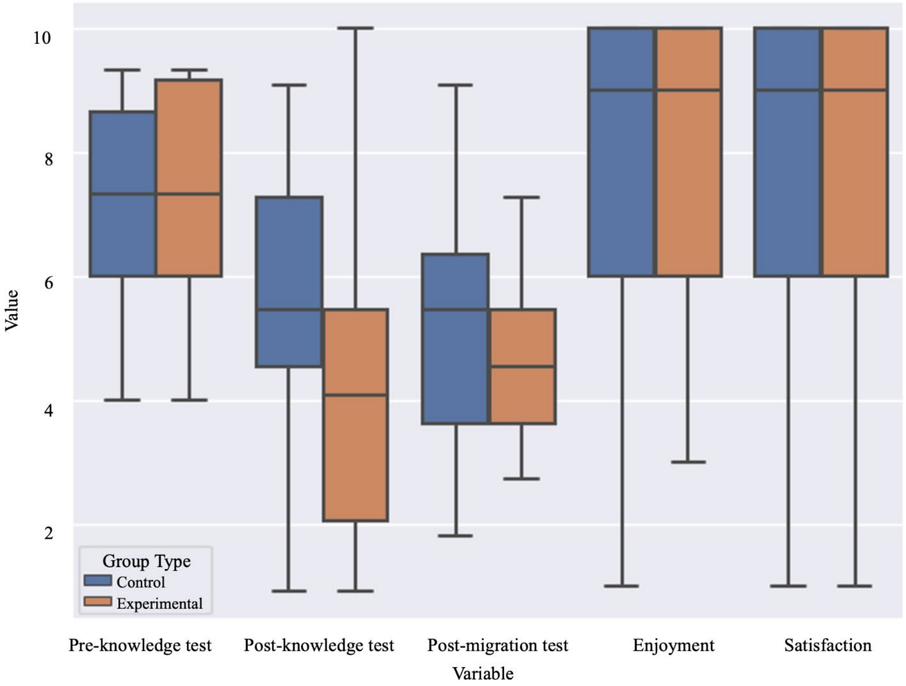


Fig. 4 Differences in subjective tests and learning experience between the experimental group and control group

Table 5 Descriptive analysis and mann whitney U test of learning motivation between experimental group and control group

		Control group (n=61)		Experimental group (n=66)		Stats test	
		M	SD	M	SD	Z	p
Pre-test	IGO	3.75	0.76	3.91	0.81	–1.033	0.301
	TV	3.86	0.65	3.87	0.77	–0.284	0.776
	CLB	3.72	0.67	3.66	0.68	–0.412	0.681
	SLP	3.71	0.70	3.73	0.86	–0.155	0.877
Post test	IGO	3.67	1.08	3.94	0.86	–0.837	0.403
	TV	3.78	1.00	3.97	0.70	–0.410	0.682
	CLB	3.75	0.90	3.70	0.74	–0.352	0.725
	SLP	3.72	0.95	3.77	0.87	–1.10	0.270
	Enjoyment	7.61	2.70	8.06	2.10	–0.78	0.436
	Satisfaction	7.77	2.62	7.98	2.60	–0.52	0.602

Note Stats Test represents Wilcoxon Signed-Rank Test. IGO represents intrinsic goal orientation, TV represents task value, CLB represents the control of learning behaviors, and SLP represents self-efficacy of learning and performance. *p<.05; **p<.01; ***p<.001

Specifically, in the pilot assignment, students in experimental groups took longer than expected to log in to the wisdomBot system. Students also spent a lot of time typing questions (to clarify their requirements, as the wisdomBot was their only information source). They preferred to type short questions to prompt wisdomBot and did

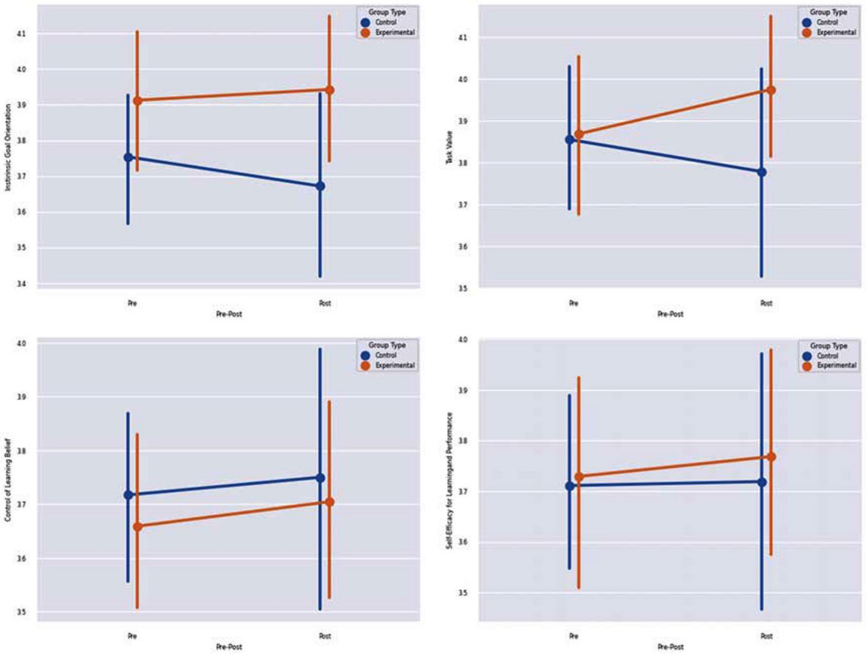


Fig. 5 Pretest and posttest changes in learning motivations between experimental group and control group. *Note* The bar represents 95% confidence interval

not like reading text generated by wisdomBot. We further found in the first assignment that students still spent most time typing questions, and some students copied and pasted materials without detailed reading and evaluating them. After assignment 1, the teacher reemphasized the potential mistakes (hallucination) wisdomBot might make and nudged them to check the next time. In the second assignment, we provided potential prompts to save students' typing time and situated learning assignments into students' daily lives to attract students. Moreover, we encouraged them to evaluate the generated content. The scaffolds on the prompt helped students to interact with wisdomBot much more easily. However, the students spent the rest of the time chatting with others rather than reading and discussing the generated text.

6.4.2 Suggesting group collaboration mode

Students in the control group were familiar with the collaboration task with the search engine, so they collaborated quite well, discussed the uncertain answers, and finished the task. However, students in the experimental group had difficulties in shifting collaboration modes when they had a new 'knowledgeable' member, who seemed to answer any questions and finish the task immediately.

Specifically, in the pilot assignment, everyone in the experimental group had access to wisdomBot. When they worked in groups to finish the pilot assignment, they did not talk to each other about the assignment or collaborate; rather they talked with wisdomBot. They shared their feelings with their team members when they saw wisdomBot's feedback. They were also likely to chat about unrelated topics with wisdomBot. This may be because children are new to the wisdomBot. In the first assignment, we limited the amount of access to wisdomBot(one group one access), and the teacher encouraged students to group work and allocate several roles/modes the group could take to collaborate. However, the teacher found that students became much more 'silent' when they finished the assignment. Some of the most skilled members in the group finished chatting with the wisdomBot, copied the answer quickly, and then either small chat with their peers or chatted with wisdomBot about other topics. In the second assignment, the teacher suggested a collaboration mode for the group (e.g. discuss the prompt together, then one student type, and others audit the generated results). Students discussed more and collaborated more this time. While some groups did not allocate the assignments properly, some students couldn't get involved and got bored quickly. Groups in the experimental group with higher digital skills(e.g. typing faster) finish assignments with higher quality. They typed faster and were more used to reading on the screen.

6.4.3 Relinquishing the authoritarian of the teacher

Teachers need to transfer to the role to facilitate students' autonomous learning with wisdomBot and nudge them to ask questions on generated content. It was observed that during the initial stages of learning how to interact with wisdomBot, students displayed a tendency to deviate from the provided instructions. Once students became immersed in their interactions with WisdomBot, there was a discernible challenge in redirecting their attention back to the guidance provided by teachers. In other words, they prefer to autonomously explore and use wisdomBot. Despite this, there was a notable inclination among students to seek assistance from teachers when encountering difficulties, whether related to the usage of WisdomBot or in prompting questions. However, compared to the pilot assignment, students did not ask follow-up questions or discuss with their group members even if they had questions or did not understand some terms generated by wisdomBot (e.g. 'internet terminal'). Only two students turned to teachers for answers, while the rest of them did not.

7 Discussion

Firstly, our results indicate several benefits of integrating LLM-based chatbots into the learning of middle school students, including increasing submission rate and a tendency to increase learning motivation (though not significant). No significant difference has been found in page aesthetics and creativity. We contextualize our results in cognitive load theory and previous studies to explain our results.

We found that students using wisdomBot (LLM-based chatbot) to assist their assignments showed higher submission rates. We attribute the higher submission rate to the fact that LLM-based chatbots decrease students' cognitive load in accomplishing assignments (Lan & Chen, 2024). LLM-based chatbots streamline human-computer interaction and alleviate extraneous cognitive burden (Adarkwah et al., 2023). According to cognitive load theory, the less extraneous cognitive load, which includes integrating information sources that are distributed in place or time and searching for information needed to complete a learning task, can help children focus on the intrinsic nature of learning tasks (Van Merrienboer & Sweller, 2005).

There are no significant differences observed in learning motivation (i.e. intrinsic goal orientation, control of learning behaviors, and self-efficacy of learning and performance), satisfaction, enjoyment, and students' ability to migrate their knowledge. Students using LLM-based chatbots to assist with their assignments showed a tendency to increase their learning motivation. This aligns with previous studies on chatbots in education (Nee et al., 2023; Okonkwo & Ade-Ibijola, 2021), where students' motivation increased when these tools are introduced in class.

Our study also confirms the high usability of LLM's natural, human-like interaction mode in a K-12 educational context. Despite the relatively long period of time students spent on typing prompts, more than half of the students in the experimental group are satisfied with completing assignments with wisdomBot and are willing to use it further. Our participants' high satisfaction and continuance use intention is consistent with previous research (Theophilou, 2023). This is aligned with previous studies on undergraduate students who apply ChatGPT to write their assignments (Playfoot et al., 2024).

Secondly, our results also indicate several concerns raised regarding integrating LLM-based chatbots into middle schools' learning. Students who completed assignments with wisdomBot had significantly lower scores on the material refinement and creativity dimensions of the scoring rubric and the post-knowledge test. They also had relatively lower scores on the migration test.

On one hand, these findings align with the concerns raised about LLM-based chatbots' potential negative impact on students' assignment completion (Adiguzel et al., 2023; Kasneci et al., 2023). It has been worried that outsourcing assignments to LLM-based chatbots may decrease students' learning motivation and critical thinking skills(Karthikeyan, 2023; Yu, 2023). Although, during this study, researchers repeatedly emphasized and reminded students to be aware of the hallucination and to critically review rather than copy AI-generated content, it may not be enough to properly guide students in using LLM-based chatbots to enhance, rather than replace, their learning. Willingness to use LLM-based chatbots was greater when the likelihood of detection was minimal and the consequences were mild, especially among individuals demonstrating a higher degree of task apathy (Playfoot et al., 2024). We observed that students were more willing to ask follow-up questions to wisdomBot in the pilot assignment(where they asked whatever questions they wanted) compared to the other two assignments (where they needed to ask questions to finish assignments despite their uncertainty about knowledge). This may indicate students' potential apathy on the assignments. Revolutionizing learning assignments and assessments, and designing guidelines and instructions for using LLMs in K-12 education, while emphasizing students' intrinsic motivation and curiosity for autonomous learning, can guide students to use LLM-based chatbots ethically and effectively (see Sect. 8).

On the other hand, the differences in post-knowledge and migration tests between the control group and experimental group may also related to different learning objectives and collaboration patterns in finishing assignments.

In terms of learning objectives, students in the control group acquire both declarative knowledge and procedural knowledge, tasked with learning objectives that demand a higher level of cognitive skill. They need to understand the problem, use search engines to retrieve information online, analyze its content, evaluate its relatedness to the problems, and create the assignment. They learn not only procedural knowledge (how to use search engines), but also factual knowledge (new facts/news found online), conceptual knowledge (new concepts and terms found online), and metacognitive knowledge (e.g. how to evaluate the relevance of retrieved information) (Morrison & Barton, 2018). In contrast, students in the experimental group focused more on articulating their needs in order to prompt the chatbots to produce accurate and comprehensive text with syntactic logic. WisdomBot enabled students to (semi)automatically complete tasks that were previously considered evidence of skill acquisition, potentially affecting grading and skill development (Jungherr, 2023). This led their study to be superficial, rather than deeply engage with and understand information (Lan & Chen, 2024). Learning objectives that require higher cognitive skill levels lead to deeper learning and transfer of knowledge (Adams, 2015). This may explain why the control group scored significantly higher on the post-knowledge test and slightly higher on the post-migration test.

In terms of collaboration patterns, we argue that wisdomBot may have authority over children's information retrieval and learning process. Authority is the power of certain people, objects, representations, or ideas that affect social communication and interaction (Bencherki & Cooren, 2019). In a collaborative learning environment, the presence of authority figures, such as teachers or textbooks, discouraging critical analysis and questioning of representations, can result in the convergence of students' learning outcomes (Hübscher-Younger & Narayanan, 2003). As the wisdomBot was pre-trained and fine-tuned in the field of AI education (see Sect. 3.1), students may find the answers generated by wisdomBot align closely with the content and format they anticipate from their textbooks and standard answers. As a result, their assignment converges to content generated by wisdomBot. This may further lead students to perceive themselves as passive recipients of knowledge rather than active contributors to knowledge construction (Hübscher-Younger & Narayanan, 2003). This argument is further validated by students' decreased follow-up questions or discussions over time.

Moreover, instruction-tuned LLMs-based chatbots, such as wisdomBot, can outperform children in introductory, standardized, and language-based tasks (van Dujin et al., 2023). When students discovered that wisdomBot outperformed them and their teammates, they may trust wisdomBot and grant it authority more and more as the algorithms demonstrated efficiency and objectivity (You et al., 2021). As a result, assignments from students in the experimental group had higher word counts and significantly higher scores in the completion, accuracy, depth, and logic dimensions. The authority of wisdomBot could be strengthened within Chinese Confucian-heritage education systems, wherein students are encouraged to respect the expertise of their teachers and adhere to their instructions and suggestions, with less emphasis placed on fostering critical thinking and creativity. (Ho et al., 2002).

Noticeably, as we use a paper-based knowledge migration test rather than an action-based assessment tool, we need to be careful to conclude the differences in higher-order learning between the control group and the experimental group. Previous studies found that knowledge tests were beneficial for knowledge learning rather than higher-order learning (including knowledge transfer and migration); students' higher-order learning is notably influenced by their participation in higher-order retrieval exercises or low-stakes quizzes featuring complex materials, where students actively retrieve and apply their knowledge (Agarwal, 2019).

8 Implication

LLM-based chatbots have the potential benefits of decreasing children's cognitive load so as to help them focus on complex problems. To enhance the potential benefits and address concerns about integrating LLM-based chatbots in middle school educational settings, our results offer implications and design considerations for both learning and LLM-based chatbot design, including revolutionizing learning objectives in the assignments and assessment, scaffolding students' collaboration with LLM-based chatbots, and design guidelines and instruction about using LLM-based chatbot in K-12.

8.1 Revolutionizing learning objectives in assignments and assessment

Our findings regarding variations in assignment scores and subjective test scores between groups indicate the importance of designing assignments that focus on high-order learning objectives. Similarly, assessments should gauge knowledge transfer, human-machine collaboration, and high-order thinking when students utilize LLM-based chatbots to complete assignments. Our implications align with arguments made in previous studies regarding the transformation of assignments and assessments, considering the impact of LLMs (Adeshola & Adepoju, 2023; de Winter, 2023; Malik, 2023). If we only assess students' assignments by the scoring rubric (in the dimensions of completion, logic, wording, and accuracy) LLM performances are likely to exceed human performances, particularly among K-12 students who have limited professional knowledge. Indeed, LLM has been proven to pass the high-school examination with a relatively high score (Grassini, 2023). In light of this, Adeshola and Adepoju (2023) proposed to create assessment techniques that take limited AI-generated text into account, e.g. assess students' analysis and evaluation of text produced by AI. Moreover, as LLM-based chatbots alleviate extraneous cognitive burden, assignments designed to integrate these tools should emphasize the intrinsic cognitive load of learning tasks, such as problem complexity (Ayres, 2006).

LLM-based chatbots can support the revolution by helping educators to generate questions and to provide in-time feedback. LLM-based chatbots can randomly generate different scenarios, multiple-choice questions, and materials for educators to design the assessment (Kumar & Lan, 2024; Olney, 2023; Xiao et al., 2023). These AI-generated materials not only offer cost-effective solutions for developing assessment tools but also enrich the diversity and customization of assignments and assessments (Cheung et al., 2023). LLM-based chatbots can also provide personalized and diverse contexts and scenarios for context-based and scenario-based assignments/

assessments, where students can apply their knowledge and higher-order thinking skills (Chi & Liu, 2023; Taasoobshirazi & Carr, 2008). For example, teachers can prompt LLMs to generate diverse daily contexts where scientific knowledge, such as Newton's second law, needs to be applied to solve problems. Students can select an interesting scenario, connect it to the science knowledge they've learned, and solve the problem by applying that knowledge. Moreover, LLMs can provide multi-round, in-time, and relatively accurate feedback during/on students' assignments and examinations (Fuller et al., 2024; Guo & Lee, 2023; Lee et al., 2024). For example, LLMs can accurately and comprehensively evaluate points of students' argumentations, especially when identifying relatively short and quantitative evaluation points such as claims, evidence, and rebuttals (Wang et al., 2024). In addition, LLMs can offer explanations and suggestions for improvement, guiding students in refining their understanding and application of the knowledge (Tanwar et al., 2024). By analyzing students' interaction log data (e.g. conversation records) with LLM-based chatbots, educators could better assess the level of students' high-order thinking skills.

8.2 Scaffolding students' collaboration with LLM-based chatbots

The potential authority of LLM-based chatbots in children's learning may lead them to perceive themselves as passive recipients of knowledge rather than active contributors to knowledge construction. The potential apathy of students toward the assignments is another concern, evidenced by students' reluctance to seek clarification. To address these concerns, we expect that both designers and educators can scaffold students' autonomous learning in their collaborations with LLM-based chatbots. Autonomous learning (also known as Self-Directed Learning, SDL) emphasizes learners' responsibility for goal-setting, material selection, learning activities, and/or assessment during their learning process (Masouleh & Jooneghani, 2012).

Educators play a crucial role in students' SDL, including organizing learning materials and activities, facilitating to motivate students, helping students to get resources, evaluating, and counseling (Yan, 2012). Educators can facilitate students in properly utilizing LLM in SDL by teaching them how to effectively prompt for their requirements, motivating students to evaluate the generated text, and evaluating their learning outcomes. Students may become more discerning in their evaluation of the content they generate by recognizing the potential for bias and error in AI systems, such as hallucination and misinformation (Lee et al., 2021). Besides instructions, educators should provide emotional interaction and support to students when they are using LLMs. Indeed, we found that students shared more about their feelings (rather than content generated by wisdomBot) with group members when they interacted with wisdomBot. It has been proved that teachers' and peers' emotional support on students' motivation and engagement positively affects their learning autonomy and social behaviors (Ruzek et al., 2016; Wentzel, 2016).

The design of educational LLMs-based chatbots could also scaffold students' autonomous learning. This requires LLMs to go beyond current assistance in information retrieval, actively inspiring students' curiosity and critical thinking. For example, LLM-based chatbots could act as "thinking assistants" to provide scaffolding queries for brainstorming and thought-provoking to encourage deep reflection and critical thinking (Park & Kulkarni, 2023). Indeed, educational chatbots that engage students in autonomous conversations have been proven to enable students to take ownership of their learning process and cultivate self-regulation skills (Ait et al., 2023; Al-Hafdi & AlNajdi, 2024). With enhanced capabilities in natural language understanding across different contexts and human-like text generation on various topics, LLM-based chatbots could further act as an opponent in Socratic dialogue to facilitate inquiry learning (Miao & Holmes, 2023), which promotes SDL (Sjödahl Hammarlund et al., 2013).

8.3 Designing guidelines and instructions for the use and regulation of LLMs in k-12 education

The experimental group exhibited a relatively high intention to continue using LLM-based chatbots. This highlights the necessity for both the use of these tools by students and the regulation of the tools themselves, with the aim of safeguarding students' safety, rights, and well-being.

To begin with, the different scores of subjective and objective tests between groups prompt the need for guidelines and instructions on using LLM in K-12 for teachers, staff, and even families to safeguard students' usage of LLM. Universities have published guidelines and instructions on how to use generative AI responsibly and wisely (Fujimaki, 2023; Hong Kong University of Science and Technology, 2023; Nanyang Technological University, 2023). Similar guidelines and instructions for K-12 could be proposed and released by educational institutions or governments.

In addition to advocating for regulations and guidelines regarding students' usage of LLM, we also propose regulations on data safety, privacy, and ethical issues on LLM-based education technologies. The distinctive characteristics of educational contexts entail handling sensitive data concerning students' learning preferences and behavioral tendencies (Levin, 2021). Companies that host LLM-based chatbots may collect and analyze students' data for various purposes, such as personalization of learning experiences, assessment of academic progress, and provision of feedback (Ekambaranathan et al., 2021). Without adequate regulation and guidelines, there is a risk of unauthorized access to students' data, misuse of personal information, and breaches of confidentiality, which could have serious consequences for students' well-being (Nowicki et al., 2020). To address these challenges, for example, a multi-platform architecture, EAIED, was proposed based on activity theory to promote equitable access to knowledge while preserving students' data privacy, ethics, and interoperability between learning systems in higher education. With the proliferation of digital technologies in K-12 education, similar guidance and instructions could be designed for this education period. The regulation of LLMs would empower children, parents, and teachers to be fully informed and to exercise control over their data in terms of its use in these educational tools.

During children's interaction with LLM, technology factors, such as response time and hallucination, might affect student engagement and learning with LLM-based chatbots. For example, shorter response time may increase students' satisfaction and productivity, while they may make more errors (Shneiderman, 1984). The hallucination may hinder the accuracy of context (e.g. assessment feedback, an introduction to knowledge) generated by LLMs (Ahmad et al., 2023). Although hallucination cannot be completely eliminated, designers can reduce hallucination by fine-tuning LLMs, such as structuring the input/output, providing mechanisms for user feedback, Retrieval Augmented Generation (RAG), domain-specific fine-tuning, and Reinforcement Learning with Human Feedback (RLHF) (Minaee et al., 2024). For instance, Sovrano et al. (2023) combined ChatGPT by integrating it with Achinstein's philosophical theory of explanations. The resulting model, ExplanatoryGPT, generated interactive, user-centered explanations while mitigating hallucinations and memory constraints. While hallucination may hinder knowledge comprehension, as students may receive incorrect or misleading information, it may also have some unexpected benefits on students' creativity endeavors (Jiang et al., 2024). Further study can explore how these technical features impact children's learning and further design chatbots based on students' needs.

Indeed, a child-centered approach that prioritizes children's rights, needs, and values should be the fundamental guideline for the design and development of LLM programs for middle school students. Previous educational chatbots for children often prioritize a child-friendly design to enhance the learning experience, fostering curiosity and participation, including gamification (Bergeret al., 2019), engaging visual (Gabrielli, 2020; Ruan, 2019), and anthropomorphic conversation strategies (Liu et al., 2022). For example, students in experimental groups spent most time typing questions, and copying and pasting materials (without detailed reading) when they interacted with wisdomBot. This may be attributed to children's insufficient typing skills and reading skills, compared to adults (Cain et al., 2004). Tying assistance, drag-and-drop interface, and visualization of generated content may help students to better collaborate with LLMs. Zha et al. (2024) applied a child-centered method to design LLM-based chatbots for middle school students in creative project-based learning (PBL). The results of the learning experiences of 31 middle school students indicated that the LLM-based chatbot, tailored to children-centered design considerations, was beneficial in each step (i.e., discover, define, develop, and deliver) of middle school students creative PBL. Similar approaches could be applied in the scenario of retrieving information and completing assignments.

9 Limitations and future research

Admittedly, our study has several limitations. First, the school that signed up for our research may already be more interested in the impact of LLM-based chatbots than the average population. Future exploration with schools from diverse backgrounds might offer supplementary insights that complement the ones we have discovered.

Second, although our instruments were adopted from well-established scales and we described a detailed design process of our assessment tools, our assignments, the scoring rubric, as well as pre and post-subjective tests may not be suitable as standardized measurements. Further research could explore assessment tools with better reliability and validity.

Third, as we wanted to guarantee the safety and control of students' experiences with wisdomBot, we did not engage children with long-period and unsupervised learning with wisdomBot (e.g. study alone at home). This may overlook the impact of home study on students' assignment completion. Future longitude research or in-home study could increase the depth of insights into the long-term impact and usage patterns of LLM-based chatbots in educational settings.

Finally, the impact of LLM-based chatbot-assisted assignments on students' learning is a broad topic related to a variety of different issues. In this study, we were only able to address a few of the topics relating to students' learning in middle school IT classes. Studies in different learning contexts, assignment types, and students' age groups may contribute additional findings complementary to the ones we found.

10 Conclusion

Middle school students are immersed in a learning environment where the prevalence of LLM-based chatbots and other AI tools is on the rise. As these advanced technologies continue to proliferate, students in this age group are increasingly likely to engage with these tools within the educational context. Despite the potential benefits, they may outsource their assignment to these tools, raising concerns about its potential harm. This paper is the first to provide empirical evidence on the impact of completing assignments with an LLM-based chatbot on middle school students' assignment performance, learning test scores, learning motivation, learning satisfaction, and learning experiences, as well as summarize teachers' reflections on designing assignments, assessment and related learning design.

We discovered that students who completed assignments with LLM-based chatbots exhibited higher submission rates, increased word counts, and achieved significantly elevated scores in assignments' completion, accuracy, depth, and logic. However, they were less adept at critically merging and analyzing materials and gained significantly lower scores on the knowledge test and relatively low scores on the migration test. No significant differences have been observed in learning motivation, satisfaction, enjoyment, and students' ability to migrate their knowledge. The majority of students expressed satisfaction and a willingness to continue using the tool. We also identified three key gaps in learning designs and guidance through teacher reflection and in-class observation. These gaps encompass a lack of assessment tools measuring students' thinking skills, students' deficiencies in reading and typing skills, and the need for a redefinition of human roles (teachers and learning peers).

Our findings provided insights for the future design of learning with LLM-based chatbots, including revolutionizing learning assessment with a focus on critical thinking skills, combining LLM-based chatbots with students' autonomous learning, emphasizing emotional support when students' learning with LLM, and Designing guidelines and instructions about using LLM in K-12, as well as the design of educational LLM-based chatbots. We hope that our findings will support educators who plan to integrate LLM-based chatbots in learning activities for K-12 students, and designers of educational LLM-based chatbots, as well as the policymakers who may propose K-12 generative artificial intelligence guidelines and instruction in the future.

Appendix

Full version of assignments scoring rubric.

Dimension\score	1	2	3	4	5
Completion: Assess the extent to which the task is fully completed.	Significant elements are missing or incomplete.	Several key components are incomplete.	Most elements are complete, but some are missing.	Nearly all elements are complete, with minor omissions.	The task is fully and thoroughly completed.
Creativity: Evaluate the level of creativity demonstrated in the work.	Little to no evidence of original thinking or creativity.	Limited creativity; ideas are somewhat derivative.	Some original elements, but overall lacking creativity.	Demonstrates creativity in certain aspects.	Exceptional creativity are evident throughout.
Accuracy: Examine the correctness of the content or information presented.	Numerous inaccuracies or errors in content.	Several factual inaccuracies are present.	Few inaccuracies; some minor errors.	Rare inaccuracies, with only minor errors.	All content is accurate and error-free.
Material Refinement: Assess the degree of secondary processing or refinement of materials(from wisdomBot or search engine).	Little to no evidence of secondary processing.	Minimal efforts in refining materials.	Some attempts at processing or refining materials.	Effective use of secondary processing in certain areas.	Comprehensive and skillful refinement of materials.
Depth: Assess the extent to which the content demonstrates thorough understanding.	Content lacks depth and is superficial, with little elaboration or analysis.	Limited depth, with some aspects lacking elaboration or analysis.	Moderate depth, with most aspects adequately elaborated and analyzed.	Substantial depth, with thorough elaboration and analysis of most aspects.	Exceptional depth, demonstrating comprehensive understanding and insightful analysis throughout.
Logic: Evaluate the logical flow and coherence of the work.	The work lacks a logical structure or flow.	Major inconsistencies disrupt the logical flow.	Overall logical flow, but with some disruptions.	Logical progression with minimal disruptions.	A highly logical and coherent presentation.
Page Aesthetics: Consider the visual attractiveness and formatting skills in creating the document.	The document lacks visual appeal and proper formatting.	Minimal attention to aesthetics and formatting.	Adequate visual appeal, but some formatting issues.	Well-designed with minor formatting improvements possible.	Visually appealing, with excellent formatting skills demonstrated.

Acknowledgements The authors would like to acknowledge the funding support of National Science and Technology Major Project (2022ZD0115904).

Author contributions Yumeng Zhu: Investigation, Data curation, Writing, Methodology, Visualisation. Caifeng Zhu: Investigation, Data curation, Writing. Tao Wu: Software and Writing-Reviewing. Shulei Wang: Software. Yiyun Zhou: Software. Jingyuan Chen: Writing-Reviewing and Editing. Wei Fu: Writing-Reviewing and Editing. Yan Li: Conceptualization, Investigation, Supervision, Writing- Reviewing and Editing.

Data availability Data will be made available on request.

Declarations

Submission declaration and verification This work described has not been published previously.

Competing interests The authors declare that they have no conflict of interest.

References

Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, *103*(3), 152. https://doi.org/10.3163/1536-5050.103.3.010

Adarkwah, M. A., Ying, C., Mustafa, M. Y., & Huang, R. (2023, August). Prediction of Learner Information-Seeking Behavior and Classroom Engagement in the Advent of ChatGPT. In *International Conference on Smart Learning Environments* (pp. 117–126). Singapore: Springer Nature Singapore https://doi.org/10.1007/978-981-99-5961-7_13

Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT. *Education Interactive Learning Environments*, 1–14. https://doi.org/10.1080/10494820.2023.2253858

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of *ChatGPT*. *Contemporary Educational Technology*, *15*(3), ep429. https://doi.org/10.30935/cedtech/13152

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209. https://doi.org/10.1037/edu0000282

Ahmad, Z., Kaiser, W., & Rahim, S. (2023). Hallucinations in ChatGPT: An unreliable tool for learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, *15*(4), 1–18. https://doi.org/10.21659/rupkatha.v15n4.17

Ait Baha, T., El Hajji, M., Es-Saady, Y., & Fadili, H. (2023). The impact of educational chatbot on student learning experience. *Education and Information Technologies*, 1–24. https://doi.org/10.1007/s10639-023-12166-w

Al-Hafdi, F. S., & AlNajdi, S. M. (2024). The effectiveness of using chatbot-based environment on learning process, students' performances and perceptions: A mixed exploratory study. *Education and Information Technologies*, 1–32. https://doi.org/10.1007/s10639-024-12671-6

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

Alemdag, E. (2023). The effect of chatbots on learning: A meta-analysis of empirical *research*. *Journal of Research on Technology in Education*, 1–23. https://doi.org/10.1080/15391523.2023.2255698

Ayres, P. (2006). Impact of reducing intrinsic cognitive load on learning in a mathematical domain. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *20*(3), 287–298. https://doi.org/10.1002/acp.1245

Bencherki, N., Matte, F., & Cooren, F. (2019). *Authority and power in social interaction: Methods and analysis*. Routledge.

Berger, E., Sæthre, T. H., & Divitini, M. (2019). PrivaCity: A Chatbot Game to Raise Privacy Awareness Among Teenagers. In Informatics in Schools. New Ideas in School Informatics: 12th International Conference on *Informatics in Schools*: *Situation, Evolution, and Perspectives, ISSEP 2019, Larnaca, Cyprus, November 18–20, 2019, Proceedings 12* (pp. 293–304). Springer International Publishing. https://doi.org/10.1007/978-3-030-33759-9_23

Bilgic, K., & Dogusoy, B. (2023). Exploring secondary school students' computational thinking experiences enriched with block-based programming activities: An action research. *Education and Information Technologies*, *28*(8), 10359–10384. https://doi.org/10.1007/s10639-023-11583-1

Buijs, M., & Admiraal, W. (2013). Homework assignments to enhance student engagement in secondary education. *European Journal of Psychology of Education*, *28*, 767–779. https://doi.org/10.1007/s10212-012-0139-0

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: *Concurrent prediction by working memory*, verbal ability, and component skills. *Journal of Educational Psychology*, *96*(1), 31. https://doi.org/10.1037/0022-0663.96.1.31

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

Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., & Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom). *PLoS One*, *18*(8), e0290691. https://doi.org/10.1371/journal.pone.0290691

Chi, S., Wang, Z., & Liu, X. (2023). Assessment of context-based chemistry problem-solving skills: Test design and results from ninth-grade students. *Research in Science Education*, *53*(2), 295–318. https://doi.org/10.1007/s11165-022-10056-8

Chinese Ministry of Education (2022, March 25). *Curriculum Standard for Comulsory Education Information Science and Technology (2022)*. Retrieved April 25, 2024 from: http://www.moe.gov.cn/srcsite/A26/s8001/202204/W020220420582361024968.pdf

de Winter, J. C. (2023). Can ChatGPT pass high school exams on English Language Comprehension? *International Journal of Artificial Intelligence in Education*, 1–16. https://doi.org/10.1007/s40593-023-00372-z

Ekambaranathan, A., Zhao, J., & Van Kleek, M. (2021, May). Money makes the world go around: Identifying Barriers to Better Privacy in Children's Apps From Developers' Perspectives. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–15). https://doi.org/10.1145/3411764.3445599

Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in human B*ehavior, *93*, 279–289. https://doi.org/10.1016/j.chb.2018.12.023

Fujimaki, A. (2023). (n.a.). *Regarding the Use of Generative AI*. Retrieved October 15, from https://en.nagoya-u.ac.jp/academics/ai/index.html

Fuller, A., Morbitzer, K., Zeeman, K. A., Persky, J. M. M., Savage, A. C., A., & McLaughlin, J. E. (2024). Exploring the use of ChatGPT to analyze student course evaluation comments. *BMC Medical Education*, *24*(1), 1–8. https://doi.org/10.1186/s12909-024-05316-2

Gabrielli, S., Rizzi, S., Carbone, S., & Donisi, V. (2020). A chatbot-based coaching intervention for adolescents to promote life skills: Pilot study. *JMIR Human Factors*, *7*(1), e16762. https://humanfactors.jmir.org/2020/1/e16762

Gan, W., Qi, Z., Wu, J., & Lin, J. C. W. (2023, December). Large language models in education: Vision and opportunities. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 4776–4785). IEEE.

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences*, *13*(7), 692. https://doi.org/10.3390/educsci13070692

Grogan, K. A. (2017). Will this be on the test? How exam structure affects perceptions of innovative assignments in a masters of science microeconomics course. *International Review of Economics Education*, *26*, 1–8. https://doi.org/10.1016/j.iree.2017.06.001

Guo, Y., & Lee, D. (2023). Leveraging chatgpt for enhancing critical thinking skills. *Journal of Chemical Education*, *100*(12), 4876–4883. https://doi.org/10.1021/acs.jchemed.3c00505

Guo, K., Zhong, Y., Li, D., & Chu, S. K. W. (2023). Effects of chatbot-assisted in-class debates on students' argumentation skills and task *motivation*. *Computers & Education*, *203*, 104862. https://doi.org/10.1016/j.compedu.2023.104862

Henthorn, R., Lowden, K., & McArdle, K. (2024). It gives meaning and purpose to what you do': Mentors' interpretations of practitioner action research in education. *Educational Action Research*, *32*(2), 169–185. https://doi.org/10.1080/09650792.2022.2106260

Ho, D. Y. F., Peng, S. Q., & Chan, S. F. F. (2002). Authority and learning in Confucian-heritage education: A relational methodological analysis. In F. Salili, Y. Y. Hong, & C. Y. Chiu, (Eds.), *Multiple competencies and self-regulated learning: Implications for multicultural education* (pp. 29–47). Greenwich, CT: Information Age Publishing.

Holmes, W., & Miao, F. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing.

Hong Kong University of Science and Technology. (n.a.). *AI Literacy for End-Users: Use AI Wisely*. Retrieved October 15 (2023). from https://libguides.hkust.edu.hk/ai-literacy

Hübscher-Younger, T., & Narayanan, N. H. (2003). Designing for divergence. In *Designing for change in networked learning environments: Proceedings of the international conference on computer support for collaborative learning 2003* (pp. 461–470). Dordrecht: Springer Netherlands.

Jauhiainen, J. S., & Guerra, A. G. (2023). Generative AI and ChatGPT in School Children's education: Evidence from a school lesson. *Sustainability*, *15*(18), 14025. https://doi.org/10.3390/su151814025

Jiang, X., Tian, Y., Hua, F., Xu, C., Wang, Y., & Guo, J. (2024). A Survey on Large Language Model Hallucination via a Creativity Perspective. *arXiv preprint arXiv:2402.06647*. https://doi.org/10.48550/arXiv.2402.06647

Jishnu, D., Srinivasan, M., Dhanunjay, G. S., & Shamala, R. (2023). Unveiling student motivations: A study of ChatGPT usage in education. *ShodhKosh: Journal of Visual and Performing Arts, 4*(2), 65–73. https://doi.org/10.29121/shodhkosh.v4.i2.2023.503.

Jungherr, A. (2023). Using ChatGPT and other large language model (LLM) applications for academic paper assignments. SocArXiv. https://doi.org/10.31235/osf.io/d84q6

Kaplan-Rakowski, R., Grotewold, K., Hartwick, P., & Papin, K. (2023). Generative AI and teachers' perspectives on its implementation in Education. *Journal of Interactive Learning Research*, *34*(2), 313–338. https://doi.org/10.21275/SR23219122412

Karthikeyan, C. (2023). Literature Review on pros and cons of ChatGPT implications in Education. *International Journal of Science and Research (IJSR)*, *12*(3), 283–291.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., & Dementieva, D. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274. Fischer, F.Kasneci, G.

Kumar, N. A., & Lan, A. (2024). Using large Language models for Student-Code guided Test Case Generation in Computer Science Education. *arXiv Preprint arXiv:2402 07081*. https://doi.org/10.48550/arXiv.2402.07081

Lan, Y. J., & Chen, N. S. (2024). Teachers' agency in the era of LLM and generative AI. *Educational Technology & Society*, *27*(1), I–XVIII. https://www.jstor.org/stable/48754837

Lee, H. (2023). The rise of ChatGPT: Exploring its potential in medical education. *Anatomical Sciences Education*. https://doi.org/10.1002/ase.2270

Lee, H. Y., Chen, P. H., Wang, W. S., Huang, Y. M., & Wu, T. T. (2024). Empowering ChatGPT with guidance mechanism in blended learning: Effect of self-regulated learning, higher-order thinking skills, and knowledge construction. *International Journal of Educational Technology in Higher Education*, *21*(1), 1–28. https://doi.org/10.1186/s41239-024-00447-4

Lee, I., Ali, S., Zhang, H., DiPaola, D., & Breazeal, C. (2021, March). Developing middle school students' AI literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 191–197). https://doi.org/10.1145/3408877.3432513

Levin, D. A. (2021). The state of K-12 cybersecurity: 2020 year in review. *K-12 cybersecurity resource center*. Retrieved April 20, 2024, from https://www.k12six.org/the-report

Li, J., & Xue, E. (2021). *Compulsory Education Policy in China: Concept and Practice*. Springer Nature. https://doi.org/10.1007/978-981-33-6358-8

Li, Y., Chen, J., Zhou, H., Yuan, H., & Yang, R. (2023). Research on motivation and behavior of ChatGPT use in middle school students. *International Journal of New Developments in Education*, *5*(19), 69–77. https://doi.org/10.25236/IJNDE.2023.051911

Lin, K. M. (2011). e-Learning continuance intention: Moderating effects of user *e-learning experience*. *Computers & Education*, *56*(2), 515–526. https://doi.org/10.1016/j.compedu.2010.09.017

Liu, C. C., Liao, M. G., & Chang, C. H. (2022). An analysis of children's interaction with an AI chatbot and its impact on their interest in reading. *Computers & Education*, *189*, 104576. https://doi.org/10.1016/j.compedu.2022.104576. Lin, H. M.

Liu, M., Ren, Y., Nyagoga, L. M., Stonier, F., Wu, Z., & Yu, L. (2023). Future of education in the era of generative artificial intelligence: Consensus among *Chinese scholars* on applications of ChatGPT in schools. *Future in Educational Research*, *1*(1), 72–101. https://doi.org/10.1002/fer3.10

Mahajan, V. (2023, October 13). *100+Incredible ChatGPT Statistics & Facts in 2023*. Retrieved November 12, 2023, from https://www.notta.ai/en/blog/chatgpt-statistics

Malik, A., Khan, M. L., & Hussain, K. (2023). How is ChatGPT transforming academia? Examining its impact on teaching, research, assessment, and learning. *Examining its Impact on Teaching, Research, Assessment, and Learning (April 9, 2023)*. https://doi.org/10.2139/ssrn.4413516

Masouleh, N. S., & Jooneghani, R. B. (2012). Autonomous learning: A teacher-less learning! *Procedia-Social and Behavioral Sciences*, *55*, 835–842. https://doi.org/10.1016/j.sbspro.2012.09.570

Meyer, J. G., Urbanowicz, R. J., & Martin, P. C. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *Big Data Mining*, *16*(1), 20. https://doi.org/10.1186/s13040-023-00339-9. O'Connor, K.Li, R., Peng, P. C., … Moore, J. H.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv Preprint arXiv:2402 06196*. https://doi.org/10.48550/arXiv.2402.06196

Moritz, S., Romeike, B., Stosch, C., & Tolks, D. (2023). Generative AI (gAI) in medical education: ChatGPT and co. *GMS Journal for Medical Education*, *40*(4). https://doi.org/10.3205/zma001636

Morrison, R., & Barton, G. (2018). Search engine use as a literacy in the middle years: The need for explicit instruction and active learners. *Literacy Learning: The Middle Years*, *26*(3), 37–47. https://doi.org/10.3316/informit.846488641829827

Nanyang Technological University. (n.a.). *NTU Position on the Use of Generative Artificial Intelligence in Research* Retrieved October 15 (2023). from https://www.ntu.edu.sg/research/resources/use-of-gai-in-research

Nee, C. K., Rahman, M. H. A., Yahaya, N., Ibrahim, N. H., Razak, R. A., & Sugino, C. (2023). Exploring the Trend and potential distribution of Chatbot in Education: A systematic review. *International Journal of Information and Education Technology*, *13*(3), 516–525. https://doi.org/10.18178/ijiet.2023.13.3.1834

Nowicki, J. M. (2020). *Data Security: Recent K-12 Data Breaches Show That Students Are Vulnerable to Harm*. Report to the Republican Leader, Committee on Education and Labor, House of Representatives. GAO-20-644. US Government Accountability Office. Retrieved April 21, 2024, from https://files.eric.ed.gov/fulltext/ED609671.pdf

Ofcom (2023, November 28). *Gen Z driving early adoption of Gen AI, our latest research shows* Retrieved November 29, 2023, from https://www.ofcom.org.uk/news-centre/2023/gen-z-driving-early-adoption-of-gen-ai

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, *2*, 100033. https://doi.org/10.1016/j.caeai.2021.100033

Olney, A. M. (2023, June). Generating multiple choice questions from a textbook: LLMs match human performance on most metrics. In *AIED Workshops*. https://ceur-ws.org/Vol-3487/paper7.pdf

Oxford English Dictionary (2023, July). s.v. *chatbot, n* Retrieved November 20, 2023, from https://doi.org/10.1093/OED/2981785869

Özer, B., Duran, V., & Tekke, M. (2020). Training of trainers: An action-based research for improving the Pedagogical skills of academicians. *International Journal of Evaluation and Research in Education*, *9*(3), 704–715. https://doi.org/10.11591/ijere.v9i3.20327

Park, S. (2023). C. Kulkarni (Ed.), Thinking assistants: LLM-Based conversational assistants that help users think by asking rather than answering. *arXiv Preprint arXiv:2312 06024*https://doi.org/10.48550/arXiv.2312.06024

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A., Kamp, E. T., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Perkins, M. (2023). Academic Integrity considerations of AI large Language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, *20*(2). https://doi.org/10.53761/1.20.02.07

Playfoot, D., Quigley, M., & Thomas, A. G. (2024). Hey ChatGPT, give me a title for a paper about degree apathy and student use of AI for assignment writing. *The Internet and Higher Education*, 100950. https://doi.org/10.1016/j.iheduc.2024.100950

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and benefits of large language models for the environment. *Environmental Science & Technology*, *57*(9), 3464–3466. https://doi.org/10.1021/acs.est.3c01106

Ruan, S., Willis, A., Xu, Q., Davis, G. M., & Jiang, L. (2019, June). Brunskill, E., & Landay, J. A. Bookbuddy: Turning digital materials into interactive foreign language lessons through a voice chatbot. In *Proceedings of the sixth (2019) ACM conference on learning@ scale* (pp.1–4). https://doi.org/10.1145/3330430.3333643

Ruzek, E. A., Hafen, C. A., Allen, J. P., Gregory, A., Mikami, A. Y., & Pianta, R. C. (2016). How *teacher emotional support* motivates students: The mediating roles of perceived peer relatedness, autonomy support, and competence. *Learning and Instruction*, *42*, 95–103. https://doi.org/10.1016/j.learninstruc.2016.01.004

Shneiderman, B. (1984). Response time and display rate in human performance with computers. *ACM Computing Surveys (CSUR)*, *16*(3), 265–285.

Shoufan, A. (2023). Exploring students' perceptions of ChatGPT: Thematic analysis and follow-up survey. *IEEE Access, 11*, 38805–38818. https://doi.org/10.1109/ACCESS.2023.3268224.

Sjödahl Hammarlund, C., Nordmark, E., & Gummesson, C. (2013). Integrating theory and practice by self-directed inquiry-based learning? A pilot study. *The European Journal of Physiotherapy*, *15*(4), 225–230. https://doi.org/10.3109/21679169.2013.836565

Sovrano, F., Ashley, K., & Bacchelli, A. (2023, July). Toward eliminating hallucinations: Gpt-based explanatory AI for intelligent textbooks and documentation. In *CEUR Workshop Proceedings* (pp. 54–65). CEUR-WS. https://ceur-ws.org/Vol-3444/itb23_s3p2.pdf

Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*, 1–14. https://doi.org/10.1080/10494820.2023.2209881

Taasoobshirazi, G., & Carr, M. (2008). A review and critique of context-based physics instruction and assessment. *Educational Research Review*, *3*(2), 155–167. https://doi.org/10.1016/j.edurev.2008.01.002

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*. https://doi.org/10.48550/arXiv.2103.14411

Tanwar, H., Shrivastva, K., Singh, R., & Kumar, D. (2024). OpineBot: Class Feedback Reimagined Using a Conversational LLM. *arXiv preprint arXiv:2401.15589*. https://doi.org/10.48550/arXiv.2401.15589

The University of North Carolina at Chapel Hill (n.a.) *Teaching About The Use Of Generative AI Guidance For Instructors* Retrieved October 19 (2023). from: https://provost.unc.edu/teaching-generative-ai-guidance/

Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., & Telari, A. (2023, November). … Ognibene, D. Learning to Prompt in the Classroom to Understand AI Limits: A pilot study. In *International Conference of the Italian Association for Artificial Intelligence* (pp. 481–496). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47546-7_33

University of OXFORD (2023, 6 February). *Unauthorised use of AI in exams and assessment* Retrieved October 15, 2023, from https://academic.admin.ox.ac.uk/article/unauthorised-use-of-ai-in-exams-and-assessment

van Duijn, M. J., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M. R., & van der Putten, P. (2023). Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7–10 on advanced tests. *arXiv preprint arXiv:2310.20320*. https://doi.org/10.48550/arXiv.2310.20320

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*, 147–177. https://doi.org/10.1007/s10648-005-3951-0

Waltzer, T., Cox, R. L., & Heyman, G. D. (2023). Testing the ability of teachers and students to Differentiate between essays generated by ChatGPT and High School Students. *Human Behavior and Emerging Technologies*, *1923981*, 1–9. https://doi.org/10.1155/2023/1923981

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity*, *51*, 101440. https://doi.org/10.1016/j.tsc.2023.101440

Wei, T. (2022). An interpretation of the revised compulsory education curriculum program and standards: A revolution in China's compulsory education. *Science Insights Education Frontiers*, *13*(1), 1845–1853. https://doi.org/10.15354/sief.22.re065

Welding, L. (2023, March 17). *Half of College Students Say Using AI on Schoolwork Is Cheating or Plagiarism*. Retrieved October 13, 2023, from https://www.bestcolleges.com/research/college-students-ai-tools-survey/

Wentzel, K. R., Russell, S., & Baker, S. (2016). Emotional support and expectations from parents, teachers, and peers predict adolescent competence at school. *Journal of Educational Psychology*, *108*(2), 242. https://doi.org/10.1037/edu0000049

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, *4*, 654924. https://doi.org/10.3389/frai.2021.654924

Wu, R., & Yu, Z. (2024). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology*, *55*(1), 10–33. https://doi.org/10.1111/bjet.13334

Wu, T. T., Lee, H. Y., Li, P. H., Huang, C. N., & Huang, Y. M. (2024). Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning aid. *Journal of Educational Computing Research*, *61*(8), 3–31. https://doi.org/10.1177/07356331231191125

Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023, July). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 610–625). https://aclanthology.org/2023.bea-1.52

Yan, S. (2012). Teachers' roles in autonomous learning. *Journal of Sociological Research*, *3*(2), 557–562. https://doi.org/10.5296/jsr.v3i2.2860

Yang, R. Research on motivation and behavior of ChatGPT use in middle school students. *International Journal of New Developments in Education* *5*(19): 69–77. https://doi.org/10.25236/IJNDE.2023.051911

You, Y., Kou, Y., Ding, X., & Gui, X. (2021, May). The medical authority of AI: A study of AI-enabled consumer-facing health technology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). https://doi.org/10.1145/3411764.3445657

Youth League (2022, November). *2021 National Research Report on Internet Use by Minors*. Retrieved November 1, 2023, from: https://www.cagd.gov.cn/data/uploads//ueditor/php/upload/file/2022/11/1669791740317797.pdf

Yu, H. (2023). Reflection on whether Chat *GPT should be* banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, *14*, 1181712. https://doi.org/10.3389/fpsyg.2023.1181712

Zha, S., Qiao, Y., Hu, Q., Li, Z., Gong, J., & Xu, Y. (2024). Designing Child-Centric AI Learning Environments: Insights from LLM-Enhanced Creative Project-Based Learning. *arXiv preprint arXiv:2403.16159*. https://doi.org/10.48550/arXiv.2403.16159

Zhai, X. (2022). ChatGPT user experience: Implications for education. *Available at SSRN 4312418*. https://doi.org/10.2139/ssrn.4312418

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*https://doi.org/10.48550/arXiv.2303.18223

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yumeng Zhu[1] · **Caifeng Zhu**[2] · **Tao Wu**[3] · **Shulei Wang**[4] · **Yiyun Zhou**[4] · **Jingyuan Chen**[1] · **Fei Wu**[3] · **Yan Li**[1]

✉ Yan Li
yanli@zju.edu.cn

[1] College of Education, Zhejiang University (Zijingang Campus), Yuhangtang Rd. #866, Hangzhou, Zhejiang 310058, China

[2] Zhejiang Hangzhou Chu Kochen Honors School, HuiJu Rd.#376, Hangzhou, Zhejiang 310053, China

[3] College of Computer Science and Technology, Zhejiang University (Yuquan Campus), Zheda Rd. #38, Hangzhou, Zhejiang 310027, China

[4] School of Software Technology, Zhejiang University (Yuquan Campus), Zheda Rd. #38, Hangzhou, Zhejiang 310027, China