

# Current Trends and Future Prospects of Large-Scale Foundation Model in K-12 Education

Qiannan Zhu<sup>a,b,†</sup>, Mei Wang<sup>a,b,†</sup>, Ting Zhang<sup>a,b,†</sup>, Hua Huang<sup>a,b</sup>

<sup>a</sup> School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China

<sup>b</sup> Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing 100816, China

© Higher Education Press 2025

**Abstract** The rapid advancement of artificial intelligence has significantly impacted education, with large-scale foundation models (LFMs) emerging as transformative tools. While LFMs have demonstrated exceptional performance across diverse domains, their integration into K-12 education remains in its early stages, requiring alignment with pedagogical principles, cognitive development, and curriculum standards. This paper provides a comprehensive technological review of LFM applications in K-12 education, examining current workflows, challenges, and future opportunities. We explore how LFMs facilitate personalized learning, teacher–student collaboration, and automated assessment while highlighting critical issues such as motivation, engagement, and age-appropriate instructional strategies. By analyzing global developments, this study offers valuable insights for educators seeking to optimize AI-driven teaching methods and for students leveraging AI for self-directed learning. Our findings aim to inform future research and drive innovation in educational AI, ensuring the effective and ethical integration of LFMs into the evolving K-12 educational landscape.

**Keywords** large-scale foundation models (LFMs), multimodal large language model, large language model, K-12 education

## 1 Introduction

*Technology can become the “wings” that will allow the*

Received December 26, 2024; revised February 19, 2025; accepted March 13, 2025

Hua Huang (✉)

E-mail: [huanghua@bnu.edu.cn](mailto:huanghua@bnu.edu.cn)

†These authors contributed equally to this work and should be considered co-first authors.

*educational world to fly farther and faster than ever before—if we allow it.*

— Jenny Arledge

Over the past few decades, education has undergone a profound transformation driven by the advancement of technology, resulting in a significant shift in teaching and learning paradigm (Batool et al., 2023; Chen et al., 2020; Chiu et al., 2023; Latif et al., 2023; Li et al., 2024; Maghsudi et al., 2021). The journey began with the electrification phase (Sovacool & Ryan, 2016), which introduced audio-visual tools to enhance teaching, marking the dawn of modern educational technology. Following this, the integration of computers and multimedia further revolutionized education (Romero & Ventura, 2013), enabling the use of digital resources, online learning, and resource sharing. Entering the 2010s, artificial intelligence (AI) began to capture significant attention and was increasingly applied in various educational (Alam, 2023; Ercikan & McCaffrey, 2022; Tapalova & Zhiyenbayeva, 2022). The history of AI in education dates back several decades. In the 1950s, the first intelligent tutoring system (ITS), known as “SAINT,” was developed (Alkhatlan & Kalita, 2018). Later, in the 2000s, AI gained prominence in educational games and simulations, exemplified by the defeat of world chess champion Garry Kasparov by IBM’s Deep Blue. With the recent widespread availability of generative AI tools, AI has gradually permeated the education sector (Denny et al., 2024; Romero & Ventura, 2020; Xiong et al., 2024). Figure 1 illustrates the evolution of educational technology, highlighting its clear progression through five distinct stages.

The integration of AI into education has long been a topic of speculation and anticipation. However, it was the launch of ChatGPT (Shafik, 2024) by OpenAI that brought the world a significant step closer to realizing the transformative potential of AI technology.

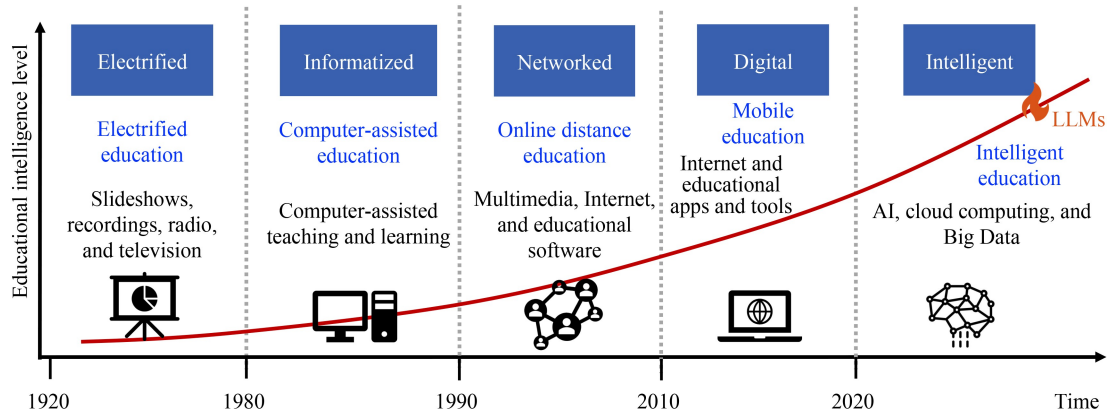


Figure 1 Evolution of educational technology progression.

ChatGPT's introduction was widely recognized, with Science (2022) naming it one of the top ten scientific breakthroughs of the year. The technology behind lie large-scale foundation models (LFMs), particularly large language models (LLMs). These models have dramatically reduced the labor time required in various fields, serving as prime examples of the emergence of new productive forces. Their impact extends far beyond the realm of education, influencing industries such as healthcare, media, and transportation. For instance, in natural language processing (NLP), LLMs have shown exceptional capabilities in tasks requiring reasoning and reading comprehension (Bhagavatula et al., 2019; Hendrycks et al., 2020; Rajpurkar et al., 2018). In protein structure prediction, AI has also outperformed human experts in CASP evaluations (Zheng et al., 2023). In the legal field, GPT-4 (OpenAI, 2023) achieved an impressive score of 297 on the Uniform Bar Examination, surpassing the high threshold of 273 set by Arizona. Furthermore, in programming competitions, OpenAI's reasoning model, o1 (Zhong et al., 2024), ranked among the top 500 in the American Invitational Mathematics Examination qualification round and outperformed human doctoral-level experts in benchmark tests for physics, biology, and chemistry. Such applications have become increasingly widespread, impacting various industries such as healthcare, media, transportation, and of course education.

The education field, with its wealth of data resources, growing demand for personalized learning, and substantial social value, has become an ideal landscape for the application of LLM technology. As a result, numerous online education companies and educational institutions have actively invested efforts aimed at creating LFMs specifically designed to address the unique challenges and requirements of the education domain. For example, iFLYTEK's Spark LFM (iFLYTEK, 2024) demonstrates its versatility in education by offering multimodal interaction, coding,

text generation, mathematical problem-solving, and knowledge-based question-answering capabilities. Another notable example is EduChat (Dan et al., 2023), which integrates Socratic teaching methods during its instruction fine-tuning phase. Furthermore, TAL's MathGPT model (Liu, 2025), with a parameter scale of 100 billion, is capable of processing formula image inputs and providing both solutions and detailed explanations, demonstrating its potential to assist students in complex mathematical learning.

Research and empirical evidence suggest that LLMs not only perform exceptionally well on benchmark tests across various disciplines (OpenAI, 2023), but also provide valuable support and enhancement in diverse educational scenarios. LLMs can function as writing and reading assistants, helping students improve both their language expression and comprehension skills (Susnjak & McIntosh, 2024). They also play a crucial role in teacher-student collaboration, personalized learning, and automated assessment (Tan et al., 2023). Furthermore, LLMs are adept at handling tasks such as responding to student inquiries (Lazaridou et al., 2022) and assisting with learning planning (Chen et al., 2023b; Wang et al., 2024a), thereby supporting students' learning processes on a broader scale. These models are fundamentally transforming the education landscape, empowering all levels of the educational process toward a new era of intelligence and personalization.

Despite huge potential, the application of LLMs in education remains in its early stages, with much of their transformative power yet to be fully realized. This is particularly true in K-12 education, where students' learning needs are highly personalized, characterized by its emphasis on holistic student development, including cognitive growth, social-emotional learning, and literacy acquisition. Specifically, there are several fundamental principles in K-12 education, demanding a distinct approach to the integration of LLMs.

First, cognitive development varies significantly

across age groups, requiring instructional materials to be developmentally appropriate. Vygotsky (1978)'s zone of proximal development suggests that optimal learning occurs when students engage with tasks that are slightly beyond their independent ability but achievable with guidance. Piaget (1952)'s stages of cognitive development highlight age-specific learning needs: Younger children in the preoperational stage (ages 2–7) benefit from hands-on, visually supported instruction, while older students in the formal operational stage (ages 11+) require opportunities for abstract reasoning and problem-solving. Effective AI-assisted learning must scaffold student understanding in alignment with these developmental principles.

Second, literacy development is central to K-12 education, serving as the foundation for all other learning domains. According to stages of reading development of Chall (1983), literacy acquisition follows a progression from phonemic awareness in early grades to reading fluency and comprehension in later years. LLMs have the potential to support this process by offering personalized reading interventions, adaptive writing feedback, and automated text simplification to align with different literacy levels (Wolf, 2007). However, without careful calibration, these models may reinforce surface-level comprehension rather than fostering deep engagement with texts, highlighting the need for structured integration with established literacy frameworks.

Third, K-12 education prioritizes real-world learning experiences that cultivate problem-solving skills, creativity, and social awareness. Constructivist theories of learning, particularly experiential learning model of Dewey (1986), stress that meaningful learning occurs through engagement with authentic, real-world problems rather than passive information absorption. In this context, LLMs should not merely function as knowledge repositories but actively facilitate problem-based learning, inquiry-based instruction, and project-based learning (Hmelo-Silver, 2004). For instance, an AI-powered learning platform could simulate historical events, guide students through scientific investigations, or generate context-based math problems that reflect real-world applications.

Forth, student engagement is a key determinant of learning success. Self-determination theory highlights the importance of autonomy, competence, and relatedness in fostering intrinsic motivation (Ryan & Deci, 2000). LLMs can enhance motivation by providing adaptive learning paths, gamified educational experiences, and personalized feedback mechanisms. However, an over-reliance on AI-generated assistance may reduce students' opportunities for productive struggle—a critical component of deep learning (Kapur, 2008). A balanced approach should ensure that AI tools complement rather than supplant student-led inquiry

and critical thinking. The convergence of large foundational models and K-12 education is an inevitable trend. Given this trajectory, it is crucial to address the question: How can LLMs be effectively integrated into K-12 education? To move forward, it is essential to clarify the key issues surrounding the promotion of these technologies within the K-12 educational context.

We observe that no existing literature has systematically summarized the application of LLMs in K-12 education from a technological perspective. To bridge this gap, this paper aims to provide a comprehensive technological review of LLMs in K-12 education, offering an in-depth analysis of the current state and emerging trends within the global K-12 LLM-supported educational ecosystem. This review aims to offer valuable insights for educators on effective teaching strategies and for students on how to enhance their self-directed learning practices. The cases discussed in this paper represent a selection of the most relevant cases collected to date, providing a foundational understanding of the landscape of LLM applications in K-12 education. We hope that this work will stimulate further innovation and foster a deeper understanding of the evolving role of LLMs in shaping the future of K-12 education.

The paper is structured as follows. Section 2 provides a detailed introduction to the emergence and current state of LLMs for education. Section 3 analyzes the application of these models in the K-12 education sector through a survey of current practices, focusing on typical scenarios. Furthermore, Section 4 highlights the current challenges and explores future potential opportunities. Section 5 presents the conclusions.

## 2 LFM for Education

### 2.1 | Large-Scale Foundation Model

LLMs are the typical LFM, trained on massive datasets with scaled model sizes, have ushered in a new era of possibilities in AI. Through pre-training with large-scale unlabelled data, instruction tuning (Chung et al., 2024) and key technologies for reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), LLMs have acquired rich knowledge of the world (Brown et al., 2020; Zhao et al., 2023), have better text comprehension and generation capabilities, can solve a variety of complex tasks effectively, and can excel at accurately tracking context in multi-turn dialogues, ensuring alignment with human values for safe and responsible use (Bubeck et al., 2023; Wei, 2022a).

The most powerful LLMs are the transformer architecture (Devlin et al., 2019; Vaswani et al., 2017),

which use stacked self-attention in encoder and cross-attention in decoder within an encoder–decoder structure. The encoder transforms an input sequence of symbols  $x = (x_1, x_2, \dots, x_n)$  into a sequence of continuous representations  $y = (y_1, y_2, \dots, y_n)$ , which the decoder then uses to generate the output sequence of symbols (i.e., human-like text), producing one element at a time. ChatGPT (Shafik, 2024) developed by OpenAI has gained widespread attention in society since its release. ChatGPT is the typical decoder-only transformer model where GPT-series (Brown et al., 2020; Radford et al., 2019 & 2020) are decoder to accurately predict the next word. It is the conversation model that can understand and generate human-like text. After that, GPT-4 and GPT-4o1 (OpenAI, 2023) were released, introducing multimodal input capabilities and demonstrating significant performance improvements across various evaluation tasks compared to ChatGPT. With ChatGPT, GPT-4 and GPT-4o as milestones of language models, the AI generated content era began (Bubeck et al., 2023).

During the development process of LLMs, several key techniques drive the success of LLMs. First, it is challenging to train LLMs due to their massive size, where distributed training algorithms with various parallel strategies, such as DeepSpeed (Rasley et al., 2020) and Megatron-LM (Shoeybi et al., 2019) are proposed to learn the parameters of LLMs. In addition, the optimization frameworks with optimization tricks, i.e., restart to overcome training loss spike (Chowdhery et al., 2023) and mixed precision training (Le Scao et al., 2023) for training stability emerged. Second, LLMs with general-purpose problem-solving abilities after pre-training on large-scale corpora, may not always be immediately evident when performing specific tasks. Ability elicitation through the design of appropriate task instructions or specific in-context learning strategies to elicit such abilities is necessary. For example, it is useful to approach complex reasoning tasks through chain-of-thought (CoT) prompting by incorporating intermediate reasoning steps. Third, alignment tuning is proposed to align LLMs with human values. This ensures they produce high-quality, harmless responses, avoiding toxic, biased, or harmful content for humans. InstructGPT (Ouyang et al., 2022) introduced RLHF technique to generate helpful, honest, and harmless context. Moreover, there are various technical advancements have also played a significant role in the success of LLMs.

Recently, a large number of LLMs have emerged to advance the capabilities of artificial intelligence in areas such as NLP (Touvron et al., 2023a & 2023b), computer vision (Awadalla et al., 2023; Li et al., 2023b; Zhu et al., 2023a), and multimodal (Radford et al., 2021; Wang et al., 2022a). These advanced LLMs can be broadly classified into LLMs and

multimodal LLMs (MLLMs), where LLMs have shown surprising performance on NLP tasks and MLLMs have the ability to receive and reason on multimodal tasks. The LLaMA series from LLaMA, LLaMA-2 (Touvron et al., 2023a & 2023b), LLaMA-3 to LLaMA-3.1 (AI Meta, 2024a & 2024b), available as open-source LLMs pre-trained on 15 trillion tokens, deliver competitive performance compared to leading closed-source models like GPT-4. The Mistral series (Jiang et al., 2023 & 2024), including Mistral (7B), Mistral NeMo (12B), Mistral Large 2 (123B), and Mixtral, delivers strong performance on benchmarks like MMLU and GSM8k with Mistral Large 2’s support for 11 languages and over 80 programming languages. The Gemma series includes lightweight and high-performing open models, such as Gemma-1 (2B and 7B) and Gemma-2 (2B, 9B, and 27B) (Gemma Team, 2024a & 2024b), trained on up to 13 trillion English tokens, where Gemma-2 demonstrates exceptional performance across benchmarks like ARC-c, MMLU, and GSM8k. The Qwen series (Qwen Team, 2024; Yang et al., 2024) is an open source collection of large-scale models supporting both English and Chinese, including Qwen (7B to 72B), Qwen1.5 (0.5B to 110B), Qwen2 (0.5B to 72B), and the latest Qwen2.5 (0.5B to 72B), pre-trained on up to 18 trillion tokens, demonstrating significant improvements in knowledge retention, coding, mathematics and instructional follow-up. Moreover, Baichuan (Yang et al., 2023a) and GLM (GLM Team, 2024) are open-source bilingual LLMs supporting both Chinese and English, excelling in various evaluations across semantics, mathematics, reasoning, coding, and knowledge domains. Despite their impressive performance on various NLP tasks, LLMs are constrained to processing discrete text and lack the ability to understand multimodal input and produce multimodal output, such as images, videos, and audio. In contrast, MLLMs are designed to achieve this capability. Generally, MLLMs consist of a pre-trained modality encoder that compresses modalities such as images or audio into a more compact representation, a pre-trained LLM, a modality interface that matches different modalities, and a modality generator for generating other modalities expected for text (Yin et al., 2023a). In the field of MLLMs, CLIP (Radford et al., 2021) integrates a visual encoder that aligns semantically with text by leveraging large-scale pretraining on image-text pairs. Some works like MiniGPT-4 (Zhu et al., 2023a) adopts EVA-CLIP (Fang et al., 2023; Sun et al., 2023) as an encoder, Osprey (Yuan et al., 2024) adopts a convolution-based ConvNext-L (Cherti et al., 2023) as encoder, and Qwen-vl (Bai et al., 2023) adopts the aligned image-caption-box visual receptor as an encoder to support flexible image resolution and parameter size input. Encoders for other modalities are also available (Deshmukh et al.,

2023; Elizalde et al., 2023; Han et al., 2023; Lin et al., 2023). For example, DouBao (Bai et al., 2024) employs representation learning-generator-renderer pipeline to encode multi-modal inputs, including lyrics, style descriptions, audio references, and voice prompts for making vocal music generation.

Bridging the gap between different modalities in large multimodal models is necessary but highly resource-intensive. A more practical approach is to design the modality interface as a learnable connector between the encoder and the LLM, where the module projects multimodal information into the space that the LLM can understand efficiently. To make LLM understandable efficiently, the token-level and feature-level fusion of multimodal information are embedded in the modality interface. For token-level fusion, the features generated by encoders are converted into tokens and combined with text tokens, which are then fed into the LLMs for further processing. For example, the BLIP-2 series (Chen et al., 2023a; Dai et al., 2023; Li et al., 2023a; Zhang et al., 2023a) are typical approaches that introduce learnable query tokens and Q-Former network to extract information from multimodal input. In addition, some methods such as LLaVA (Liu et al., 2023 & 2024) use an MLP-based interface to align features between multimodal information (Pi et al., 2023; Su et al., 2023; Zhang, 2023c). The feature-level fusion based approaches mainly focus on deep interaction and fusion between multimodal features (Guo et al., 2023; Wang et al., 2023; Yin et al., 2023b; Zhang et al., 2023b). For example, Flamingo (Alayrac et al., 2022) integrates external visual cues into language features by adding cross-attention layers within the frozen Transformer layers of LLMs. Among these LLMs, as data sets become larger and model architectures more sophisticated, LLMs have developed remarkable emergent capabilities, including instruction following (Chung et al., 2024; Wei et al., 2021), in-context learning (Liu et al., 2021; Rubin et al., 2021), CoT (Fu et al., 2022; Zhang et al., 2022) and planning (Chen et al., 2023d; Yao et al., 2023), which enable them to perform complex tasks with increasing skill and adaptability.

Research on LFM is not only hot in academia, but industry has also conducted research on practical applications and released a variety of products in real-life scenarios. For example, Moonshot AI has developed Kimi 3, which can handle long conversations with a capacity of up to 200,000 characters in input and output, and can also receive documents to chat with users. OpenAI enabled Sora to generate new videos based on text, image, and video inputs. Qwen is Alibaba Group's LLM, works as an assistant that can do scientific answering, writing, creative work, coding, etc. Today, more and more LLM products can parse web pages and URLs to capture up-to-date information for

providing users with the information they need. However, current LLMs are limited in processing multimodal information with long context, weak ability of agents to interact with the real world, being tricked into giving biased or undesirable responses, etc.

## 2.2 | LFM of Education

The ability of LLMs to explain, generalize, and create has attracted considerable interest and led to extensive research and discussion in domain applications such as education (Kasneci et al., 2023; Li et al., 2023c), healthcare (Thirunavukarasu et al., 2023; Yang et al., 2023c), and finance (Wu et al., 2023a; Yang et al., 2023b). Focusing on education, the LLMs can make potential in teacher-student collaboration, personalized learning, and assessment automation (Wang et al., 2024b).

The research and development of educational LLMs has been given great attention, many research institutions and enterprises have released their educational LLMs for enhancing educational quality. For example, iFLYTEK Spark helps students with various tasks such as grading homework, practicing spoken language, answering encyclopedic questions, and writing code. MathGPT is specialized in mathematical LLM, which can do mathematical calculation and reasoning, code understanding and generation, and math knowledge question answering. Ziyue is a specialized English LLM that supports speaking coach, AI writing, and AI translation. OpenAI o1 can reason through complex tasks by spending more time thinking before responding, achieving a higher ability to solve harder problems than previous models in science, coding, and math. Developing educational LLMs are an effective way to support digital transformation in education. More educational LLM are summarized into Table 1.

## 3 Current Trends

This paper focuses on K-12 Education and explores how LLMs can enhance educational quality. The application of LLMs in K-12 education requires tailored approaches, such as providing guided responses, staying within students' knowledge, and adhering to curriculum standards. These applications should encourage critical thinking, match students' cognitive levels, and avoid overly complex or irrelevant information to ensure effective and reliable support for learning. Therefore, the successful integration of LLMs into K-12 education requires a comprehensive understanding of students' learning characteristics, cognitive development, and curriculum requirements.

**Table 1** Educational LLMs for typical applications scenario

Educational LLMs	Typical applications cenario
iFLYTEK	Grading homework, practicing spoken language, and writing code.
MathGPT	Mathematical calculation and reasoning, and math knowledge question.
Ziyue	Speaking coach, AI writing, and AI translation.
OpenAI o1	Solving complex problems of science, coding, and math.
EduChat	Open question answering, essay assessment, and Socratic teaching.
FeiXiang	Homework assignment and correction, virtual classroom.
Zhihai	Syllabus generation and question generation.

From an application needs perspective, LLMs must support differentiated instruction and adaptive learning while ensuring pedagogical alignment with established educational frameworks. Given the diverse learning paces and abilities of K-12 students, LLMs should facilitate personalized learning by providing scaffolded guidance, timely feedback, and explanations that align with students' cognitive development stages. Currently, LLMs demonstrate strong capabilities in automating content creation and other related areas. However, challenges still exist in many aspects, such as ensuring content accuracy and mitigating biases. Therefore, LLMs should be considered as auxiliary tools rather than replacements for traditional teaching strategies in order to maximize their educational benefits.

This section reviews the board applications of LLM in K-12 education scenario supporting for students and teachers, where the advanced technologies, case, and ricks in each perspective are highlighted.

### 3.1 | Teaching Resource Recommendation

Lesson planning is the process in which teachers design and organize the content, methods, and materials for a specific lesson to ensure effective teaching and learning (Iqbal et al., 2021; Milkova, 2025). It involves setting clear objectives, selecting appropriate teaching strategies, preparing resources, and outlining activities to engage students and achieve desired learning outcomes. Importantly, finding the high-quality resources that cover the topics of the lesson is the key step and is highly used in pedagogical practice (Ding & Carlson, 2013). Educational resource recommendation technologies (Manouselis et al., 2010; Wu et al., 2020) offer a practical solution by leveraging teacher-specific requirements to identify and suggest relevant teaching and learning materials from educational databases. The previous approaches (Karpicke & Grimaldi, 2012; Urdaneta-Ponte et al., 2021) usually either used the deep neural-based retrieval network (Dou et al., 2007; Zhu et al., 2023b) to return a list of relevant resources to the user's submitted query, or explore the collaborative

filtering approach (Herlocker et al., 2004; Schafer et al., 2007), the content-based approach (Zhu et al., 2019; Zhu et al., 2020), and the hybrid approach (Deschênes, 2020; Zhu et al., 2022) to recommend useful information that is relevant to their individual needs. With the advent of LLMs, exploring the great potential of LLMs to help teachers create high-quality teaching materials has attracted more attention (Wang, 2024b; Xu et al., 2024; Zhu et al., 2023c). Leiker et al. (2023) used LLM with human-in-the-loop based prompt engineering guidance process to find high-quality content. Koraiishi (2023) used ChatGPT with zero-shot prompt strategy to optimize the materials of the English.

Here we give a real example to help readers understand how to use LLMs for resource recommendation. Figure 2 gives the details, where we use Qwen2.5 to perform the recommendation task. When user input their query, Qwen2.5 searches the website for information and uses its ability to summarize and refine the information to return the relevant resources and their URL to the user. Although LLM can effectively reduce teachers' resource preparation burden, improve efficiency, and support high-quality lesson planning, it is unable to effectively analyze teaching and learning situations and recommend personalized learning content. That is, it cannot well provide personalized recommendations for teachers by fully understanding students' learning abilities and prior knowledge.

### 3.2 | Teaching Content Creation

Educational content, such as lesson plans, teaching designs, syllabus, presentation slides, and questions, are related to the educational goals, and objectives and are necessary materials in pedagogical practice (Bell, 1993; Davis, 2017). How to generate such materials is a popular research topic in the application of LLM to education.

The lesson plans, teaching designs and syllabus are most the operational documents that guide the specific implementation of the lesson, the course, and the unit (Jury et al., 2024; Wang et al., 2024b).

Traditional methods rely on education specialists or teachers to determine the teaching or learning objectives, activities and strategies to check student understanding, and write the lesson plan, teaching design and syllabus (Milkova, 2025), which lacks flexibility and individualization. Within the era of Big Data, a large amount of data on student learning and teacher teaching has been accumulated, deriving some tools or systems that utilize deep learning models (Han et al., 2021) to mine the data and help teachers produce instructional content (Yu et al., 2019). However, these tools or systems suffered from the inability of human-machine interaction, causing the separation of lesson preparation and teaching process and their adaptive adjustments, which seriously reduces teaching efficiency (Jia et al., 2018). LLMs have strong

comprehension, generation, and interaction capabilities. Teachers can provide them with customized template, content, and background knowledge, enabling LLMs to generate lesson plans, teaching designs, and syllabi that encompass teaching objectives, learning points, and key knowledge points, among other elements, thus broadening teachers' teaching design ideas.

Here we give a real example to help readers understand how to use LLMs for generating lesson plans, teaching designs, and syllabus. Figure 3 gives the details, where we use Qwen2.5 to perform the task. When a user submits the query, Qwen2.5 claims that teaching land reforms in ancient China, particularly the Equal-Field System of Emperor Xiaowen of the Northern Wei Dynasty and Wang Anshi's Equal-Field

**Input:** I am a middle school math teacher and I am presently teaching the knowledge point "linear equation in one variable," can you recommend any relevant curriculum resources?

**Output:** The following are recommended resources to help you prepare for linear equation in one variable.

- Khan Academy. Khan Academy offers a complete collection of instructional videos and practice problems on linear equations, including quadratic equations. These resources are ideal for students to study on their own or as supplemental materials for the classroom.
- Illustrative Mathematics. Illustrative Mathematics provides Common Core Standards-aligned instructional resources, including detailed problem sets and teaching guides, which are very helpful in teaching quadratic equations!
- Teaching Design of "Understanding Quadratic Equations", the first prize of National Quality Lesson for Young Teachers of Middle School Mathematics, which provides detailed frameworks on guiding students to comprehend the concept of equations, their solutions, and constructing equations from real-world problems.
- ...

**Figure 2** LLMs in recommending educational resources. To facilitate the presentation of the generated content, we have streamlined LLMs' generated content by removing unimportant information and retaining important information related to the recommended resources. Same for other cases. LLMs: large language models.

**Input:** I am a high school history teacher. I will be teaching students about land reforms in ancient China, specifically covering the Equal-Field System of Emperor Xiaowen of the Northern Wei Dynasty and Wang Anshi's Equal-Field Taxation Law. I want students to have a basic understanding of these two historical facts and ideally, grasp the underlying thoughts behind them. I hope to maintain high levels of student engagement during the lecture. How can I structure this class?

**Output:** Here is a suggested structure for your class that could help you achieve your goals.

- 1. Introduction (5 minutes)**
  - Engage with a hook: Start with a brief story or scenario that illustrates the problems...
  - Set objectives: Clearly outline what students will learn during the lesson...
- 2. Brief historical context (10 minutes)**
  - Provide a concise overview of the time periods associated with these reforms...
- 3. Detailed examination of reforms (20 minutes)**
  - Equal-field system: Explain how the system worked, including land allocation and taxation.
  - Discuss the goals of Emperor Xiaowen, such as increasing agricultural productivity...
  - Highlight any successes and limitations of the system...
- 4. Interactive activities (10 minutes)**
  - Group discussion: Divide students into groups representing different societal roles...
  - Role play: Have students act out scenarios where they must implement one of the reforms.
  - Case study analysis: Present case studies of individual families affected by these reforms...
- 5. Homework assignment**

**Figure 3** Large language models in generating lesson plan.

Taxation Law, can be both enlightening and engaging for high school students, and gives the lesson plan composed of “introduction,” “brief historical context,” “detailed examination of reforms,” “interactive activities,” and “homework assignment.”

Presentation slides are a series of visual aids created to support teachers in delivering information, ideas, and concepts to students. The traditional method is to manually search or create videos, pictures, and text, then organize such materials to create slides (Strauss et al., 2011). This method of relying on manual labor to integrate different media types requires extra effort and technical skills, is time consuming and inefficient. MLLMs can generate various slides based on input conditioning, such as customized syllabus, teaching content, and background knowledge. Figure 4 gives a slides example. MLLMs transforms the content of the lesson plan into the slides. However, it is obvious that most of the generated slides are text-based and lack graphics, which cannot be directly used for pedagogical practice.

Question generation has become one of the most popular research topics in LLMs’ application for education (Du et al., 2017; Duan et al., 2017). Manually creating questions is a complex process that demands specialized training and practical experience. Automatic question generation (AQG) techniques (Kurdi et al., 2023) can reduce the costs of manual question creation and meet the demand for a continuous supply of new questions. AQG aims to generate questions of controlled difficulty, enrich question forms and structures, and automate template construction (Zhang et al., 2021). For example, Heilman and Smith (2010) proposed a statistical ranking network to generate and rank fact-based questions based on the content of a reading practice or assessment. The highest-ranked questions were either revised by educators or directly provided to students for practice. Recently, many LLM-based approaches proposed various strategies, such as fine-tuning LLMs with supplemental reading materials (Xiao et al., 2023), aligning with specific learning objectives (Doughty et al., 2024) and making implicit diversity controls toward the equation of question (Zhou et al., 2023b) to generate questions. Figure 5 provides the example of generating math exercise about the knowledge point “opposite vertex.” At present, simple questions can be generated, and the generation of complex questions is limited.

### 3.3 | Intelligent Question Retrieval

Intelligent question retrieval aims to analyze the content of a given problem, and rapidly match relevant questions and answers from a vast database. This

technology provides solutions, detailed explanations, and reference materials for the question input, helping students consolidate their knowledge and enhance their learning efficiency.

Most current efforts still rely on traditional AI methods due to the high cost and slow processing speed of LLMs. Furthermore, the majority of existing solutions have been developed by online education companies, such as Readboy, ZuoYeBang, TAL, and iFLYTEK, which have access to extensive question datasets. In addition to accumulating large question datasets, these approaches primarily focus on several key technical aspects: (1) enhancing the effectiveness and efficiency of retrieval strategies; (2) improving the accuracy of retrieval for complex illustrations and diagrams; (3) integrating factors such as question subject, type, and difficulty level to customize search results to meet individual needs; (4) ensuring that incomplete or defective questions are still retrieved with high precision.

The intelligent question retrieval system typically follows several stages. First, question input processing accepts the user’s question, either as text or an image for text recognition. The text undergoes preliminary processing, including word segmentation, automatic error correction, and stop word filtering. In the query semantic understanding phase, the system predicts the subject category, question type, and difficulty level by analyzing both the image and text. It also conducts word weight analysis, emphasizing high-weight words in the recall process while excluding irrelevant or low-weight terms. Synonyms are rewritten according to predefined rules to expand the recall scope. In the retrieval and reranking phase, relevant documents are first filtered from the question gallery, then reranked based on a refined score considering factors such as relevance, quality, difficulty, and popularity to select the optimal results. Figure 6 illustrates retrieval examples in an intelligent question retrieval system.

Although intelligent question retrieval systems have become a mature application widely adopted by online education enterprises, they still face several significant challenges. First, recognition accuracy is suboptimal when dealing with illegible handwriting. Second, retrieval performance is severely limited when questions contain images, as extracting key information and performing effective matching proves difficult. Third, for new types of interdisciplinary questions, where the question database coverage is insufficient, providing precise answers remains a challenge. These issues underscore the need for further advancements in both technology and methodology to improve the reliability of intelligent question retrieval systems.

### 3.4 | Intelligent Problem Solving

Intelligent problem solving is defined as the systematic generation of problem-solving processes based on the

content of a question, providing accurate and effective solutions to support personalized learning. Recent advancements in intelligent problem solving have been driven by LFM, which, having been pre-trained on vast

(a)

**Input:** I am a high school history teacher in China. I will be teaching students about land reforms in ancient China, specifically covering the Equal-Field System of Emperor Xiaowen of the Northern Wei Dynasty and Wang Anshi's Equal-Field Taxation Law. I want students to have a basic understanding of these two historical facts and ideally, grasp the underlying thoughts behind them. I hope to maintain high levels of student engagement during the lecture. How can I structure this class? Please give me the response in Chinese.

**Output:**



中国古代土地改革

1

知识引入

**知识引入**

在中国古代，土地改革一直是一个重要的政治议题。在北魏时期，孝文帝推行了均田制，而在宋代，王安石则提出了方田均税法。这些土地改革措施对于中国历史的发展产生了深远的影响。本节课我们将重点讲解北魏孝文帝的均田制和王安石的方田均税法，了解它们的具体内容以及对历史的影响。

2

知识讲解

**知识讲解**

1. 北魏孝文帝的均田制  
北魏孝文帝在位期间，为了解决土地兼并问题，推行了均田制。均田制的核心思想是将土地按照家庭人口数进行平均分配，以达到平均分配土地的目的。均田制的实施，使得土地兼并得到了有效的遏制，同时也促进了农业生产的发展。
2. 王安石的方田均税法  
宋代时期，王安石提出了方田均税法。方田均税法的核心思想是通过土地进行重新测量、重新分配，以及对税收进行统一规划，实现对土地和税收的平均分配。方田均税法的实施，一方面促进了农业生产的发展，另一方面也推动了商业和手工业的发展。
3. 均田制和方田均税法的异同  
均田制和方田均税法都是中国古代土地改革的重要措施，它们的核心思想都是平均分配土地和税收。但是，均田制更加注重土地的平均分配，而方田均税法则更加注重税收的平均分配。

3

例题

**例题**

1. 北魏孝文帝的均田制的核心思想是什么？
2. 王安石的方田均税法的实施对于中国古代经济的发展有哪些影响？
3. 均田制和方田均税法的异同点有哪些？

答案：

1. 均田制的核心思想是将土地按照家庭人口数进行平均分配，以达到平均分配土地的目的。
2. 王安石的方田均税法的实施促进了农业生产、商业和手工业的发展。
3. 均田制更加注重土地的平均分配，而方田均税法则更加注重税收的平均分配。均田制的实施时间较早，而方田均税法则是在均田制的基础上进一步完善的。

**知识点生活应用**

土地改革对于中国古代经济的发展产生了深远的影响，而在现代社会中，土地问题仍然是一个重要的议题。通过了解中国古代土地改革的历史，我们可以更好地理解现代土地改革的意义和重要性，同时也可以更好地参与到现代土地改革的进程中。

(b)

**Input:** I am a high school history teacher in China. I will be teaching students about land reforms in ancient China, specifically covering the Equal-Field System of Emperor Xiaowen of the Northern Wei Dynasty and Wang Anshi's Equal-Field Taxation Law. I want students to have a basic understanding of these two historical facts and ideally, grasp the underlying thoughts behind them. I hope to maintain high levels of student engagement during the lecture. How can I structure this class? Please give me the response in Chinese.

**Output:**



**Ancient Chinese land reforms**

**1 Knowledge introduction**

**2 Knowledge explanation**

**3 Example questions**

**Knowledge application in daily life**

Land reform has had a profound impact on the development of ancient Chinese economy. In modern society, land issues are still an important issue. Understanding the history of ancient Chinese land reform can help us better understand the significance of modern land reform. At the same time, we can also participate in the process of modern land reform.

**Knowledge introduction**

In ancient China, land reform has always been an important political issue. During the Northern Wei Dynasty, Emperor Xiaowen implemented the Equal-Field System, and in the Song Dynasty, Wang Anshi introduced the Equal-Field Tax Law. These land reform measures had a profound impact on the development of Chinese history. This section will focus on explaining Emperor Xiaowen's Equal-Field System and Wang Anshi's Equal-Field Tax Law, understanding their specific content and their impact on Chinese history.

**Knowledge explanation**

1. Emperor Xiaowen's Equal-Field System  
In order to solve the land consolidation issue, Emperor Xiaowen of the Northern Wei Dynasty implemented the Equal-Field System, which allocated land to farmers according to family size and the number of able-bodied men, achieving the purpose of equitable land distribution. This system made land distribution more equitable and also promoted the development of agricultural production.

2. Wang Anshi's Equal-Field Tax Law  
During the Song Dynasty, Wang Anshi proposed the Equal-Field Tax Law, which aimed to re-measure and re-distribute land, as well as to adjust taxes according to the actual situation of the land, achieving equitable land and tax distribution. The implementation of this law promoted the development of agricultural production and handicraft industries, and also promoted the development of commerce.

3. The similarities and differences between the Equal-Field System and the Equal-Field Tax Law  
Both the Equal-Field System and the Equal-Field Tax Law were important land reform measures in ancient Chinese history. Their core ideas were to achieve equitable distribution of land and taxation. However, the Equal-Field System focused more on the equitable distribution of land, while the Equal-Field Tax Law placed more emphasis on the equitable distribution of taxes.

**Example questions**

1. What was the core idea behind Emperor Xiaowen's Equal-Field System?  
2. What impact did Wang Anshi's Equal-Field Tax Law have on the development of ancient Chinese economy?  
3. What are the similarities and differences between the Equal-Field System and the Equal-Field Tax Law?

**Answers:**

1. The core idea of the Equal-Field System was to distribute land equally according to family size and the number of able-bodied men to achieve equitable land distribution.  
2. Wang Anshi's Equal-Field Tax Law promoted the development of agricultural production, commerce, and handicraft industries.  
3. The Equal-Field System focused more on the equitable distribution of land, while the Equal-Field Tax Law placed more emphasis on the equitable distribution of taxes.

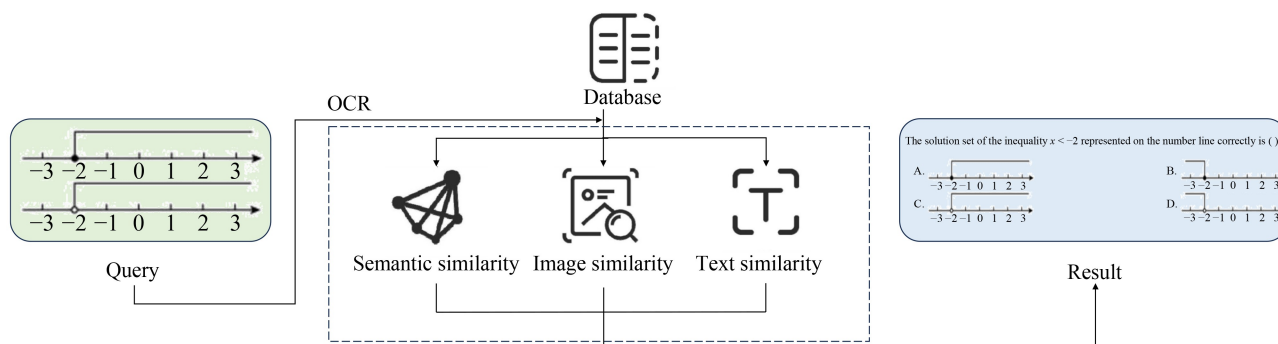
**Figure 4** Large language models in generating presentation slides. (a) Original slides created by LLMs; (b) translation of the slides.

**Input:** You are a middle school math teacher. Generate me a test on “opposite angles.” Medium difficulty.

**Output:**

- When two straight lines intersect, which pair of the four angles formed are known as opposite angles?
  - Two adjacent angles
  - Any two angles
  - Two angles in the direction of the diagonal
  - Two angles on the same side
- If  $\angle A$  and  $\angle B$  are opposite angles and  $\angle A = 5x + 30^\circ$  and  $\angle B = 4x + 60^\circ$ , what is  $x$  equal to?
  - $10^\circ$
  - $20^\circ$
  - $30^\circ$
  - $40^\circ$
- If there are three points  $A$ ,  $B$ , and  $C$  on a line, where  $B$  is the midpoint of  $AC$ , make a line  $BD$  perpendicular to  $AC$  through  $B$ . If  $\angle ABD = 70^\circ$ , then the opposite angle  $\angle CBD = \underline{\hspace{2cm}}$   $^\circ$ .
- Given two intersecting lines  $m$  and  $n$ , and an additional line  $p$  passing through the intersection such that it forms a dyadic angle with  $m$  and  $n$ , respectively. If one of the dyadic angles formed by the line  $p$  with  $m$  is  $50^\circ$  and the other dyadic angle formed with  $n$  is  $20^\circ$  greater than the former, ask: What are the degrees of each of the other two dyadic angles formed between the two intersecting lines  $m$  and  $n$ ?
  - REFERENCE ANSWERS: 1. C, 2. C, 3.  $70^\circ$ , 4. Let the angle of the dyad formed with  $m$  be  $50^\circ$  and the angle of the dyad formed with  $n$  be  $50^\circ + 20^\circ = 70^\circ$ . Since the opposite vertex angles are equal, the other two opposite vertex angles between  $m$  and  $n$  are also  $50^\circ$  and  $70^\circ$ , respectively.

**Figure 5** Large language models in generating question.



**Figure 6** An intelligent question retrieval system. OCR: optical character recognition.

amounts of text data, possess robust natural language understanding and generation capabilities. These models can effectively parse complex questions, identify relevant knowledge points, and generate systematic problem-solving steps and final answers.

LFMs, when not specifically optimized for intelligent problem-solving, often struggle with tasks requiring rigorous logical reasoning, such as mathematics. To improve their performance in this domain, various strategies have been adopted, which can be broadly classified into three categories. The first category involves prompting LLMs. Wu et al. (2023b) pioneered the use of various prompting techniques, including vanilla prompts, program of thoughts prompts, and program synthesis prompts, to enhance problem-solving capabilities. Wei et al. (2022b) introduced CoT prompting, a method that guides LLMs through step-by-step reasoning, followed by multiple CoT reasoning introduced in (Wang et al., 2022b). Zhou et al. (2023a) proposed self-verification based on explicit code, a novel prompting method that allows models to verify their answers through code. The

second category focuses on improving problem-solving performance through external strategies. An et al. (2023) enhanced LLM performance by converting mathematical expressions into English, while Yamauchi et al. (2023) employed external tools, particularly the Python REPL, to correct errors in CoT reasoning and boost performance. The third approach involves fine-tuning LFMs. Some studies (Yang et al., 2023d) enhance the model’s understanding of relevant concepts by introducing intermediate steps during the fine-tuning process, thereby improving performance. Other research incorporates additional relevant datasets, including those generated by the model itself, for further training.

The intelligent problem-solving system begins by semantically parsing the input question to identify the relevant knowledge points and concepts, drawing on the extensive knowledge base built during the model’s pre-training phase. Using this understanding, the model generates a structured sequence of problem-solving steps, including the application of formulas, logical derivations, and intermediate calculations,

ensuring the rigor and coherence of the solution process. These steps and the final answer undergo multiple rounds of internal validation and optimization to improve accuracy and reliability, often incorporating heuristic searches or optimization algorithms to verify the reasonableness of the results. Figure 7 illustrates examples of problem solving using LLMs.

LLMs still face significant limitations in intelligent problem-solving. First, while current research has largely focused on organizing large-scale datasets (He et al., 2024; Zhang et al., 2024), there remains a lack of robust generalization across different datasets, grade levels, and question types. This highlights the need for continual learning approaches, similar to the process of human skill acquisition, to improve the generalization capabilities of LLMs. Second, LLMs exhibit substantial shortcomings in reasoning, particularly in handling inconsistent performance across various textual formats (e.g., words and numbers), and in providing different answers to the same question through varying reasoning paths across multiple attempts. Third, current reasoning methods, such as CoT, fail to adequately address students' needs and comprehension levels. Studies (Gattupalli et al., 2023; Yen & Hsu, 2023) indicated that LLMs often misinterpret students' questions in dialogues and fail to provide adaptive feedback. Moreover, they tend to overlook the comprehension abilities of younger students, generating overly complex responses that may cause confusion. Thus, it is essential to incorporate human-factors design into AI research to better meet the nuanced needs of K-12 education.

### 3.5 | Intelligent Question Answering

Intelligent question answering involves the automated and immediate response to student inquiries. Its core

function is to comprehend the student's question, take into account their individual characteristics, and generate contextually relevant answers.

In K-12 education, LLMs must offer precise responses tailored to specific subjects and tasks. To meet this demand, fine-tuning techniques have become crucial for enhancing the effectiveness of intelligent question answering systems (Chen et al., 2023c). In addition, methods such as prompt engineering, CoT, and the Socratic method have been employed to enhance model interaction and learning outcomes (Westerlund & Shcherbakov, 2024). Retrieval-augmented generation techniques have further improved LLM functionality by incorporating external databases (Miladi et al., 2024; Sequeda, et al., 2024), thereby alleviating the errors and hallucinations, an important feature in educational contexts where accuracy is essential. As technology evolves, intelligent question answering systems are increasingly incorporating multimodal capabilities (Lim et al., 2024; Luo et al., 2023). This expansion goes beyond text-based inputs and outputs, enabling the simultaneous processing of text, voice, images, and other forms of data, thus offering richer and more diverse support for students' learning experiences.

Figure 8 presents two practical applications of intelligent question-answering technology in education (Dan et al., 2023). Figure 8(a) involves an open-domain question-answering system based on the retrieval-augmented generation method. In this scenario, the system retrieves the correct fact—that Cao Cao is the author of the poem—and generates a detailed response that not only provides the author's information but also explains the poem's creation background and literary value, including a reference link for further reading. Figure 8(b) demonstrates an interactive learning system using the Socratic method for intelligent question-

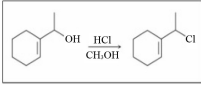
Question: Which of the statement best describes the reaction in the following figure?

Option:

(A) It is likely to be SN1.

(B) It is likely to be SN2.

(C) None of the above.



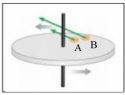
Question: Two coins rotate on a turntable. Coin B is twice as far from the axis as coin A.

Option:

(A) The speed of A is twice that of B.

(B) The speed of A equals that of B.

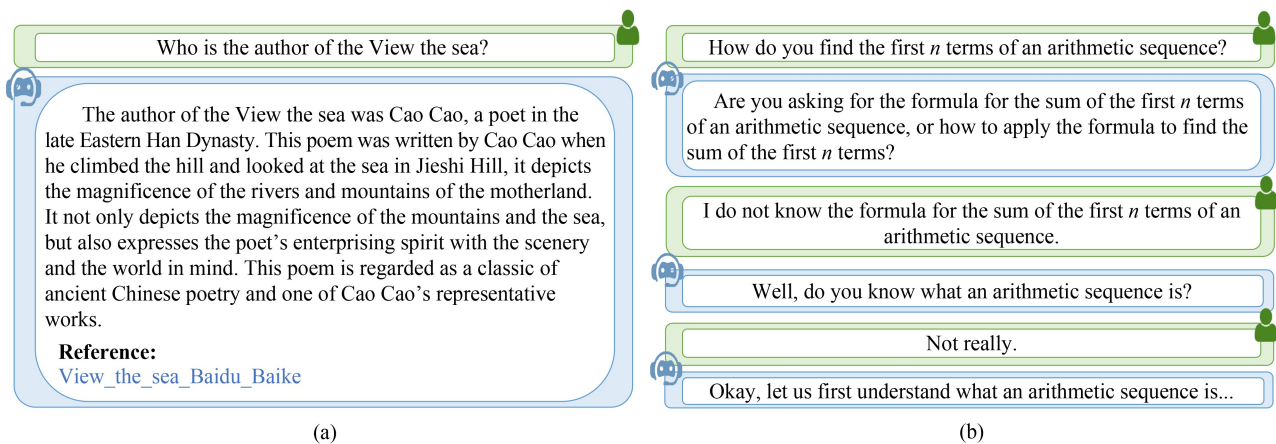
(C) The speed of A is half that of B.



The reaction depicted involves the conversion of an alcohol to an alkyl chloride in the presence of HCl and CH<sub>3</sub>OH (methanol). Given the reagents and conditions: (A) It is likely to be SN1. The alcohol initially forms a carbocation intermediate after losing the hydroxyl group in the presence of HCl. Then, chloride ion attacks the carbocation to form the alkyl chloride product. The mechanism of this reaction is consistent with the SN1 pathway.

For objects rotating in a circle, the linear speed ( $v$ ) is related to the angular speed ( $\omega$ ) by:  $v = \omega r$ , where  $v$  is the linear speed of the object,  $\omega$  is the angular speed of the turntable (same for all objects on the turntable), and  $r$  is the distance of the object from the axis of rotation. Given that coin B is twice as far from the axis as coin A, let us denote the distance of coin A from the axis as  $r$  and the distance of coin B from the axis as  $2r$ . The speed of coin A is:  $v_A = \omega r$ . The speed of coin B is:  $v_B = \omega(2r) = 2\omega r$ . Comparing the two speeds:  $v_B = 2v_A$ . Thus, the speed of A is half that of B. The correct answer is: (C) The speed of A is half that of B.

Figure 7 Large language models in intelligent problem solving.



**Figure 8** Large language models in intelligent question answering. (a) An open-domain question-answering system; (b) an interactive learning system using the Socratic method.

answering. When the question been raised, the system does not provide a direct answer but instead guides the user through a series of probing questions, helping them progressively understand the concept.

LFMs have demonstrated impressive capabilities in intelligent question answering, yet they continue to face significant challenges in practical applications, particularly in handling complex mathematical calculations and teaching methodologies. Despite notable advancements, most developments have focused on higher education, with customized models tailored for K-12 education still lacking widespread adoption. As a result, the coverage and depth of these models fall short of addressing the diverse and personalized needs of K-12 learners. Furthermore, existing educational LFMs have not systematically incorporated the deeper knowledge and principles of pedagogy, limiting their ability to fully support personalized learning, provide emotional feedback, or deliver in-depth subject knowledge for both teachers and students.

### 3.6 | Automated Scoring

Automated scoring is a revolutionary application in the field of education, shown in Figure 9. This technology not only alleviates the workload of teachers but also enhances grading efficiency and objectivity. Automatic scoring is typically divided into scoring for objective questions and subjective questions. The scoring process for objective questions is well-established and is typically carried out by comparing responses to standardized answers, enabling rapid and accurate grading.

However, the challenges encountered by automated scoring technology in scoring subjective questions are more complex. The main reason lies in the higher requirements for semantic analysis, contextual reasoning, creativity, and critical thinking

assessment. Traditional scoring for subjective questions relies on human evaluators, who, while capable of providing detailed feedback and personalized assessments, tends to be less efficient and is easily influenced by subjective factors. In recent years, automatic grading technologies based on LLMs have begun to emerge. Kundu & Barbosa (2024) evaluated the effectiveness of LLMs such as ChatGPT and Llama in the automated essay scoring task, with a particular focus on their consistency with human scores. They found that LLMs often scored more stringently than human raters and exhibited weak correlation with human grading. Additionally, they explored the ability of LLMs to identify spelling and grammatical errors, discovering that LLMs could reliably detect these mistakes and take them into account during scoring. Kim & Jo (2024) introduced a novel hybrid method that combines LLMs and comparative judgment (CJ). This method used zero-shot prompting to enable LLMs to select the better of two essays, simulating the human evaluator's comparative judgment process. The study results indicated that the CJ approach outperformed traditional rubric-based scoring methods in LLM-driven essay assessments. Rationale-based multiple trait scoring (Chu et al., 2024) combined prompt-based LLMs with fine-tuned small LLMs to accurately predict multi-faceted scores based on feature-specific rationales generated by LLMs linked to scoring criteria.

These studies indicate that while LLMs demonstrate potential in automated grading tasks, there is still room for improvement in their understanding of scoring criteria and their ability to generate consistent scores. Future research must further explore how to optimize LLMs to better mimic the behavior of human evaluators and enhance the accuracy and reliability of automated grading systems. Moreover, in the scoring of subjective questions, the results generated by LLMs may be influenced by factors such as race, gender, and culture, which can lead to biased outcomes.

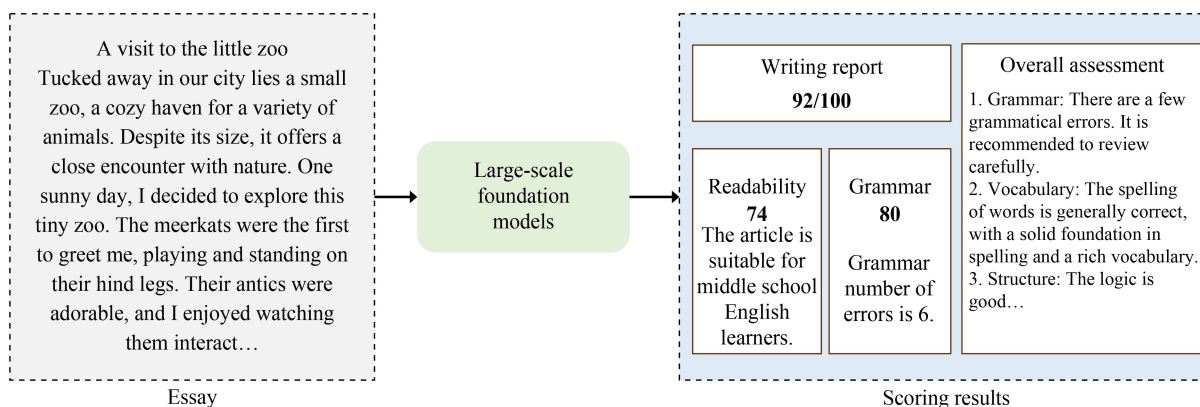


Figure 9 Large language models in automated scoring.

### 3.7 | Exam Paper Generation

Exam paper generation refers to the process of automatically generating exam papers that meet educational requirements. Traditionally, teachers need to manually select appropriate questions from a large pool of items in the question bank. This process is often cumbersome and time-consuming, while automated methods effectively address this issue. As shown in Figure 10, the model integrates various constraints, including students' knowledge levels, teaching objectives, and question types, to identify and extract relevant questions from the question bank, subsequently organizing them into a coherent and structured exam paper.

Commonly used methods include random question selection algorithms, backtracking algorithms, and genetic algorithms (Cen et al., 2010; Joshi et al., 2016; Umardand & Gaikwad, 2017). Random selection algorithms are typically used in certification exams. Backtracking algorithms are commonly applied in adaptive assessments, such as the TOEFL language proficiency test. Genetic algorithms, which iteratively optimize an initial set of question combinations, progressively generate the most optimal test set that satisfies the constraints, making them particularly effective in solving high-dimensional, multi-constraint problems.

In recent years, with the advancement of deep learning, neural network-based exam paper generation algorithms have begun to emerge. Cen et al. (2010) employed an adaptive genetic algorithm based on artificial intelligence and information processing to optimize the question selection and paper organization process. Joshi et al. (2016) suggested that the use of NLP in the exam generation process could directly extract textual information from questions, preventing the frequent selection of certain questions. Wang et al. (2020) proposed a question selection strategy for cognitive diagnosis, integrating artificial intelligence

recognition models to enhance the efficiency and accuracy of the question selection. The advent of LfMs has brought new breakthroughs to exam paper generation systems. Unlike traditional methods, LfMs are capable of processing unstructured raw question text, automatically analyzing and understanding the semantics, context, and underlying key points within the questions, without the need for prior complex labeling or manual feature extraction. This capability allows LfMs to generate exam papers that more flexibly meet specific educational needs, further enhancing the intelligence of the exam paper generation process.

However, there are still several limitations. First, the performance heavily depends on the accurate modeling of the generation objectives. If the objective modeling is flawed, such as inaccurate estimates of difficulty distribution, it could result in exam papers that are unreasonable in terms of difficulty control. Second, in addition to basic requirements for difficulty and key points coverage, real-world educational scenarios may involve more complex exam paper requirements, such as assessing creative thinking and critical analysis abilities. Therefore, exam paper generation systems need to integrate additional dimensions to ensure that the generated papers are comprehensive and reasonable.

### 3.8 | Question Quality Evaluation

Test items and examination papers are essential tools widely used in evaluating students' academic performance and teachers' instructional quality. They not only directly reflect students' learning outcomes but also serve as important instruments for assessing teaching effectiveness and guiding instructional improvements. Therefore, the quality assessment of test items and examination papers is a critical process in educational measurement. High-quality test items and examination papers need to balance several key aspects.

(1) **Fairness.** All candidates are tested under the

### Paper generation result

Part 1: Multiple Choice

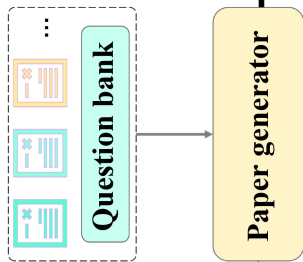
1. Given the complex number  $z = \frac{2}{(1-i)^2} - 1$ , the imaginary part of  $z$  is ( )  
 A. 2      B. 1      C. -1      D. -2 Star Replace Add

2. Given that the complex number  $z$  satisfies  $z(2-i) = 1+i$ , the imaginary part of  $z$  is ( )  
 A.  $-\frac{3}{5}$       B.  $\frac{3}{5}$       C.  $-\frac{3}{2}$       D.  $\frac{3}{2}$  Star Replace Add

3. Given that the complex number  $z$  satisfies  $z + 1 = 3i + |z|$ , then  $|z| = ( )$   
 A. 5      B. 4      C. 3      D. 2 Star Replace Add

4. If the vector  $\vec{a}, \vec{b}$  satisfies  $|\vec{a}| = 2, |\vec{b}| = 1, (\vec{a} + 2\vec{b}) \cdot \vec{a} = 6$ , then  $\cos(\vec{a}, \vec{b}) = ( )$ .  
 A.  $\frac{\sqrt{3}}{2}$       B.  $\frac{1}{2}$       C.  $-\frac{1}{2}$       D.  $-\frac{\sqrt{3}}{2}$  Star Replace Add

5. Given the complex number  $z = \frac{(i-2)^2}{i}$ , the point corresponding to  $z$  in the complex plane is located at ( )  
 A. First quadrant      B. Second quadrant      C. Third quadrant      D. Fourth quadrant Star Replace Add



### Paper generation setting

Scenario: Class Exercises | Phase Tests | College Entrance Examination

Selected the Range of Examination  
 Chapter 6: Plane Vectors | Chapter 7: Complex Numbers

Preferences  
 Difficulty:  Unrestricted  Easy  Moderately Easy  Moderate  Moderately Difficult  Difficult  Overall

Priority Area: All Regions

Monthly Exam  
 Multiple Choice(8)  
 Short Answer(4)  
 Scenario-Based(6)

Midterm/Final Exam  
 Multiple Choice(8)  
 True/False(4)  
 Problem Solving(3)  
 Scenario-Based(1)

Custom Template  
 +

Generate Paper

Figure 10 Large language models in exam paper generation.

same conditions.

(2) **Reliability.** Students get similar scores when tested at different times.

(3) **Validity.** The questions should cover the necessary knowledge points, and maintain an appropriate level of difficulty.

(4) **Teaching improvement.** By analyzing students' answers, teachers can understand the strengths and weaknesses of each student, and adjust teaching strategies.

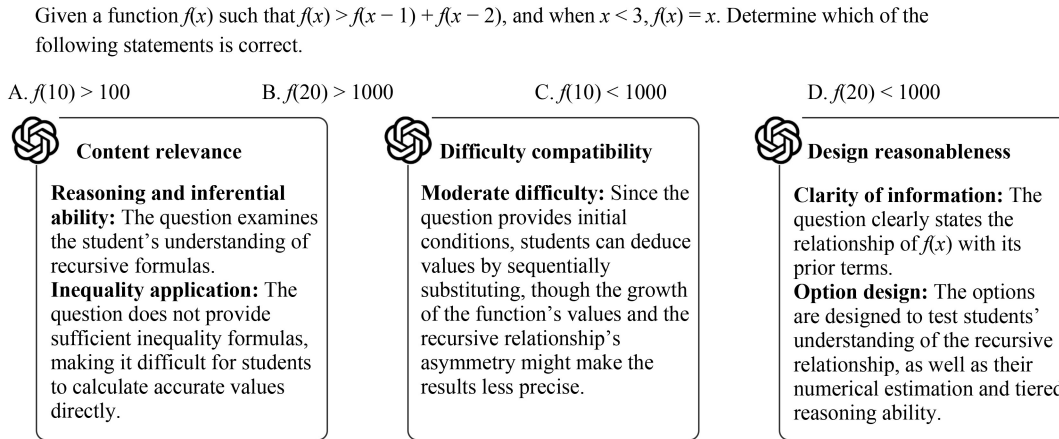
The quality assessment of test items and exam papers mainly includes single-item assessment and whole-paper assessment. In single-item assessment, the focus is on ensuring that each question effectively measures specific knowledge and skills. The assessment dimensions typically include content relevance, difficulty compatibility, discrimination power, design reasonableness, and uniqueness of answers. In whole-paper assessment, the focus is on thoroughly evaluating the entire test paper to ensure its logical structure, appropriate difficulty level, comprehensive content coverage, and alignment with curriculum requirements.

Currently, most quality assessments rely heavily on manual review. While manual review ensures high accuracy, it suffers from low efficiency and strong subjectivity, making it unsuitable for large-scale assessments. Metrics like bilingual evaluation understudy, recall-oriented understudy for gisting evaluation, and metric for evaluation of translation with explicit ordering, commonly used in machine translation tasks, have been applied to quality assessments. However, these rely on manually crafted test items as benchmarks and merely compare structural similarities, lacking deeper correlations. Consequently, they offer limited practical value for high-quality evaluations of test items and examination papers (Görgün & Bulut, 2024). Nowadays, LFM have been used in this field, aiding in the improvement of assessments. First, we can build student profiles from educational data using deep learning methods such as clustering, decision trees, and knowledge graphs. These profiles, combined with test items and examination papers, serve as inputs for LFM, which can produce multi-dimensional evaluation results (see Figure 11).

However, there are still several limitations. First, LFM may lack in-depth understanding of specialized knowledge, leading to biases when assessing domain-specific items. Second, current evaluation standards lack strong associations and integration with curriculum standards, making it difficult to conduct a comprehensive evaluation of quality.

### 3.9 | Discussion

LFMs for K-12 education are typically developed by fine-tuning pre-trained models on extensive educational



**Figure 11** Large language models in the quality assessment for test items and examination papers.

datasets, such as exam questions and textbooks. While these models capitalize on the robust comprehension and generative capabilities to deliver promising results in various educational scenarios, their underlying approach remains rooted in data fitting. They heavily depend on massive datasets, extensive parameters, and immense computational power, epitomizing a brute-force methodology. However, this approach falls short of true intelligence in the strictest sense and continues to face notable limitations and challenges.

### 3.9.1 Lack of Multimodal Capabilities

Currently, LFM's mainly focus on breakthroughs in NLP, but their abilities in text-to-image and text-to-video generation are still far from practical use. This makes it hard to meet the multimodal needs in education. For example, LFM's are not good at generating teaching materials that combine text and visuals. Educational data naturally includes rich and complex multimodal information, where different modalities complement and cannot replace each other. LFM's for K-12 education must be able to accurately understand and reason across multiple modalities.

### 3.9.2 Complex Reasoning Remains a Weakness

LFMs rely on the accumulation of massive correlations, which is essentially akin to the "Clever Hans" effect. They use a data-driven paradigm to identify correlations but fail to achieve reasoning that requires causal relationships. Xu et al. (2023) showed that as the number of reasoning steps increases, the reasoning performance of LFM's (such as ChatGPT and text-davinci-003) drops significantly.

### 3.9.3 Poor Understanding of Diagrams

Diagrams are a unique type of visual content widely

used in K-12 education, differing significantly from natural images in terms of sample quantity, low-level features, and high-level semantics. LFM's are primarily trained on massive datasets of natural images, fitting well to the distribution of these images. However, their performance remains poor on diagrams, which have distinct characteristics. This limitation affects their ability to achieve satisfactory results in solving problems in subjects like mathematics and physics.

### 3.9.4 Poor Personalization Capability.

Currently, educational applications based on LFM's mainly focus on personalization concepts inherited from prior computer-assisted learning experiences. These applications are typically limited to a few aspects, such as recommending learning paths. However, effective teaching requires consideration of each student's unique characteristics and personalized optimization based on their profiles.

### 3.9.5 Lack of Educational Expertise.

Educational activities are highly specialized knowledge transfer and talent development processes that must follow systematic and rigorous educational theories. While some LFM's fail to integrate educational and pedagogical theories into the entire process of model construction and application. As a result, issues still arise, such as inaccuracies in the knowledge and facts used for reasoning and a reasoning process that does not comply with the standards of educational applications.

## 4 Future Prospects

According to the research above, we summarize some

directions for improvement and development in the field of LFMs for K-12 education.

(1) **Knowledge understanding.** To construct LFMs for K-12 education, it is essential to build multimodal databases tailored to educational Big Data and create diverse knowledge graphs. This would provide a foundation for deep understanding, complex reasoning, and accurate generation. Additionally, based on the multimodal databases and knowledge graphs, designing appropriate training tasks can help the LFMs better understand subject-specific knowledge and educational content.

(2) **Teaching expertise.** Integrating educational and teaching principles and theory into the design and optimization of LFMs can enable them to better assist teachers in their teaching. For example, how to integrate learning motivation theory and emotional factors in educational psychology into the training and application of the model, so as to better meet the diverse needs of students in K-12 education. After that, accurate profiles can be generated, which reflect students' learning behaviors, knowledge levels, and emotional states. Based on these profiles, adaptive teaching content planning algorithms should be developed to dynamically generate personalized teaching paths. These paths would consider both the student's cognitive level and emotional state, thereby better supporting the implementation of teaching activities of LFMs.

(3) **Studying assistance.** Education has shifted from a knowledge-transfer, lecture-based model to a more heuristic approach that focuses on developing core competencies. Therefore, LFMs should not only assist students with learning, but also emphasize skill development. For example, models could incorporate Socratic-style guiding interactions and empathetic response mechanisms to meet the needs of exploratory, interactive, and personalized learning.

(4) **Privacy and Ethics.** Recently, privacy and ethical concerns have gained increasing attention, yet they remain insufficiently addressed in current research. To enhance data security, institutions can adopt differential privacy techniques to prevent data leakage. Additionally, they can implement homomorphic encryption for secure computations on sensitive educational data and leverage federated learning to reduce risks associated with centralized data storage. To address ethical challenges in model-generated content, such as bias, discrimination, and factual inaccuracies, techniques like adversarial debiasing and RLHF can be integrated. These technical methods are essential for improving the trustworthiness and security of LLMs in educational applications while ensuring alignment with pedagogical principles and ethical standards.

(5) **Evaluation benchmarks and methods.**

Since LFMs are intended for students whose cognitive abilities are still developing, the reliability and safety of their outputs become even more critical. However, many existing LFMs still present significant application risks. Issues such as inaccurate content and the generation of biased, discriminatory, abusive, or unlawful statements remain unresolved. Therefore, it is crucial to establish reasonable benchmarks, which would involve rigorous testing of the model's outputs to ensure their safety and reliability. Building on these benchmarks, evaluation methods for various application scenarios should be tailored to align with educational goals and students' needs. For instance, in intelligent question-answering systems, metrics such as students' comprehension levels and their improvement in problem-solving skills can serve as key evaluation indicators. This approach ensures a more precise assessment of the practical impact and effectiveness of LFMs in educational settings.

## 5 Conclusions

The rapid advancement of LFMs is reshaping education and opening up new possibilities in various education applications. This review provides an in-depth exploration of LFMs applications, categorized into student and teacher support such as recommending instructional resources, generating instructional content, solving problems, answering questions, etc. Along with a discussion of the challenges and limitations, we suggest directions for future research and experts to guide and inspire further progress in using LFMs to improve education.

**Conflict of Interest** The authors declare that they have no conflict of interest related to the content of this paper.

**Data Availability Statements** The authors confirm that all data generated or analyzed during this study are included in this published article.

**Authors Contributions** Qiannan Zhu wrote Sections 2 and 5 and Subsections 3.1 and 3.2; Mei Wang wrote Subsections 3.6 to 3.9 and Section 4; Ting Zhang wrote Abstract, Section 1, and Subsections 3.3 to 3.5; Hua Huang revised the manuscript and provided funding support. All authors whose names appear on the submission made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work; drafted the work or revised it critically for important intellectual content; approved the version to be published; and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

## References

- AI Meta. (2024a, April 18). *Introducing meta Llama 3: The most capable openly available LLM to date*. Available from AI Meta website.
- AI Meta. (2024b, July 23). *Introducing Llama 3.1: Our most capable models to date*. Available from AI Meta website.
- Alam, A. (2023). Harnessing the power of AI to create intelligent tutoring systems for enhanced classroom experience and improved learning outcomes. In: *Proceedings of Intelligent Communication Technologies and Virtual Mobile Networks*. Singapore: Springer, 571–591.
- Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millicah, K., Reynolds, M., & et al. (2022). Flamingo: A visual language model for few-shot learning. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. New York: Curran Associates Inc., 23716–23736.
- Alkhatlan, A., & Kalita, J. (2018). Intelligent tutoring systems: A comprehensive historical survey with recent developments. *arXiv Preprint*, arXiv:1812.09628.
- An, J., Lee, J., & Gweon, G. (2023). Does ChatGPT comprehend the place value in numbers when solving math word problems? In: *Proceedings of Human-AI Math Tutoring@AIED 2023*.
- Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., & et al. (2023). OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv Preprint*, arXiv:2308.01390.
- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv Preprint*, arXiv:2308.12966.
- Bai, Y., Chen, H., Chen, J., Chen, Z., Deng, Y., Dong, X., Hantrakul, L., Hao, W., Huang, Q., Huang, Z., & et al. (2024). Seed-music: A unified framework for high quality and controlled music generation. *arXiv Preprint*, arXiv:2409.09214.
- Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1), 905–971.
- Bell, A. (1993). Principles for the design of teaching. *Educational Studies in Mathematics*, 24(1), 5–34.
- Bhagavatula, C., Le Bras, R., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S. W., & Choi, Y. (2019). Abductive commonsense reasoning. *arXiv Preprint*, arXiv:1908.05739.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & et al. (2020). Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. New York: Curran Associates Inc., 1877–1901.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., & et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv Preprint*, arXiv:2303.12712.
- Cen, G., Dong, Y., Gao, W., Yu, L., See, S., Wang, Q., Yang, Y., & Jiang, H. (2010). A implementation of an automatic examination paper generation system. *Mathematical and Computer Modelling*, 51(11-12), 1339–1342.
- Chall, J. S. (1983). *Stages of reading development*. Belmont: Wadsworth Publishing Company.
- Chen, F., Han, M., Zhao, H., Zhang, Q., Shi, J., Xu, S., & Xu, B. (2023a). X-LLM: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv Preprint*, arXiv:2305.04160.
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8, 75264–75278.
- Chen, X., Shen, J., Xia, W., Jin, J., Song, Y., Zhang, W., Liu, W., Zhu, M., Tang, R., Dong, K., & et al. (2023b). Set-to-sequence ranking-based concept-aware learning path recommendation. In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. Cambridge: AAAI Press, 5027–5035.
- Chen, Y., Chen, H., & Su, S. (2023c). Fine-tuning large language models in education. In: *Proceedings of 2023 13th International Conference on Information Technology in Medicine and Education*. IEEE, 718–723.
- Chen, Z., Zhou, K., Zhang, B., Gong, Z., Zhao, W. X., & Wen, J.-R. (2023d). ChatCot: Tool-augmented chain-of-thought reasoning on chat-based large language models. *arXiv Preprint*, arXiv:2305.14323.
- Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., & Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2818–2829.
- Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4, 100118.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., & et al. (2023). PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(1), 11324–11436.
- Chu, S., Kim, J., Wong, B., & Yi, M. (2024). Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs. *arXiv Preprint*, arXiv:2410.14202.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., & Brahma, S. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(1), 3381–3433.
- Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). InstructBLIP: Towards general-

- purpose vision-language models with instruction tuning. *arXiv Preprint*, arXiv:2305.06500.
- Dan, Y., Lei, Z., Gu, Y., Li, Y., Yin, J., Lin, J., Ye, L., Tie, Z., Zhou, Y., Wang, Y., & et al. (2023). Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv Preprint*, arXiv:2308.02773.
- Davis, M. (2017). *Teaching design: A guide to curriculum and pedagogy for college design faculty and teachers who use design in their classrooms*. New York: Simon and Schuster.
- Denny, P., Prather, J., Becker, B. A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B. N., Santos, E. A., & Sarsa, S. (2024). Computing education in the era of generative AI. *Communications of the ACM*, 67(2), 56–67.
- Deschênes, M. (2020). Recommender systems to support learners' agency in a learning context: A systematic review. *International Journal of Educational Technology in Higher Education*, 17(1), 50.
- Deshmukh, S., Elizalde, B., Singh, R., & Wang, H. (2023). Pengi: An audio language model for audio tasks. In: *Proceedings of Advances in Neural Information Processing Systems*. New York: Curran Associates Inc., 18090–18108.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 4171–4186.
- Dewey, J. (1986). Experience and education. *The Educational Forum*, 50(3), 241–252.
- Ding, M., & Carlson, M. A. (2013). Elementary teachers' learning to construct high-quality mathematics lesson plans: A use of the IES recommendations. *The Elementary School Journal*, 113(3), 359–385.
- Dou, Z., Song, R., & Wen, J.-R. (2007). A large-scale evaluation and analysis of personalized search strategies. In: *Proceedings of Proceedings of the 16th International ACM Conference on World Wide Web*. New York: Association for Computing Machinery, 581–590.
- Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., & et al. (2024). A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. In: *Proceedings of the 26th Australasian Computing Education Conference*. New York: Association for Computing Machinery, 114–123.
- Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *arXiv Preprint*, arXiv:1705.00106.
- Duan, N., Tang, D., Chen, P., & Zhou, M. (2017). Question generation for question answering. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. New York: ACM, 866–874.
- Elizalde, B., Deshmukh, S., Al Ismail, M., & Wang, H. (2023). Clap learning audio concepts from natural language supervision. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1–5.
- Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement*, 59(3), 272–287.
- Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., & Cao, Y. (2023). Eva: Exploring the limits of masked visual representation learning at scale. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 19358–19369.
- Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2022). Complexity-based prompting for multi-step reasoning. *arXiv Preprint*, arXiv:2210.00720.
- Gattupalli, S., Lee, W., Allessio, D. A., Crabtree, D., Arroyo, I., Woolf, B., & Woolf, B. P. (2023). Exploring pre-service teachers' perceptions of large language models-generated hints in online mathematics learning. In: *Proceedings of International Conference of Artificial Intelligence in Education, IEEE*, 151–162.
- Gemma Team. (2024a). Gemma: Open models based on gemini research and technology. *arXiv Preprint*, arXiv:2403.08295.
- Gemma Team. (2024b). Gemma 2: Improving open language models at a practical size. *arXiv Preprint*, arXiv:2408.00118.
- GLM Team. (2024). ChatGLM: A family of large language models from GLM-130b to GLM-4 all tools. *arXiv Preprint*, arXiv:2406.12793.
- Görgün, G., & Bulut, O. (2024). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. *Education and Information Technologies*, 29(6), 6911–6947.
- Guo, J., Li, J., Li, D., Tiong, A. M. H., Li, B., & Tao, D. (2023). From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 10867–10877.
- Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., & et al. (2023). ImageBind-LLM: Multi-modality instruction tuning. *arXiv Preprint*, arXiv:2309.03905.
- Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. *arXiv Preprint*, arXiv:2103.00112.
- He, Z., Wu, X., Zhou, P., Xuan, R., Liu, G., Yang, X., Zhu, Q., & Huang, H. (2024). CMMU: A benchmark for Chinese multi-modal multi-type question understanding and reasoning. *arXiv Preprint*, arXiv:2401.14011.
- Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. In: *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, 609–617.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv Preprint*, arXiv:2009.03300.
- Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1),

- 5–53.
- Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, 16, 235–266.
- iFLYTEK. (2024, March 30). *AutoSpark*. Available from iFLYTEK website. (in Chinese).
- Iqbal, M. H., Siddiqie, S. A., & Mazid, M. A. (2021). Rethinking theories of lesson plan for effective teaching and learning. *Social Sciences & Humanities Open*, 4(1), 100172.
- Jia, S., Liu, J., & Wu, F. (2018). Design and application of a lesson planning system to promote deep learning based on learning analytics. In: *Proceedings of Seventh International Conference of Educational Innovation Through Technology*. IEEE, 150–155.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., & et al. (2023). Mistral 7B. *arXiv Preprint*, arXiv:2310.06825.
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., & Bressand, F. (2024). Mixtral of experts. *arXiv Preprint*, arXiv:2401.04088.
- Joshi, A. C., Kudnekar, P., Joshi, M. K., & Doiphode, S. R. (2016). A survey on question paper generation system. In: *Proceedings of National Conference on Role of Engineers in National Building*. Available on VIVA website.
- Jury, B., Lorusso, A., Leinonen, J., Denny, P., & Luxton-Reilly, A. (2024). Evaluating LLM-generated worked examples in an introductory programming course. In: *Proceedings of Australasian Computing Education Conference*. ACM, 77–86.
- Kapur, M. (2008). Productive failure. *Cognition and Instruction*, 26(3), 379–424.
- Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*, 24(3), 401–418.
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., & et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.
- Kim, S., & Jo, M. (2024). Is GPT-4 alone sufficient for automated essay scoring? A comparative judgment approach based on rater cognition. *arXiv Preprint*, arXiv:2407.05733.
- Koraishi, O. (2023). Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education and Technology*, 3(1), 55–72.
- Kundu, A., & Barbosa, D. (2024). Are large language models good essay graders? *arXiv Preprint*, arXiv:2409.13120.
- Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30, 121–204.
- Latif, E., Mai, G., Nyaaba, M., Wu, X., Liu, N., Lu, G., Li, S., Liu, T., & Zhai, X. (2023). Artificial general intelligence (AGI) for education. *arXiv Preprint*, arXiv:2304.12479.
- Lazaridou, A., Gribovskaya, E., Stokowiec, W., & Grigorev, N. (2022). Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv Preprint*, arXiv:2203.05115.
- Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., & Gallé, M., & et al. (2023). BLOOM: A 176B-parameter open-access multilingual language model. *arXiv Preprint*, arXiv:2211.05100.
- Leiker, D., Finnigan, S., Gyllen, A. R., & Cukurova, M. (2023). Prototyping the use of large language models (LLMs) for adult learning content creation at scale. *arXiv Preprint*, arXiv:2306.01815.
- Li, H., Xu, T., Zhang, C., Chen, E., Liang, J., Fan, X., Li, H., Tang, J., & Wen, Q. (2024). Bringing generative AI to adaptive learning in education. *arXiv Preprint*, arXiv:2402.14601.
- Li, J., Li, D., Savarese, S., & Hoi, S. (2023a). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *Proceedings of the 40th International Conference on Machine Learning*. ACM, 19730–19742.
- Li, K. C., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., & Qiao, Y. (2023b). VideoChat: Chat-centric video understanding. *arXiv Preprint*, arXiv:2305.06355.
- Li, Q., Fu, L., Zhang, W., Chen, X., Yu, J., Xia, W., Zhang, W., Tang, R., & Yu, Y. (2023c). Adapting large language models for education: Foundational capabilities, potentials, and challenges. *arXiv Preprint*, arXiv:2401.08664.
- Lim, Q. Z., Lee, C. P., Lim, K. M., & Samingan, A. K. (2024). UniRaG: Unification, retrieval, and generation for multimodal question answering with pre-trained language models. *IEEE Access*, 12, 71505–71519.
- Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., & et al. (2023). SPHINX: The joint mixing of weights, tasks, and visual embeddings for multimodal large language models. *arXiv Preprint*, arXiv:2311.07575.
- Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). Visual instruction tuning. *arXiv Preprint*, arXiv:2304.08485.
- Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. *arXiv Preprint*, arXiv:2310.03744.
- Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). What makes good in-context examples for GPT-3? *arXiv Preprint*, arXiv:2101.06804.
- Liu, T., Chen, Z., Fang, Z., Luo, W., Tian, M., & Liu, Z. (2025). MathEval: A comprehensive benchmark for evaluating large language models on mathematical reasoning capabilities. *Frontiers of Digital Education*, 2(2), 16.
- Luo, H., Shen, Y., & Deng, Y. (2023). Unifying text, tables, and images for multimodal question answering. In: *Proceedings of Findings of the Association for Computational Linguistics*. ACL, 9355–9367.
- Maghsudi, S., Lan, A., Xu, J., & van der Schaar, M. (2021). Personalized education in the artificial intelligence era: What to expect next? *IEEE Signal Processing Magazine*, 38(3), 37–50.
- Manouselis, N., Vuorikari, R., & Van Assche, F. (2010). Collaborative recommendation of e-learning resources: An

- experimental investigation. *Journal of Computer Assisted Learning*, 26(4), 227–242.
- Miladi, F., Psyché, V., & Lemire, D. (2024). Comparative performance of GPT-4, RAG-augmented GPT-4, and students in MOOCs. In: *Proceedings of International Conference on Breaking Barriers with Generative Intelligence*. Cham: Springer, 81–92.
- Milkova, S. (2025). *Strategies for effective lesson planning*. Available from Center for Research on Learning and Teaching website.
- OpenAI. (2023). GPT-4 technical report. *arXiv Preprint*, arXiv:2303.08774.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & et al. (2022). Training language models to follow instructions with human feedback. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. ACM, 27730–27744.
- Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Kong, L., & Zhang, T. (2023). DetGPT: Detect what you need via reasoning. *arXiv Preprint*, arXiv:2305.14167.
- Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.
- Qwen Team. (2024, September 19). *Qwen2.5: A party of foundation models!* Available from Qwenlm website.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., & et al. (2021). Learning transferable visual models from natural language supervision. *arXiv Preprint*, arXiv:2103.00020.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2020). *Improving language understanding by generative pre-training*. Available from OpenAI website.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. Available from OpenAI website.
- Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *arXiv Preprint*, arXiv:1806.03822.
- Rasley, J., Rajbhandari, S., Ruwase, O., & He, Y. (2020). DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 3505–3506.
- Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12–27.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355.
- Rubin, O., Herzig, J., & Berant, J. (2021). Learning to retrieve prompts for in-context learning. *arXiv Preprint*, arXiv:2112.08633.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68.
- Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In: Brusilovsky, P., Kobsa, A., & Nejdl, W., eds. *The adaptive web: Methods and strategies of web personalization*. Berlin: Springer, 291–324.
- Science. (2022). *The biggest science breakthroughs in 2022*. Available from Science website.
- Sequera, J., Allemang, D., & Jacob, B. (2024). A benchmark to understand the role of knowledge graphs on large language model's accuracy for question answering on enterprise SQL databases. In: *Proceedings of the 7th Joint Workshop on Graph Data Management Experiences & Systems and Network Data Analytics*. ACM, 1–12.
- Shafik, W. (2024). Introduction to ChatGPT. In: *Advanced applications of generative AI and natural language processing models*. Hershey: IGI Global, 1–25.
- Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv Preprint*, arXiv:1909.08053.
- Sovacool, B. K., & Ryan, S. E. (2016). The geography of energy and education: Leaders, laggards, and lessons for achieving primary and secondary school electrification. *Renewable and Sustainable Energy Reviews*, 58, 107–123.
- Strauss, J., Corrigan, H., & Hofacker, C. F. (2011). Optimizing student learning: Examining the use of presentation slides. *Marketing Education Review*, 21(2), 151–162.
- Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., & Cai, D. (2023). PandaGPT: One model to instruction-follow them all. *arXiv Preprint*, arXiv:2305.16355.
- Sun, Q., Fang, Y., Wu, L., Wang, X., & Cao, Y. (2023). EVA-CLIP: Improved training techniques for clip at scale. *arXiv Preprint*, arXiv:2303.15389.
- Susnjak, T., & McIntosh, T. R. (2024). ChatGPT: The end of online exam integrity. *Education Sciences*, 14(6), 656.
- Tan, K., Pang, T., Fan, C., & Yu, S. (2023). Towards applying powerful large AI models in classroom teaching: Opportunities, challenges and prospects. *arXiv Preprint*, arXiv:2305.03433.
- Tapalova, O., & Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIED for personalised learning pathways. *Electronic Journal of e-Learning*, 20(5), 639–653.
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930–1940.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & et al. (2023a). Llama: Open and efficient foundation language models. *arXiv Preprint*, arXiv:2302.13971.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., & et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv Preprint*, arXiv:2307.09288.
- Umardand, A., & Gaikwad, A. (2017). A survey on automatic question paper generation system. *International Advanced Research Journal in Science, Engineering and Technology*, 4(4),

- 18–20.
- Urdaneta-Ponte, M. C., Mendez-Zorrilla, A., & Oleagordia-Ruiz, I. (2021). Recommendation systems for education: Systematic review. *Electronics*, *10*(14), 1611.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv Preprint*, arXiv:1706.03762.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.
- Wang, N., Wang, D., & Zhang, Y. (2020). Design of an adaptive examination system based on artificial intelligence recognition model. *Mechanical Systems and Signal Processing*, *142*, 106656.
- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., & Yang, H. (2022a). OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv Preprint*, arXiv:2202.03052.
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024a). Tutor CoPilot: A human-AI approach for scaling real-time expertise. *arXiv Preprint*, arXiv:2410.03017.
- Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., & Wen, Q. (2024b). Large language models for education: A survey and outlook. *arXiv Preprint*, arXiv:2403.18105.
- Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., & et al. (2023). CogVLM: Visual expert for pretrained language models. *arXiv Preprint*, arXiv:2311.03079.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022b). Self-consistency improves chain of thought reasoning in language models. *arXiv Preprint*, arXiv:2203.11171.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv Preprint*, arXiv:2109.01652.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., & et al. (2022a). Emergent abilities of large language models. *arXiv Preprint*, arXiv:2206.07682.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022b). Chain-of-thought prompting elicits reasoning in large language models. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. ACM, 24824–24837.
- Westerlund, M., & Shcherbakov, A. (2024). LLM integration in workbook design for teaching coding subjects. In: *Proceedings of International Conference on Science and Technology Education*. Cham: Springer, 77–85.
- Wolf, M. (2007). *Proust and the squid: The story and science of the reading brain*. New York: HarperCollins Publishers.
- Wu, L., Liu, Q., Zhou, W., Mao, G., Huang, J., & Huang, H. (2020). A semantic web-based recommendation framework of educational resources in e-learning. *Technology, Knowledge and Learning*, *25*, 811–833.
- Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023a). BloombergGPT: A large language model for finance. *arXiv Preprint*, arXiv:2303.17564.
- Wu, Y., Jia, F., Zhang, S., Li, H., Zhu, E., Wang, Y., Lee, Y. T., Peng, R., Wu, Q., & Wang, C. (2023b). An empirical study on challenging math problem solving with GPT-4. *arXiv Preprint*, arXiv:2306.01337.
- Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023). 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*. ACL, 610–625.
- Xiong, Z., Li, H., Liu, Z., Chen, Z., Zhou, H., Rong, W., & Ouyang, Y. (2024). A review of data mining in personalized education: Current trends and future prospects. *Frontiers of Digital Education*, *1*(1), 26–50.
- Xu, F., Lin, Q., Han, J., Zhao, T., Liu, J., & Cambria, E. (2023). Are large language models really good logical reasoners? A comprehensive evaluation from deductive, inductive and abductive views. *arXiv Preprint*, arXiv:2306.09841.
- Xu, L., Zhang, J., Li, B., Wang, J., Chen, S., Zhao, W. X., & Wen, J.-R. (2024). Prompting large language models for recommender systems: A comprehensive framework and empirical analysis. *arXiv Preprint*, arXiv:2401.04997.
- Yamauchi, R., Sonoda, S., Sannai, A., & Kumagai, W. (2023). LPML: LLM-prompting markup language for mathematical reasoning. *arXiv Preprint*, arXiv:2309.13078.
- Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., & Yan, D. (2023a). Baichuan 2: Open large-scale language models. *arXiv Preprint*, arXiv:2309.10305.
- Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., & Huang, F. (2024). Qwen2 technical report. *arXiv Preprint*, arXiv:2407.10671.
- Yang, H., Liu, X.-Y., & Wang, C. D. (2023b). FinGPT: Open-source financial large language models. *arXiv Preprint*, arXiv:2306.06031.
- Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., & Liu, N. (2023c). Large language models in health care: Development, applications, and challenges. *Health Care Science*, *2*(4), 255–263.
- Yang, Z., Ding, M., Lv, Q., Jiang, Z., He, Z., Guo, Y., Bai, J., & Tang, J. (2023d). GPT can solve mathematical problems without a calculator. *arXiv Preprint*, arXiv:2309.03241.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *arXiv Preprint*, arXiv:2210.03629.
- Yen, A.-Z., & Hsu, W.-L. (2023). Three questions concerning the use of large language models to facilitate mathematics learning. *arXiv Preprint*, arXiv:2310.13615.
- Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023a). A survey on multimodal large language models. *arXiv Preprint*, arXiv:2306.13549.
- Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., & Chen, E. (2023b). Woodpecker: Hallucination correction for multimodal large language models. *arXiv Preprint*, arXiv:2310.16045.

- Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, 31(7), 1235–1270.
- Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., & Zhu, J. (2024). Osprey: Pixel understanding with visual instruction tuning. *arXiv Preprint*, arXiv:2312.10032.
- Zhang, G., Du, X., Chen, B., Liang, Y., Luo, T., Zheng, T., Zhu, K., Cheng, Y., Xu, C., Guo, S., & et al. (2024). Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv Preprint*, arXiv:2401.11944.
- Zhang, H., Li, X., & Bing, L. (2023a). Video-Llama: An instruction-tuned audio-visual language model for video understanding. *arXiv Preprint*, arXiv:2306.02858.
- Zhang, R., Guo, J., Chen, L., Fan, Y., & Cheng, X. (2021). A review on question generation from natural language text. *ACM Transactions on Information Systems*, 40(1), 1–43.
- Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., & Qiao, Y. (2023b). Llama-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv Preprint*, arXiv:2303.16199.
- Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., & Xie, W. (2023c). PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv Preprint*, arXiv:2305.10415.
- Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). Automatic chain of thought prompting in large language models. *arXiv Preprint*, arXiv:2210.03493.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., & et al. (2023). A survey of large language models. *arXiv Preprint*, arXiv:2303.18223.
- Zheng, W., Wuyun, Q., Freddolino, L., & Zhang, Y. (2023). Integrating deep learning, threading alignments, and a multi-MSA strategy for high-quality protein monomer and complex structure prediction in CASP15. *Proteins*, 91(12), 1684–1703.
- Zhong, T., Liu, Z., Pan, Y., Zhang, Y., Zhou, Y., Liang, S., Wu, Z., Lyu, Y., Shu, P., Yu, X., & et al. (2024). Evaluation of OpenAI o1: Opportunities and challenges of AGI. *arXiv Preprint*, arXiv:2409.18486.
- Zhou, A., Wang, K., Lu, Z., Shi, W., Luo, S., Qin, Z., Lu, S., Jia, A., Song, L., Zhan, M., & Li, H. (2023a). Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification. *arXiv Preprint*, arXiv:2308.07921.
- Zhou, Z., Ning, M., Wang, Q., Yao, J., Wang, W., Huang, X., & Huang, K. (2023b). Learning by analogy: Diverse questions generation in math word problem. *arXiv Preprint*, arXiv:2306.09064.
- Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023a). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv Preprint*, arXiv:2304.10592.
- Zhu, Q., Zhang, H., He, Q., & Dou, Z. (2022). A gain-tuning dynamic negative sampler for recommendation. In: *Proceedings of ACM Web Conference*. ACM, 277–285.
- Zhu, Q., Zhang, H., He, Q., & Dou, Z. (2023b). Query-aware explainable product search with reinforcement knowledge graph reasoning. *IEEE Transactions on Knowledge and Data Engineering*, 36, 1260–1273.
- Zhu, Q., Zhou, X., Song, Z., Tan, J., & Guo, L. (2019). DAN: Deep attention neural network for news recommendation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 5973–5980.
- Zhu, Q., Zhou, X., Wu, J., Tan, J., & Guo, L. (2020). A knowledge-aware attentional reasoning network for recommendation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 6999–7006.
- Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., & Wen, J.-R. (2023c). Large language models for information retrieval: A survey. *arXiv Preprint*, arXiv:2308.07107.