Large Language Models for Education: A Survey and Outlook

Shen Wang[1*], Tianlong Xu[1*], Hang Li[2*], Chaoli Zhang[3], Joleen Liang[4],
Jiliang Tang[2], Philip S. Yu[5], Qingsong Wen[1†]

[1]Squirrel AI, USA   [2]Michigan State University, USA   [3]Zhejiang Normal University, China
[4]Squirrel AI, China   [5]University of Illinois Chicago, USA

ABSTRACT

The advent of large language models (LLMs) has brought in a new era of possibilities in the realm of education. This survey paper summarizes the various technologies of LLMs in educational settings from multifaceted perspectives, encompassing student and teacher assistance, adaptive learning, and commercial tools. We systematically review the technological advancements in each perspective, organize related datasets and benchmarks, and identify the risks and challenges associated with the deployment of LLMs in education. Furthermore, we outline future research opportunities, highlighting the potential promising directions. Our survey aims to provide a comprehensive technological picture for educators, researchers, and policymakers to harness the power of LLMs to revolutionize educational practices and foster a more effective personalized learning environment.

1 INTRODUCTION

During the past decades, artificial intelligence (AI) for education has received a great deal of interest and has been applied in various educational scenarios [20, 26, 40, 92, 98, 118]. Specifically, educational data mining methods have been widely adopted in different aspects such as cognitive diagnosis, knowledge tracing, content recommendations, as well as learning analysis [12, 84, 148–151, 187].

As large language models (LLMs) have become a powerful paradigm in different areas [23, 47, 73, 199], they also achieved state-of-the-art performances in multiple educational scenarios [78, 101, 192]. Existing work has found that LLMs can achieve student-level performance on standardized tests [136] in a variety of mathematics subjects (e.g., physics, computer science) on both multiple-choice and free-response problems. In addition, empirical studies have shown that LLMs can serve as a writing or reading assistant for education [120, 164]. A recent study [164] reveals that ChatGPT is capable of generating logically consistent answers across disciplines, balancing both depth and breadth. Another quantitative analysis [120] shows that students using ChatGPT (by keeping or refining the results from LLMs as their own answers) perform better than average students in some courses from the field of computer security. Recently, several perspective papers [77, 166] also explore various application scenarios of LLMs in classroom teaching, such as teacher-student collaboration, personalized learning, and assessment automation. However, the application of LLMs in education may lead to a series of practical issues, e.g., plagiarism, potential bias in AI-generated content, overreliance on LLMs, and inequitable access for non-English speaking individuals [78].

To provide researchers with a broad overview of the domain, numerous exploratory and survey papers have been proposed. For example, Qadir [145], Rahman and Watanobe [146] and Rahman

and Watanobe [146] conclude the applications of ChatGPT to engineering education by analyzing the responses of ChatGPT to the related pedagogical questions. Jeon and Lee [70] and Mogavi et al. [127] collect the opinions from different ChatGPT user groups, e.g., educator, learner, researcher, through in-person interviews, online post replies and user logs, and conclude the practical applications of LLMs in education scenarios. Baidoo-Anu and Owusu Ansah [10] and Zhang and Tur [201] focus on the literature review over the published papers and summarize the progress of the area with structured tables. Although the above works have covered a wide range of existing applications of LLMs in education scenarios and provided their long-term visions for future studies, we argue that none of the literature has systematically summarized LLMs for education from a technological perspective. To bridge this gap, this survey aims to provide a comprehensive technological review of LLMs for education, which provides a novel technology-centric taxonomy and summary of existing publicly available datasets and benchmarks. Furthermore, we summarize current challenges as well as further research opportunities, in order to foster innovations and understanding in the dynamic and ever-evolving landscape of LLMs for education. In summary, our contributions lie in the following three main parts:

1, Comprehensive and up-to-date survey. We offer a comprehensive and up-to-date survey on LLMs for a wide spectrum of education, containing academic research, commercial tools, and related datasets and benchmarks.

2, New technology-centric taxonomy. We provide a new taxonomy that offers a thorough analysis from a technological perspective on LLMs for education, encompassing student and teacher assistance, adaptive learning, and commercial tools.

3, Current challenges and future research directions We discuss current risks and challenges, as well as highlight future research opportunities and directions, urging researchers to dive deeper into this exciting area.

2 LLM IN EDUCATION APPLICATIONS

2.1 Overview

The education application can be categorized based on its users' role in education and its usage scenario in education. In this paper, we summarize the appearance of LLMs in different applications and discuss the benefits brought by LLMs compared to the original methods. We present our primary summary of education applications with LLMs using a taxonomy illustrated in Figure 1.

2.2 Study Assisting

Providing students with timely learning support has been widely recognized as a crucial factor in improving student engagement and learning efficiency during their independent studies [41]. Due

---

Shen Wang[1*], Tianlong Xu[1*], Hang Li[2*], Chaoli Zhang[3], Joleen Liang[4],
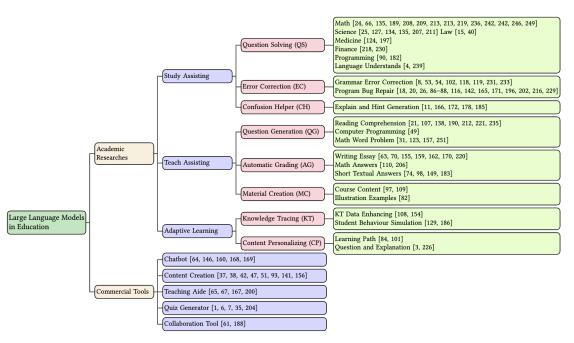Jiliang Tang[2], Philip S. Yu[5], Qingsong Wen[1†]

Figure 1: A taxonomy of LLMs for education applications with representative works.

to the limitation of prior algorithms in generating fixed-form responses, many of the existing study-assisting approaches face poor generalization challenges while being implemented in real-world scenarios [85]. Fortunately, the appearance of LLMs brings revolutionary changes to this field. Using finetuned LLMs [137] to generate human-like responses, recent studies in LLM-based educational support have demonstrated promising results. These studies provide real-time assistance to students by helping them solve challenging questions, correcting errors, and offering explanations or hints for areas of confusion.

2.2.1 Question Solving (QS). Contributing to the large-scale parameter size of LLMs and the enormous sized and diverse web corpus used during the pre-training phase, LLMs have been proven to be a powerful question zero-shot solver to questions spread from a wide spread of subjects, including math [184, 198], law [15, 35], medicine [107, 168], finance [183, 194], programming [79, 156], language understands[4, 202]. In addition, to further improve LLM's problem-solving performance while facing complicated questions, a variety of studies have been actively proposed. For example, Wei et al. [178] propose the Chain-of-Thought (CoT) prompting method, which guides LLMs to solve a challenging problem by decomposing it into simpler sequential steps. Other works [162, 176] exploit the strong in-context learning ability of LLMs and propose advanced few-shot demonstration-selection algorithms to improve LLM's problem-solving performance to general questions. Chen et al. [21] and Gao et al. [53] leverage external programming tools to avoid calculation errors introduced during the textual problem solving process of raw LLMs. Wu et al. [181] regard chat-optimized LLMs as powerful agents and design a multi-agent conversation to solve those complicated questions through a collaborative process. Cobbe et al. [28] and Zhou et al. [209] propose the external verifier module to rectify intermediate errors during generation, which improves

LLM problem solving performance on challenging math questions. Overall, with the proposition of all these novel designs, the use of LLMs for question solving has achieved impressive progress. Furthermore, students can find high-quality answers to their blocking questions in a timely manner.

2.2.2 Error Correction (EC). Error correction focuses on providing instant feedback to students' errors that they make during the learning process. This action is helpful for students in the early stages of learning. Zhang et al. [204] explore using four prompt strategies: zero-shot, zero-shot-CoT, few-shot, and few-shot-CoT to correct common grammar errors in Chinese and English text. From their experiment, they find that LLMs have tremendous potential in the correction task, and some simple spelling errors have been perfectly solved by the current LLMs. GrammarGPT [48] leverages LLM for addressing native Chinese grammatical errors. By fine-tuning open-source LLMs with hybrid annotated dataset, which involves both human annotation and ChatGPT generation, the proposed framework performs effectively in native Chinese grammatical error correction. Zhang et al. [200] propose to use a large language model trained in code, such as Codex, to build an APR system – MMAPR – for introductory Python programming assignments. With MMAPR evaluated on real student programs and comparing it to the prior state-of-the-art Python syntax repair engine, the author found that MMAPR can fix more programs and produce smaller patches on average. Do Viet and Markov [43] develop a few-shot example generation pipeline, which involves code summarize generation and code modification to create few-shot examples. With the generated few-shot examples, the performance and stability of LLMs on bug fixing performance on student programs receive a great boost.

2.2.3 Confusion Helper (CH). Unlike QS and AC, studies in the confusion helper direction avoid providing correct problem solutions directly. Instead, these works aim to use LLMs to generate pedagogical guidance or hints that help students solve problems themselves. Shridhar et al. [158] propose various guided question generation schemes based on input conditioning and reinforcement learning, and explore the ability of LLMs to generate sequential questions to guide the solution of math word problems. Prihar et al. [144] explore using LLMs to generate explanations for math problems in two ways: summarizing question-related tutoring chat logs and learning a few shots from existing explanation text. Based on their experiment, they find the synthetic explanations cannot outperform teacher-written explanations, as some terms may not be known by students, and the advice is sometimes too general. The research by Pardos and Bhandari [139] evaluates the learning gain differences between ChatGPT and human-tutor-generated algebra hints. By observing the changes in participants' pre-test and post-test scores between the controlled groups, the author draws a similar conclusion that hints generated by LLMs are less effective in guiding students to find worked solutions. Balse et al. [11] evaluate the validity of using LLMs to generate explaining text to logical errors in students' computer programming homework. By ranking the synthetic explanations between the ones written by course TAs, the author finds that synthetic explanations are competitive to human-generated results but have shortages in correctness and information missing problems. Rooein et al. [152] experiment with generating adaptive explanations for different groups of students. By introducing controlling conditions, such as age group, education level, and detail level, to the instructional prompt, the proposed method adapts the generated explanations to students with diverse learning profiles.

2.3 Teach Assisting

Contributing to LLM's unprecedented logical reasoning and problem-solving capability, developing LLM-based teach-assisting models has become another popular topic in education research recently. With the help of these assisting algorithms, instructors are able to get rid of prior burdensome routine workloads and focus their attention on tasks like in-class instructions, which cannot be replaced by existing machine learning models.

2.3.1 Question Generation (QG). Due to its highly frequent usage in pedagogical practice, Question Generation (QG) has become one of the most popular research topics in LLMs' application for education. Xiao et al. [186] leverage LLMs to generate reading comprehension questions by first fine-tuning it with supplemental reading materials and textbook exercise passages, then by employing a plug-and-play controllable text generation approach, fine-tuned LLMs are guided in generating more coherent passages based on specified topic keywords. Doughty et al. [44] analyze the ability of LLM (GPT-4) to produce multiple-choice questions (MCQs) aligned with specific learning objectives (LO) of Python programming classes in higher education. By integrating several generation control modules with the prompt assembly process, the proposed framework is capable of producing MCQs with clear language, a single correct choice, and high-quality distractors. Lee et al. [94] focused

on aligning prompting questions and reading comprehension taxonomy using a 2D matrix-structured framework. Using aligned prompts, LLM questions can cover a broad range of question types and difficulty levels in a balanced manner. Zhou et al. [211] work on generating diverse math word problems with implicit diversity controls toward the equation of question and achieve the goal of generating high-quality diverse questions.

2.3.2 Automatic Grading (AG). Research on automatic assignment graders has been proposed much earlier to the recent emergence of LLMs. However, due to the limitations in the learning capability of prior models, most existing auto-grading algorithms [111] focus on exploring semantic comparisons between golden solutions and student responses, which dismisses the logical considerations behind manual scoring processes. Apart from that, as the quality of the provided solution heavily influences the result, the applications of previous works are restricted to some well-annotated problems. Fortunately, with the appearance of LLMs, the above challenges have become easy to solve. Studies by Yancey et al. [193] and Pinto et al. [143] first explore the usage of LLMs for automatic scoring in open-ended questions and writing essays using prompt tuning algorithms. By including comprehensive contexts, clear rubrics, and high-quality examples, LLMs demonstrate its satisfactory performance on both grading tasks. Xiao et al. [185] further integrates CoT within the grading process. This approach instructs LLMs to first analyze and explain the provided materials before making final score determinations. With such modifications, the LLMs will not only generate the score results, but also provide detailed comments to the students' responses, which helps students to learn how to improve for the next time. Li et al. [97] extend the grading object from the textual responses of the students to those with handwritten responses. With the use of the advanced multimodal LLM framework, for example, CLIP and BLIP, the work demonstrates that the incorporation of the student's text and image, as well as the text and image of the question, improves the model's grading performance. Funayama et al. [51] propose a pre-finetuning method for the cross-prompt to learn the shared relationship between different rubrics and annotated examples, then by further tuning the pre-finetuned LLMs on the target scoring task, the model can achieve comparable performance under the limitation of labeled samples.

2.3.3 Material Creation (MC). Despite the above tasks, pioneering researchers also find the great potential of LLMs to help teachers create high-quality educational materials. For example, Leiker et al. [96] presents an investigation into the use of LLMs in asynchronous course creation, particularly within the context of adult learning, training, and upskilling. To ensure the accuracy and clarity of the generated content, the author integrates LLMs with a robust human-in-the-loop process. Koraishi [86] leverage GPT-4 with a zero-shot prompt strategy to optimize the materials of the English as a Foreign Language (EFL) course. In their exploration, the authors examine how ChatGPT can be used in material development, streamlining the process of creating engaging and contextually relevant resources tailored to the needs of individual learners, as well as other more general uses. Jury et al. [74] present a novel tool, 'WorkedGen', which uses LLMs to generate interactive worked examples. Through the use of strategies such as prompt chaining
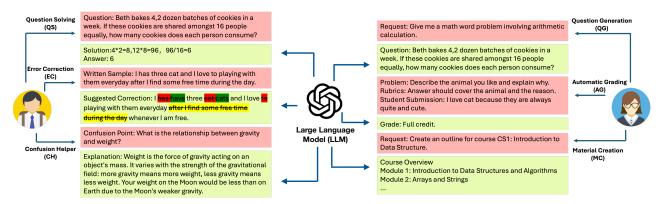
Shen Wang[1*], Tianlong Xu[1*], Hang Li[2*], Chaoli Zhang[3], Joleen Liang[4],
Jiliang Tang[2], Philip S. Yu[5], Qingsong Wen[1†]

Figure 2: LLMs in student and teacher assisting.

and one-shot learning to optimize the output, the generated work examples receive positive feedback from students.

2.4 Adaptive Learning

Based on the specific problems solved by the proposed methods, existing work on adaptive learning can be classified into two categories: knowledge tracing [2] and content personalization [130]. To be specific, knowledge tracing targets estimating the students' knowledge mastery status based on the correctness of students' responses to questions during their study processes. Content personalizing focuses on providing customized learning content to students based on personalized factors such as learning status, preferences, and goals. During the past few decades, a variety of machine learning algorithms, including traditional statistical methods [88] and advanced deep learning models [109], have been explored by different studies, and some promising results have been achieved for both problems [110]. With the recent surge of powerful LLMs in various applications, novel opportunities are also emerging for research in these directions.

2.4.1 Knowledge Tracing (KT). The current usage of LLMs in knowledge tracing focuses on generating auxiliary information to both the question text and the student records data. In recent work by Ni et al. [132], the author uses LLM to extract the knowledge keyword for each text of questions in the student-question response graph. Contributing to LLM's strong generous capabilities to deal with unseen text, the proposed framework proves especially advantageous in addressing cold start scenarios characterized by limited student question practice data. In addition to that, Lee et al. [95] proposes a framework, DCL4KT+LLM, which predicts the difficulties of questions based on the text of the question stem and the concepts of knowledge associated with LLM. Using the predicted question difficulties, DCL4KT+LLM overcomes the missing difficulty information problem of existing knowledge tracing algorithms when faced with unseen questions or concepts. Finally, Sonkar and Baraniuk [159] explores the capabilities of LLM in logical reasoning with distorted facts. By utilizing the prompts designed by the study, the LLMs demonstrate the possibility of simulating students' incorrect responses when given the appropriate knowledge profiles of the students.

2.4.2 Content Personalizing (CP). As most advanced LLMs are generative models, the use of LLMs to create personalized learning content has been explored in many recent education researches. For example, Kuo et al. [89] attempts to generate a dynamic learning path for students based on their most recent knowledge mastery diagnosis result. Kabir and Lin [76] incorporate the knowledge concept structures during the generation. Specifically, if the student masters a topic for a given Learning Object (LO), a question from the next LO will be automatically generated. Yadav et al. [191] explore the potential of LLMs in creating contextualizing algebraic questions based on student interests. By conducting iterative prompt engineering on the few-shot learning approach, the system aptly accommodates novel interests such as TikTok and NBA to the generated question stem text, which helps to improve student engagement and outcomes during the study. Despite generating content, other studies [3] also try to leverage chat-based LLMs to generate explanations of learning recommendations. By utilizing Knowledge Graphs (KGs) as a source of contextual information, the approach demonstrates its capability to generate convincing answers for a learner who has inquiries about the learning path recommended by ITS systems.

2.5 Education Toolkit

Besides leveraging LLMs to empower well-formulated education applications in academia, several LLM-powered commercial education tools have been developed in the industry. In particular, they can be categorized into five categories, including Chatbot, Content Creation, Teaching Aide, Quiz Generator and Collaboration Tool.

2.5.1 Chatbot. Using a Large Language Model (LLM) chatbot as an educational tool offers a range of advantages and opportunities. LLM chatbots can adapt their responses to the individual needs of learners, providing personalized feedback and support. This customization can accommodate different learning styles, speeds, and preferences. They offer 24/7 availability, making learning accessible anytime, anywhere. This can be particularly beneficial for learners in different time zones or with varying schedules. The interactive nature of chatbots can make learning more engaging and fun. They can simulate conversations, create interactive learning scenarios, and provide instant feedback, which can be more effective than passive learning methods. Chatbots can handle thousands of queries

simultaneously, making them a scalable solution for educational institutions to support a large number of learners without a corresponding increase in teaching staff. They can automate repetitive teaching tasks, such as grading quizzes or providing basic feedback, allowing educators to focus on more complex and creative teaching responsibilities. There are some representative chatbots, such as ChatGPT [135], Bing Chat [123], Google Bard [58], Perplexity [141], Pi Pi.ai [142].

2.5.2 Content Creation. Curipod [37] takes user input topics and generates an interactive slide deck, including polls, word clouds, open-ended questions, and drawing tools. Diffit [42] provides a platform on which the user can find leveled resources for virtually any topic. It enables teachers to adapt existing materials to suit any reader, create customized resources on any subject, and then edit and share these materials with students. MagicSchool [119] is an LLM-powered educational platform that aims to help teachers save time by automating tasks such as lesson planning, grading, and creating educational content. It provides access to more than 40 AI tools, which can be searched by keywords and organized into categories for planning, student support, productivity, and community tools. Education Copilot [32] offers LLM-generated templates for a variety of educational needs, including lesson plans, writing prompts, handouts, student reports, project outlines, and much more, streamlining the preparation process for educators. Nolej [133] specializes in creating a wide range of interactive educational content, including comprehensive courses, interactive videos, assessments, and plug-and-play content, to enhance the learning experience. Eduaide.ai [46] is an LLM-powered teaching assistant created to support teachers in lesson planning, instructional design, and the creation of educational content. It features a resource generator, teaching assistant, feedback bot, and AI chat, providing comprehensive assistance for educators. Khanmigo [82], developed by Khan Academy, is an LLM-powered learning tool that serves as a virtual tutor and debate partner. It can also assist teachers in generating lesson plans and handling various administrative tasks, enhancing both learning and teaching experiences. Copy.ai [33]is an LLM-powered writing tool that uses machine learning to produce a wide range of content types, such as blog headlines, emails, social media posts, and web copy.

2.5.3 Teaching Aide. gotFeedback [59] is developed to assist teachers in providing more personalized and timely feedback to their students, seamlessly integrating into the gotLearning platform. It is based on research emphasizing that effective feedback should be goal-referenced, tangible and transparent, actionable, user-friendly, timely, ongoing, and consistent, ensuring that it meets students' needs effectively. Grammarly [60] serves as an online writing assistant, employing LLM to help students write bold, clear, and error-free writing. Grammarly's AI meticulously checks grammar, spelling, style, tone, and more, ensuring your writing is polished and professional. Goblin Tools [171] offers a suite of simple single-task tools specifically designed to assist neurodivergent individuals with tasks that may be overwhelming or challenging. This collection includes Magic ToDo, Formalizer, Judge, Estimator, and Compiler, each tool catering to different needs and simplifying daily tasks to enhance productivity and ease. ChatPDF [140] is an LLM-powered tool designed to let users interact with PDF documents through a conversational interface. This innovative approach enables easier navigation and interaction with PDF content, making it more accessible and user-friendly.

2.5.4 Quiz Generator. QuestionWell [1] is an LLM-based tool that generates an unlimited supply of questions, allowing teachers to focus on what is most important. By entering reading material, AI can create essential questions, learning objectives, and aligned multiple choice questions, streamlining the process of preparing educational content and assessments. Formative [6], a platform for assignments and quizzes that accommodated a wide range of question types, has now enhanced its capabilities by integrating ChatGPT. This addition enables the generation of new standard-aligned questions, hints for learners, and feedback for students, leveraging the power of LLM to enrich the educational experience and support customized learning paths. Quizizz AI [7] is an LLM-powered feature that specializes in generating multiple-choice questions, which has the ability to automatically decide the appropriate number of questions to generate based on the content supplied. Furthermore, Quizizz AI can modify existing quizzes through its Enhance feature, allowing for the customization of activities to meet the specific needs of students. Conker [30] is a tool that enables the creation of multiple-choice, read-and-respond, and fill-in-the-blank quizzes tailored to students of various levels on specific topics. It also supports the usage of user input text, from which it can generate quizzes, making it a versatile resource for educators aiming to assess and reinforce student learning efficiently. Twee [174] is an LLM-powered tool designed to streamline lesson planning for English teachers, generating educational content, including questions, dialogues, stories, letters, articles, multiple choice questions, and true/false statements. This comprehensive support helps teachers enrich their lesson plans and engage students with a wide range of learning materials.

2.5.5 Collaboration Tool. summarize.tech [161] is a ChatGPT-powered tool that can summarize any long YouTube video, such as a lecture, a live event, or a government meeting. Parlay Genie [56] serves as a discussion prompt generator that creates higher-order thinking questions for classes based on a specific topic, a YouTube video, or an article. It uses the capabilities of ChatGPT to generate engaging and thought-provoking prompts, facilitating deep discussions and critical thinking among students.

3 DATASET AND BENCHMARK

LLMs revolutionized the field of natural language processing (NLP) by enabling a wide range of text-rich downstream tasks, which leverage the extensive knowledge and linguistic understanding embedded within LLMs to perform specific functions requiring comprehension, generation, or text transformation. Therefore, many datasets and benchmarks are constructed for text-rich educational downstream tasks. The majority of datasets and benchmarks lie in the tasks of question-solving (QS), error correction (EC), question generation (QG), and automatic grading (AG), which cover use cases that benefit different users, subjects, levels, and languages. Some of these datasets mainly benefit the student, while others help the teacher.

Datasets and benchmarks for educational applications vary widely in scope and purpose, targeting different aspects of the educational process, such as student performance data [147], text and resource

Shen Wang[1*], Tianlong Xu[1*], Hang Li[2*], Chaoli Zhang[3], Joleen Liang[4],
Jiliang Tang[2], Philip S. Yu[5], Qingsong Wen[1†]

databases [16], online learning data [155], language learning database [169], education game data [113], demographic and socioeconomic data [31], learning management system (LMS) data [29], special education and needs data [128]. Specifically, the question-solving ones [8, 22, 28, 64, 67, 83, 114, 116, 122, 177, 207], take a significant amount as a prevalent task for both Education and NLP fields. In particular, many datasets [8, 28, 64, 67, 116, 122, 177, 207] are constructed for math question-solving, aiming to provide an abstract expression from a narrative description. Some datasets also take the images [81, 83, 114, 122] and tables [116] into consideration. On the other hand, another bunch of datasets and benchmarks [22, 81, 83], are constructed for science textbook question-solving, which requires a comprehensive understanding of the textbook and provide the answer corresponding to the key information in the question. There are also large amounts of datasets and benchmarks constructed for error correction. They are used for foreign language training [17, 34, 39, 45, 61, 129, 131, 153, 154, 165, 172, 189, 205] and computer science programming language training [62, 75, 93, 105, 108, 173]. The foreign language training datasets and benchmarks contains grammatical errors and spelling errors that are needed to be identified and corrected. The programming training datasets and benchmarks include several code bugs that require sufficient coding understanding for proper correction. On the other hand, there are several datasets and benchmarks for teacher-assisting tasks. [13, 19, 57, 63, 90, 106, 180, 190] are constructed for question construction task that aims to evaluate the ability of the LLM for generating the educational questions from a given context. [14, 160, 170, 194] are built for automatic grading the student assignments. We summarize the commonly used publicly available datasets and benchmarks for evaluating LLMs on education applications in Table 1 of the Appendix[1].

4   RISKS AND POTENTIAL CHALLENGES

This section discusses risks and challenges along with the rise of generative AI and LLMs and summarizes some early proposals for the implementation of guardrails and responsible AI. Given the importance of education as a critically important domain, more caution should be used when implementing the implications of LLMs. A well-established framework on responsible AI [124] has outlined six foundational elements: fairness, inclusiveness, reliability & safety, privacy & security, transparency, and accountability. Besides these, for education domain, the concern of overreliance is also a major concern as over-depending on LLMs will harm some key capabilities of students such as critical thinking, academic writing, and even creativity.

4.1   Fairness and Inclusiveness

Limited by LLM training data, where representations of specific groups of individuals and social stereotypes might be dominant, bias could develop [212]. Li et al. [103] summarized that for the education domain, critical LLM fairness discussions are based on demographic bias and counterfactual concerns. Fenu et al. [49] has introduced some bias LLMs that fail to generate as much useful content for some groups of people not represented in the data. Also

of concern is the fact that people in some demographic groups may not have equal access to educational models of comparable quality levels. Weidinger et al. [179] shows the lack of LLM ability in generating content for groups whose languages are not selected for training. Oketunji et al. [134] argue that LLMs inherently generate bias, and propose a large language model bias index to quantify and address biases, enhancing the reliability of LLMs. Li et al. [104] introduced a systematic methodology to evaluate fairness and bias that could be displayed in LLMs, where a number of biased prompts are fed into LLMs and the probabilistic metrics for both individuals and groups that indicate the level of fairness are computed. In the domain of education, reinforced statements like "You should be unbiased for the sensitive feature (race or gender in experiments)" are helpful in mitigating the biased responses from LLMs. Chhikara et al. [25] show some gender bias from LLMs and explore possible solutions using few-shot learning as well as retrieval augmented generation. Caliskan et al. [18] examine social bias among scholars by evaluating LLMs (llama 2, Yi, Tulu, etc.) with various input prompts and argue that fine-tuning is the most effective approach to maintaining fairness. Li et al. [102] think that LLMs frequently present dominant viewpoints while ignoring alternative perspectives from minority parties (underrepresented in the training data), resulting in potential biases. They proposed a FAIRTHINKING pipeline to automatically generate roles that enable LLMs to articulate diverse perspectives for fair expressions. Li et al. [100] analyzes reasoning bias in decision-making systems of education and health care and devises a guided-debiasing framework that incorporates a prompt selection mechanism.

4.2   Reliability and Safety

LLMs have encountered reliability issues, including hallucinations, the production of toxic output, and inconsistencies in responses. These challenges are particularly significant in the educational sector. Hallucinations, where LLMs generate fictitious content, are a critical concern highlighted by Ji et al. [71]. Zhuo et al. [212] have outlined ethical considerations regarding the potential of LLMs to create content containing offensive language and explicit material. Cheng et al. [24] have discussed the issue of temporal misalignment in LLM data versions, introducing a novel tracer to track knowledge cutoffs. Shoaib et al. [157] underscore the risks of misinformation and disinformation through seemingly authentic content, suggesting the adoption of cyber-wellness education to boost public awareness and resilience. Liu et al. [112] explore the application of text-to-video models, like Sora, as tools to simulate real-world scenarios. They caution, however, that these models, despite their advanced capabilities, can sometimes lead to confusion or mislead students due to their limitations in accurately representing physical realism and complex spatial-temporal contexts. To improve the reliability of LLMs, Tan et al. [167] have developed a metacognitive strategy that enables LLMs to identify and correct errors autonomously. This method aims to detect inaccuracies with minimal human intervention and signals when adjustments are necessary. Additionally, the use of retrieval augmented generation (RAG) has been identified by Gao et al. and Zhao et al. as an effective way to address the issues of hallucinations and response inconsistencies, improving the reliability and accuracy of LLMs in content generation [54, 206].

---

[1]While there exist various educational applications, including those that assist with confusion helper, material creation, knowledge tracking, and content personalizing, we did not discuss them in this section due to the absence of publicly accessible datasets.

4.3 Transparency and Accountability

LLM, by design, runs as a black-box mechanism, so it comes with transparency and accountability concerns. Milano et al. [125] and BaHammam et al. [9] raise several challenges about LLMs to higher education, including plagiarism, inaccurate reporting, cheating in exams, as well as several other operational, financial, pedagogical issues. As a further thought of students using generative AI in homework or exams, Macneil et al. [117] discuss the respective impact to the traditional assessment methods, and think we educators should come up with new assessment framework to take into account the usage of Chat-GPT-like tools. Zhou et al. [210] brought up academic integrity ethical concerns that specifically confused teachers and students and calls for rethinking of policy-making. As specific measures, Gao et al. [52] introduce a novel concept called mixcase, representing a hybrid text form involving both machine-generated and human-generated and developed detectors that can distinguish human and machine texts. To tackle LLM ethical concerns in intellectual property violation, Huang et al. [68] propose incorporating citation while training LLMs, which could help enhance content transparency and verifiability. Finlayson, et al. [50] developed a systematic framework to efficiently discover the LLM's hidden size, obtaining full-vocabulary outputs, detecting and disambiguating different model updates, which could help users hold providers accountable by tracking model changes, thus enhancing the accountability.

4.4 Privacy and Security

Privacy and security protection have become increasingly important topics with the rise of LLMs, especially in the education sector where they deserve heightened scrutiny. Latham et al. [91] conducted a case study to explore the general public's perceptions of AI in education, revealing that while research has largely focused on the effectiveness of AI, critical areas such as learner awareness and acceptance of tracking and profiling algorithms remain underexplored. This underscores the need for more research into the ethical and legal aspects of AI in education. Das et al. [38] conducted an extensive review on the challenges of protecting personally identifiable information in the context of LLM use, highlighting widespread security and privacy concerns. Shoaib et al. [157] addressed the threats to personal privacy posed by deepfake content, proposing solutions like the use of detection algorithms and the implementation of standard protocols to bolster protection. Ke et al. [80]raised concerns about data privacy and the ethical implications of employing LLMs in psychological research, emphasizing the importance of safeguarding participant privacy in research projects. This highlights the necessity for researchers to understand the limitations of LLMs, adhere to ethical standards, and consider the potential consequences of their use. Suraworachet et al. [163] provided a comparative analysis of student information disclosure using LLMs versus traditional methods. Their findings point to challenges in valid evaluation, respecting privacy, and the absence of meaningful interactions in using LLMs to assess student performance. In terms of mitigation strategies, Hicke et al. [66] suggested frameworks that combine Retrieval-Augmented Generation (RAG) and fine-tuning techniques for enhanced privacy protection. Meanwhile, Masikisiki et al. [121] highlighted the significance of offering users the option
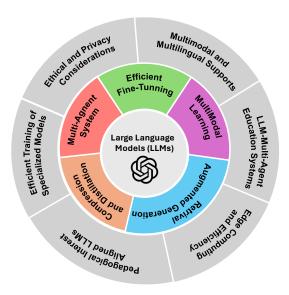
Figure 3: Future directions for LLMs in education.

to delete their interactions, emphasizing the importance of user control over personal data.

4.5 Overly Dependency on LLMs

Given the impressive performance of LLM generative ability, there is great concern that students blindly rely on LLMs for most of their work, leaving their ability to think independently disappearing. Milano et al. [125] discuss the problem of overreliance caused by Chat-GPT-like applications that students may use to compose essays and academic publications without improving their writing skills, which is essential to cultivate critical thinking. The concern might have even more impact on foreign-language students or students who are educationally disadvantaged while placing less emphasis on learning how to craft well-written texts. Krupp et al. [87] discuss the challenges of overreliance on the implications of LLMs in education and propose some moderated approaches to mitigate such effects. Similarly, Zuber et al. [213] discuss the risk that overreliance can bring to democracy and suggest cultivating thinking skills in children, fostering coherent thought formulation, and distinguishing between machine-generated output and genuine. They argue that LLMs should be used to augment but not substitute human thinking capacities. Adewumi et al. [5] also present scenarios that students tend to rely on LLMs for essay writing rather than writing their own, and demonstrates that using a probing-chain-of-thought tool could substantially stimulate critical thinking in the accompany with LLMs.

5 FUTURE DIRECTIONS

Here, we discuss the future opportunities for LLMs in education and summarize the promising directions in Figure 3. For each direction, we discuss the potential application of advanced LLM-based techniques and conclude their impacts on the future of education.

5.1 Pedagogical Interest Aligned LLMs

Although advanced LLMs like GPT-4 have demonstrated promising performance while being applied to experimental applications

Shen Wang[1*], Tianlong Xu[1*], Hang Li[2*], Chaoli Zhang[3], Joleen Liang[4],
Jiliang Tang[2], Philip S. Yu[5], Qingsong Wen[1†]

in education, applying LLMs directly in real-world instruction is still challenging, as delivering high-quality education is a complicated task that involves multi-discipline knowledge and administration constraints [126]. To solve these issues, future researchers in this direction could leverage advanced generation techniques like Retrieval-Augmented Generation (RAG) [55] to inform LLMs with the necessary prior information and guide LLMs to generate the pedagogical interest-aligned results. Apart from that, collecting large-size pedagogical instruction datasets from real-world instruction scenarios and fine-tuning existing LLMs to align with human instructors' behavior will also be an interesting direction to explore for future research. With learning from the preference of human instructors, LLMs could encoder the pedagogical restriction and knowledge patterns within LLMs' parameter space, and generate pedagogical interest-aligned results without too much intervention from the external information.

5.2 LLM-Multi-Agent Education Systems

The generous usage of LLMs in language comprehensive, reasoning, planning, and programming inspires works like AutoGen [181] in developing a collaboration framework that involves multiple LLMs to solve complicated tasks through conversation-format procedures. Similarly, problems in education commonly involve multi-step processing logic, which will be a good fit for using the multi-agent-based LLMs system. The recent work by Yang et al. [195] has demonstrated the great potential of the multi-agent framework for grading tasks. In this work, a human-like grading procedure is achieved by leveraging multiple LLM-based grader agents and critic agents, and the discrepancy errors made by individual judges are corrected through the group discussion procedure. For future research in this direction, more types of LLM-based agents can be included and their functions may spread from specific command executors to high-level plan makers. More importantly, human instructors, who are viewed as the special agents in the system, can also get involved with LLMs' interaction directly and provide any necessary interventions to the system flexibly.

5.3 Multimodal and Multilingual Supports

The high similarity between different human languages naturally enables LLMs to effectively support multilingual tasks. In addition to that, the recent findings of the alignment between multimodal and language tokens [182] further extend LLMs beyond textual analysis, venturing into multimodal learning analytics. By accepting diverse inputs, LLMs can take advantage of the mutual information between different data resources and provide advanced support to challenging tasks in education. For future works in multimodal direction, more attention could be placed on developing LLMs capable of interpreting and integrating these varied data sources, offering more nuanced insights into student engagement, comprehension, and learning styles. Such advances could pave the way for highly personalized and adaptive learning experiences that are tailored to the unique needs of each student. On the other hand, multilingual LLMs provide convenient access to quality global education resources for every individual with his/her proficient language. With developing robust models that not only translate, but also understand cultural nuances, colloquial expressions, and regional educational standards, researches in this direction would help learners around the world to benefit from LLMs in their native languages and significantly improve equity and inclusion in global education.

5.4 Edge Computing and Efficiency

Integrating LLMs with edge computing presents a promising avenue to enhance the efficiency and accessibility of educational technologies. By processing data closer to the end user, edge computing can reduce latency, increase content delivery speed, and enable offline access to educational resources. Future efforts could explore optimizing LLMs for edge deployment, focusing on lightweight models that maintain high performance while minimizing computational resources, which would be particularly beneficial in areas with limited internet connectivity, ensuring equitable access to educational tools. Additionally, processing data locally reduces the need to transmit sensitive information over the Internet, enhancing privacy and security. Edge computing could be a potential framework to utilize LLMs while adhering to stringent data protection standards.

5.5 Efficient Training of Specialized Models

The development of specialized LLMs tailored to specific educational domains or subjects represents a significant opportunity for future research. This direction involves creating models that not only grasp general language understanding, but also possess deep knowledge in fields such as mathematics, science, or literature. The point is that specialized LLMs could achieve a deep understanding of specific subjects, offering insights and support that are highly relevant and accurate but also more cost-effective. The challenge lies in the efficient training of these models, which requires innovations in data collection, model architecture, and training methodologies. Specialized models could offer more accurate and contextually relevant help, improving the educational experience for both students and educators.

5.6 Ethical and Privacy Considerations

Ethical and privacy considerations take center stage as LLMs become increasingly integrated into educational settings. Future research must address the responsible use of LLMs, including issues related to data security, student privacy, and bias mitigation. The development of frameworks and guidelines for ethical LLM deployment in education is crucial. This includes ensuring transparency in model training processes, safeguarding sensitive information, and creating inclusive models that reflect the diversity of the student population. Addressing these considerations is essential to build trust and ensure the responsible use of LLMs in education.

6 CONCLUSION

The rapid development of LLMs has revolutionized education. In this survey, we provide a comprehensive review of LLMs specifically applied for various educational scenarios from a multifaceted taxonomy, including student and teacher assistance, adaptive learning, and miscellaneous tools. In addition, we also summarize the related datasets and benchmarks, as well as current challenges and future directions. We hope that our survey can facilitate and inspire more innovative work within LLMs for education.

REFERENCES

[1] 2024. QuestionWell. https://www.questionwell.org/

[2] Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. 2023. Knowledge tracing: A survey. *Comput. Surveys* 55, 11 (2023), 1–37.

[3] Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. 2024. Knowledge Graphs as Context Sources for LLM-Based Explanations of Learning Recommendations. *arXiv preprint arXiv:2403.03008* (2024).

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[5] Tosin Adewumi, Lama Alkhaled, Claudia Buck, Sergio Hernandez, Saga Brilioth, Mkpe Kekung, Yelvin Ragimov, and Elisa Barney. 2023. ProCoT: Stimulating Critical Thinking and Writing of Students through Engagement with Large Language Models (LLMs). arXiv:2312.09801 [cs.CL]

[6] Formative AI. 2024. Formative AI. https://www.formative.com/

[7] Quizizz AI. 2024. Quizizz AI. https://quizizz.com/home/quizizz-ai

[8] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319* (2019).

[9] Ahmed S. BaHammam, Khaled Trabelsi, Seithikurippu R. Pandi-Perumal, and Hiatham Jahrami. 2023. Adapting to the Impact of AI in Scientific Writing: Balancing Benefits and Drawbacks while Developing Policies and Regulations. arXiv:2306.06699 [q-bio.OT]

[10] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Available at SSRN 4337484* (2023).

[11] Rishabh Balse, Viraj Kumar, Prajish Prasad, and Jayakrishnan Madathil Warriem. 2023. Evaluating the Quality of LLM-Generated Explanations for Logical Errors in CS1 Student Programs. In *Proceedings of the 16th Annual ACM India Compute Conference*. 49–54.

[12] Saba Batool, Junaid Rashid, Muhammad Wasif Nisar, Jungeun Kim, Hyuk-Yoon Kwon, and Amir Hussain. 2023. Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies* 28, 1 (2023), 905–971.

[13] Semere Kiros Bitew, Amir Hadifar, Lucas Sterckx, Johannes Deleu, Chris Develder, and Thomas Demeester. 2022. Learning to reuse distractors to support multiple choice question generation in education. *IEEE Transactions on Learning Technologies* (2022).

[14] Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series* 2013, 2 (2013), i–15.

[15] Michael Bommarito II and Daniel Martin Katz. 2022. GPT takes the bar exam. *arXiv preprint arXiv:2212.14402* (2022).

[16] Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. 42–47.

[17] Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*. 52–75.

[18] Chahat Raj1 Anjishnu Mukherjee1 Aylin Caliskan and Antonios Anastasopoulos1 Ziwei Zhu. [n. d.]. A Psychological View to Social Bias in LLMs: Evaluation and Mitigation. ([n. d.]).

[19] Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. LearningQ: a large-scale dataset for educational question generation. In *Proceedings of the international AAAI conference on web and social media*, Vol. 12.

[20] Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial intelligence in education: A review. *IEEE Access* 8 (2020), 75264–75278.

[21] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588* (2022).

[22] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. Theoremqa: A theorem-driven question answering dataset. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[23] Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. 2023. Exploring the potential of large language models (llms) in learning on graphs. *arXiv preprint arXiv:2307.03393* (2023).

[24] Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated Data: Tracing Knowledge Cutoffs in Large Language Models. arXiv:2403.12958 [cs.CL]

[25] Garima Chhikara, Anurag Sharma, Kripabandhu Ghosh, and Abhijnan Chakraborty. 2024. Few-Shot Fairness: Unveiling LLM's Potential for Fairness-Aware Classification. arXiv:2402.18502 [cs.CL]

[26] Thomas KF Chiu, Qi Xia, Xinyan Zhou, Ching Sing Chai, and Miaoting Cheng. 2023. Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence* 4 (2023), 100118.

[27] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457* (2018).

[28] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[29] Rianne Conijn, Chris Snijders, Ad Kleingeld, and Uwe Matzat. 2016. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies* 10, 1 (2016), 17–29.

[30] Conker. 2024. Conker. https://www.conker.ai/

[31] Grant Cooper, Amanda Berry, and James Baglin. 2020. Demographic predictors of students' science participation over the age of 16: An Australian case study. *Research in Science Education* 50, 1 (2020), 361–373.

[32] Education Copilot. 2024. Education Copilot. https://educationcopilot.com/

[33] Copy.ai. 2024. Copy.ai. https://www.copy.ai/

[34] Teodor-Mihai Cotet, Stefan Ruseti, and Mihai Dascalu. 2020. Neural grammatical error correction for romanian. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 625–631.

[35] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092* (2023).

[36] Xiliang Cui and Bao-lin Zhang. 2011. The principles for building the "international corpus of learner chinese". *Applied Linguistics* 2 (2011), 100–108.

[37] Curipod. 2024. Curipod. http://www.curipod.com/ai

[38] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2024. Security and Privacy Challenges of Large Language Models: A Survey. arXiv:2402.00888 [cs.CL]

[39] Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H Sanchez Gutierrez, and Kenji Sagae. 2020. Developing NLP tools with a new corpus of learner Spanish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 7238–7243.

[40] Paul Denny, James Prather, Brett A Becker, James Finnie-Ansley, Arto Hellas, Juho Leinonen, Andrew Luxton-Reilly, Brent N Reeves, Eddie Antonio Santos, and Sami Sarsa. 2024. Computing education in the era of generative AI. *Commun. ACM* 67, 2 (2024), 56–67.

[41] David G Dewhurst, Hamish A MacLeod, and Tracey AM Norris. 2000. Independent student learning aided by computers: an acceptable alternative to lectures? *Computers & Education* 35, 3 (2000), 223–241.

[42] Diffit. 2024. Diffit. https://beta.diffit.me/

[43] Tung Do Viet and Konstantin Markov. 2023. Using Large Language Models for Bug Localization and Fixing. In *2023 12th International Conference on Awareness Science and Technology (iCAST)*. IEEE, 192–197.

[44] Jacob Doughty, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, 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*. 114–123.

[45] Hanyue Du, Yike Zhao, Qingyuan Tian, Jiani Wang, Lei Wang, Yunshi Lan, and Xuesong Lu. 2023. FlaCGEC: A Chinese Grammatical Error Correction Dataset with Fine-grained Linguistic Annotation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 5321–5325.

[46] Eduaide.ai. 2024. Eduaide.ai. https://www.eduaide.ai/

[47] Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046* (2023).

[48] Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. Grammargpt: Exploring open-source llms for native chinese grammatical error correction with supervised fine-tuning. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 69–80.

[49] Gianni Fenu, Roberta Galici, and Mirko Marras. 2022. Experts' view on challenges and needs for fairness in artificial intelligence for education. In *International Conference on Artificial Intelligence in Education*. Springer, 243–255.

[50] Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. 2024. Logits of API-Protected LLMs Leak Proprietary Information. arXiv:2403.09539 [cs.CL]

[51] Hiroaki Funayama, Yuya Asazuma, Yuichiroh Matsubayashi, Tomoya Mizumoto, and Kentaro Inui. 2023. Reducing the cost: Cross-prompt pre-finetuning for short answer scoring. In *International conference on artificial intelligence in education*. Springer, 78–89.

Shen Wang[1*], Tianlong Xu[1*], Hang Li[2*], Chaoli Zhang[3], Joleen Liang[4],
Jiliang Tang[2], Philip S. Yu[5], Qingsong Wen[1†]

[52] Chujie Gao, Dongping Chen, Qihui Zhang, Yue Huang, Yao Wan, and Lichao Sun. 2024. LLM-as-a-Coauthor: The Challenges of Detecting LLM-Human Mixcase. arXiv:2401.05952 [cs.CL]

[53] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*. PMLR, 10764–10799.

[54] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL]

[55] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

[56] Parlay Genie. 2024. Parlay Genie. https://parlayideas.com/

[57] Huanli Gong, Liangming Pan, and Hengchang Hu. 2022. Khanq: A dataset for generating deep questions in education. In *Proceedings of the 29th International Conference on Computational Linguistics*. 5925–5938.

[58] Google. 2024. Google Bard. https://bard.google.com

[59] gotFeedback. 2024. gotFeedback. https://www.gotlearning.com/gotfeedback/

[60] Grammarly. 2024. Grammarly. https://app.grammarly.com/

[61] Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The wiked error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *Advances in Natural Language Processing: 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings 9*. Springer, 478–490.

[62] Qi Guo, Junming Cao, Xiaofei Xie, Shangqing Liu, Xiaohong Li, Bihuan Chen, and Xin Peng. 2024. Exploring the potential of chatgpt in automated code refinement: An empirical study. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–13.

[63] Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. Eduqg: A multi-format multiple-choice dataset for the educational domain. *IEEE Access* 11 (2023), 20885–20896.

[64] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[65] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* (2021).

[66] Yann Hicke, Anmol Agarwal, Qianou Ma, and Paul Denny. 2023. AI-TA: Towards an Intelligent Question-Answer Teaching Assistant using Open-Source LLMs. arXiv:2311.02775 [cs.LG]

[67] Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 887–896.

[68] Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A Key to Building Responsible and Accountable Large Language Models. arXiv:2307.02185 [cs.CL]

[69] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems* 36 (2024).

[70] Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies* (2023), 1–20.

[71] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[72] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.

[73] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position Paper: What Can Large Language Models Tell Us about Time Series Analysis. *arXiv preprint arXiv:2402.02713* (2024).

[74] Breanna Jury, Angela Lorusso, Juho Leinonen, Paul Denny, and Andrew Luxton-Reilly. 2024. Evaluating LLM-generated Worked Examples in an Introductory Programming Course. In *Proceedings of the 26th Australasian Computing Education Conference*. 77–86.

[75] René Just, Darioush Jalali, and Michael D Ernst. 2014. Defects4J: A database of existing faults to enable controlled testing studies for Java programs. In *Proceedings of the 2014 international symposium on software testing and analysis*. 437–440.

[76] Md Rayhan Kabir and Fuhua Lin. 2023. An LLM-Powered Adaptive Practicing System. In *AIED 2023 workshop on Empowering Education with LLMs-the Next-Gen Interface and Content Generation, AIED*.

[77] Firuz Kamalov and Ikhlaas Gurrib. 2023. A new era of artificial intelligence in education: A multifaceted revolution. *arXiv preprint arXiv:2305.18303* (2023).

[78] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[79] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara Jane Ericson, David Weintrop, and Tovi Grossman. 2023. How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*. 1–12.

[80] Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. 2024. Exploring the Frontiers of LLMs in Psychological Applications: A Comprehensive Review. arXiv:2401.01519 [cs.LG]

[81] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 235–251.

[82] Khanmigo. 2024. Khanmigo. https://blog.khanacademy.org/teacher-khanmigo/

[83] Daesik Kim, Seonhoon Kim, and Nojun Kwak. 2018. Textbook question answering with multi-modal context graph understanding and self-supervised open-set comprehension. *arXiv preprint arXiv:1811.00232* (2018).

[84] Kenneth R Koedinger, Sidney D'Mello, Elizabeth A McLaughlin, Zachary A Pardos, and Carolyn P Rosé. 2015. Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science* 6, 4 (2015), 333–353.

[85] Claudia M König, Christin Karrenbauer, and Michael H Breitner. 2023. Critical success factors and challenges for individual digital study assistants in higher education: A mixed methods analysis. *Education and Information Technologies* 28, 4 (2023), 4475–4503.

[86] Osama Koraishi. 2023. Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education and Technology* 3, 1 (2023).

[87] Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2023. Challenges and Opportunities of Moderating Usage of Large Language Models in Education. arXiv:2312.14969 [cs.HC]

[88] Danijel Kučak, Vedran Juričić, and Goran Đambić. 2018. MACHINE LEARNING IN EDUCATION-A SURVEY OF CURRENT RESEARCH TRENDS. *Annals of DAAAM & Proceedings* 29 (2018).

[89] Bor-Chen Kuo, Frederic TY Chang, and Zong-En Bai. 2023. Leveraging LLMs for Adaptive Testing and Learning in Taiwan Adaptive Learning Platform (TALP). (2023).

[90] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683* (2017).

[91] A. Latham and S. Goltz. 2019. A Survey of the General Public's Views on the Ethics of Using AI in Education. In *Artificial Intelligence in Education (Lecture Notes in Computer Science, Vol. 11625)*, S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, and R. Luckin (Eds.). Springer, Cham. https://doi.org/10.1007/978-3-030-23204-7_17

[92] Ehsan Latif, Gengchen Mai, Matthew Nyaaba, Xuansheng Wu, Ninghao Liu, Guoyu Lu, Sheng Li, Tianming Liu, and Xiaoming Zhai. 2023. Artificial general intelligence (AGI) for education. *arXiv preprint arXiv:2304.12479* (2023).

[93] Claire Le Goues, Neal Holtschulte, Edward K Smith, Yuriy Brun, Premkumar Devanbu, Stephanie Forrest, and Westley Weimer. 2015. The ManyBugs and IntroClass benchmarks for automated repair of C programs. *IEEE Transactions on Software Engineering* 41, 12 (2015), 1236–1256.

[94] Unggi Lee, Haewon Jung, Younghoon Jeon, Younghoon Sohn, Wonhee Hwang, Jewoong Moon, and Hyeoncheol Kim. 2023. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies* (2023), 1–33.

[95] Unggi Lee, Sungjun Yoon, Joon Seo Yun, Kyoungsoo Park, YoungHoon Jung, Damji Stratton, and Hyeoncheol Kim. 2023. Difficulty-Focused Contrastive Learning for Knowledge Tracing with a Large Language Model-Based Difficulty Prediction. *arXiv preprint arXiv:2312.11890* (2023).

[96] Daniel Leiker, Sara Finnigan, Ashley Ricker Gyllen, and Mutlu Cukurova. 2023. Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale. *arXiv preprint arXiv:2306.01815* (2023).

[97] Hai Li, Chenglu Li, Wanli Xing, Sami Baral, and Neil Heffernan. 2024. Automated Feedback for Student Math Responses Based on Multi-Modality and Fine-Tuning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 763–770.

[98] Hang Li, Tianlong Xu, Chaoli Zhang, Eason Chen, Jing Liang, Xing Fan, Haoyang Li, Jiliang Tang, and Qingsong Wen. 2024. Bringing Generative AI to Adaptive Learning in Education. *arXiv preprint arXiv:2402.14601* (2024).

[99] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212* (2023).

[100] Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. 2024. Steering LLMs Towards Unbiased Responses: A Causality-Guided Debiasing Framework. arXiv:2403.08743 [cs.CL]

[101] Qingyao Li, Lingyue Fu, Weiming Zhang, Xianyu Chen, Jingwei Yu, Wei Xia, Weinan Zhang, Ruiming Tang, and Yong Yu. 2023. Adapting Large Language Models for Education: Foundational Capabilities, Potentials, and Challenges. *arXiv preprint arXiv:2401.08664* (2023).

[102] Tianlin Li, Xiaoyu Zhang, Chao Du, Tianyu Pang, Qian Liu, Qing Guo, Chao Shen, and Yang Liu. 2024. Your Large Language Model is Secretly a Fairness Proponent and You Should Prompt it Like One. arXiv:2402.12150 [cs.CL]

[103] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024. A Survey on Fairness in Large Language Models. arXiv:2308.10149 [cs.CL]

[104] Yunqi Li and Yongfeng Zhang. 2023. Fairness of ChatGPT. arXiv:2305.18569 [cs.LG]

[105] Zhiyu Li, Shuai Lu, Daya Guo, Nan Duan, Shailesh Jannu, Grant Jenks, Deep Majumder, Jared Green, Alexey Svyatkovskiy, Shengyu Fu, et al. 2022. Automating code review activities by large-scale pre-training. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering.* 1035–1047.

[106] Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C Lee Giles. 2018. Distractor generation for multiple choice questions using learning to rank. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications.* 284–290.

[107] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2023. Can large language models reason about medical questions? *Patterns* (2023).

[108] Derrick Lin, James Koppel, Angela Chen, and Armando Solar-Lezama. 2017. QuixBugs: A multi-lingual program repair benchmark set based on the Quixey Challenge. In *Proceedings Companion of the 2017 ACM SIGPLAN international conference on systems, programming, languages, and applications: software for humanity.* 55–56.

[109] Yuanguo Lin, Hong Chen, Wei Xia, Fan Lin, Pengcheng Wu, Zongyue Wang, and Yong Li. 2023. A comprehensive survey on deep learning techniques in educational data mining. *arXiv preprint arXiv:2309.04761* (2023).

[110] Min Liu, Emily McKelroy, Stephanie B Corliss, and Jamison Carrigan. 2017. Investigating the effect of an adaptive learning intervention on students' learning. *Educational technology research and development* 65 (2017), 1605–1625.

[111] Tiaoqiao Liu, Wenbiao Ding, Zhiwei Wang, Jiliang Tang, Gale Yan Huang, and Zitao Liu. 2019. Automatic short answer grading via multiway attention networks. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part II 20.* Springer, 169–173.

[112] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. 2024. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. arXiv:2402.17177 [cs.CV]

[113] Zi-Yu Liu, Zaffar Ahmed Shaikh, and Farida Gazizova. 2020. Using the Concept of Game-Based Learning in Education. *International Journal of Emerging Technologies in Learning* (2020).

[114] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165* (2021).

[115] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610* (2022).

[116] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214* (2021).

[117] Stephen MacNeil, Scott Spurlock, and Ian Applebaum. 2024. Imagining Computing Education Assessment after Generative AI. arXiv:2401.04601 [cs.HC]

[118] Setareh Maghsudi, Andrew Lan, Jie Xu, and Mihaela van Der Schaar. 2021. Personalized education in the artificial intelligence era: what to expect next. *IEEE Signal Processing Magazine* 38, 3 (2021), 37–50.

[119] MagicSchool.ai. 2024. MagicSchool.ai. https://www.magicschool.ai/

[120] Kamil Malinka, Martin Peresíni, Anton Firc, Ondrej Hujnák, and Filip Janus. 2023. On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree?. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1.* 47–53.

[121] Baphumelele Masikisiki, Vukosi Marivate, and Yvette Hlope. 2023. Investigating the Efficacy of Large Language Models in Reflective Assessment Methods through Chain of Thoughts Prompting. arXiv:2310.00272 [cs.CL]

[122] Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2021. A diverse corpus for evaluating and developing English math word problem solvers. *arXiv preprint arXiv:2106.15772* (2021).

[123] Microsoft. 2024. Bing Chat. https://learn.microsoft.com/en-us/training/modules/enhance-teaching-learning-bing-chat/

[124] Microsoft. 2024. *What is responsible AI?* Accessed: 2024-02-08.

[125] S. Milano, J.A. McGrane, and S. Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence* 5 (2023), 333–334. https://doi.org/10.1038/s42256-023-00644-2

[126] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence* 5, 4 (2023), 333–334.

[127] Reza Hadi Mogavi, Chao Deng, Justin Juho Kim, Pengyuan Zhou, Young D Kwon, Ahmed Hosny Saleh Metwally, Ahmed Tlili, Simone Bassanelli, Antonio Bucchiarone, Sujit Gujar, et al. 2023. Exploring user perspectives on chatgpt: Applications, perceptions, and implications for ai-integrated education. *arXiv preprint arXiv:2305.13114* (2023).

[128] Mary E Morningstar, Jennifer A Kurth, and Paul E Johnson. 2017. Examining national trends in educational placements for students with significant disabilities. *Remedial and Special Education* 38, 1 (2017), 3–12.

[129] Jakub Náplava, Milan Straka, Jana Straková, and Alexandr Rosen. 2022. Czech grammar error correction with a large and diverse corpus. *Transactions of the Association for Computational Linguistics* 10 (2022), 452–467.

[130] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. 2019. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091* (2019).

[131] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the eighteenth conference on computational natural language learning: shared task.* 1–14.

[132] Lin Ni, Sijie Wang, Zeyu Zhang, Xiaoxuan Li, Xianda Zheng, Paul Denny, and Jiamou Liu. 2023. Enhancing student performance prediction on learnersourced questions with sgnn-llm synergy. *arXiv preprint arXiv:2309.13500* (2023).

[133] Nolej. 2024. Nolej. https://nolej.io/

[134] Abiodun Finbarrs Oketunji, Muhammad Anas, and Deepthi Saina. 2023. Large Language Model (LLM) Bias Index–LLMBI. *arXiv preprint arXiv:2312.14769* (2023).

[135] OpenAI. 2024. OpenAI Chat. https://chat.openai.com

[136] R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article* 2, 5 (2023).

[137] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[138] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning.* PMLR, 248–260.

[139] Zachary A Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871* (2023).

[140] Chat PDF. 2024. Chat PDF. https://www.chatpdf.com/

[141] Perplexity AI. 2024. Perplexity AI. https://perplexity.ai

[142] Pi.ai. 2024. Pi. http://www.Pi.ai

[143] Gustavo Pinto, Isadora Cardoso-Pereira, Danilo Monteiro, Danilo Lucena, Alberto Souza, and Kiev Gama. 2023. Large language models for education: Grading open-ended questions using chatgpt. In *Proceedings of the XXXVII Brazilian Symposium on Software Engineering.* 293–302.

[144] Ethan Prihar, Morgan Lee, Mia Hopman, Adam Tauman Kalai, Sofia Vempala, Allison Wang, Gabriel Wickline, Aly Murray, and Neil Heffernan. 2023. Comparing different approaches to generating mathematics explanations using large language models. In *International Conference on Artificial Intelligence in Education.* Springer, 290–295.

[145] Junaid Qadir. 2023. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON).* IEEE, 1–9.

[146] Md Mostafizer Rahman and Yutaka Watanobe. 2023. ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences* 13, 9 (2023), 5783.

[147] Adams Ray, Wu Margaret, et al. 2003. *PISA Programme for international student assessment (PISA) PISA 2000 technical report: PISA 2000 technical report.* oecd Publishing.

[148] Cristobal Romero and Sebastian Ventura. 2007. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications* 33, 1 (2007), 135–146.

[149] Cristóbal Romero and Sebastián Ventura. 2010. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews)* 40, 6 (2010), 601–618.

Shen Wang[1*], Tianlong Xu[1*], Hang Li[2*], Chaoli Zhang[3], Joleen Liang[4],
Jiliang Tang[2], Philip S. Yu[5], Qingsong Wen[1†]

[150] Cristobal Romero and Sebastian Ventura. 2013. Data mining in education. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery* 3, 1 (2013), 12–27.

[151] Cristobal Romero and Sebastian Ventura. 2020. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 10, 3 (2020), e1355.

[152] Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know Your Audience: Do LLMs Adapt to Different Age and Education Levels? *arXiv preprint arXiv:2312.02065* (2023).

[153] Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. *arXiv preprint arXiv:2106.03830* (2021).

[154] Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics* 7 (2019), 1–17.

[155] José A Ruipérez-Valiente, Thomas Staubitz, Matt Jenner, Sherif Halawa, Jiayin Zhang, Ignacio Despujol, Jorge Maldonado-Mahauad, German Montoro, Melanie Peffer, Tobias Rohloff, et al. 2022. Large scale analytics of global and regional MOOC providers: Differences in learners' demographics, preferences, and perceptions. *Computers & Education* 180 (2022), 104426.

[156] Jaromir Savelka, Arav Agarwal, Marshall An, Chris Bogart, and Majd Sakr. 2023. Thrilled by your progress! Large language models (GPT-4) no longer struggle to pass assessments in higher education programming courses. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*. 78–92.

[157] Mohamed R. Shoaib, Zefan Wang, Milad Taleby Ahvanooey, and Jun Zhao. 2023. Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. arXiv:2311.17394 [cs.CR]

[158] Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. [n. d.]. Automatic Generation of Socratic Questions for Learning to Solve Math Word Problems. ([n. d.]).

[159] Shashank Sonkar and Richard G Baraniuk. 2023. Deduction under Perturbed Evidence: Probing Student Simulation (Knowledge Tracing) Capabilities of Large Language Models. (2023).

[160] Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*. 1501–1510.

[161] summarize.tech. 2024. summarize.tech. https://www.summarize.tech/

[162] Hao Sun. 2023. Offline prompt evaluation and optimization with inverse reinforcement learning. *arXiv preprint arXiv:2309.06553* (2023).

[163] Wannapon Surawrachet, Jennifer Seon, and Mutlu Cukurova. 2024. Predicting challenge moments from students' discourse: A comparison of GPT-4 to two traditional natural language processing approaches. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)*. ACM. https://doi.org/10.1145/3636555.3636905

[164] Teo Susnjak. 2022. ChatGPT: The end of online exam integrity? *arXiv preprint arXiv:2212.09292* (2022).

[165] Oleksiy Syvokon and Olena Nahorna. 2021. UA-GEC: Grammatical error correction and fluency corpus for the ukrainian language. *arXiv preprint arXiv:2103.16997* (2021).

[166] Kehui Tan, Tianqi Pang, and Chenyou Fan. 2023. Towards applying powerful large ai models in classroom teaching: Opportunities, challenges and prospects. *arXiv preprint arXiv:2305.03433* (2023).

[167] Zhen Tan, Jie Peng, Tianlong Chen, and Huan Liu. 2024. Tuning-Free Accountable Intervention for LLM Deployment – A Metacognitive Approach. arXiv:2403.05636 [cs.AI]

[168] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[169] Jörg Tiedemann. 2020. The Tatoeba Translation Challenge–Realistic Data Sets for Low Resource and Multilingual MT. *arXiv preprint arXiv:2010.06354* (2020).

[170] Maria Tigina, Anastasiia Birillo, Yaroslav Golubev, Hieke Keuning, Nikolay Vyahhi, and Timofey Bryksin. 2023. Analyzing the quality of submissions in online programming courses. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*. IEEE, 271–282.

[171] Goblin Tools. 2024. Goblin Tools. https://goblin.tools/

[172] Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*. 32–37.

[173] Michele Tufano, Cody Watson, Gabriele Bavota, Massimiliano Di Penta, Martin White, and Denys Poshyvanyk. 2019. An empirical study on learning bug-fixing patches in the wild via neural machine translation. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 28, 4 (2019), 1–29.

[174] Twee. 2024. Twee. https://twee.com/

[175] Shyam Upadhyay and Ming-Wei Chang. 2016. Annotating derivations: A new evaluation strategy and dataset for algebra word problems. *arXiv preprint arXiv:1609.07197* (2016).

[176] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

[177] Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. 845–854.

[178] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[179] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. arXiv:2112.04359 [cs.CL]

[180] Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209* (2017).

[181] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155* (2023).

[182] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519* (2023).

[183] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[184] Yiran Wu, Feiran Jia, Shaokun Zhang, Qingyun Wu, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, and Chi Wang. 2023. An empirical study on challenging math problem solving with gpt-4. *arXiv preprint arXiv:2306.01337* (2023).

[185] Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. *arXiv preprint arXiv:2401.06431* (2024).

[186] Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 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 (BEA 2023)*. 610–625.

[187] Zhang Xiong, Haoxuan Li, Zhuang Liu, Zhuofan Chen, Hao Zhou, Wenge Rong, and Yuanxin Ouyang. 2024. A Review of Data Mining in Personalized Education: Current Trends and Future Prospects. *arXiv preprint arXiv:2402.17236* (2024).

[188] Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020* (2023).

[189] Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. FCGEC: fine-grained corpus for Chinese grammatical error correction. *arXiv preprint arXiv:2210.12364* (2022).

[190] Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic Questions and Where to Find Them: FairytaleQA–An Authentic Dataset for Narrative Comprehension. *arXiv preprint arXiv:2203.13947* (2022).

[191] Gautam Yadav, Ying-Jui Tseng, and Xiaolin Ni. 2023. Contextualizing problems to student interests at scale in intelligent tutoring system using large language models. *arXiv preprint arXiv:2306.00190* (2023).

[192] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* 55, 1 (2024), 90–112.

[193] Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. 576–584.

[194] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031* (2023).

[195] Kaiqi Yang, Yuchen Chu, Taylor Darwin, Ahreum Han, Hang Li, Hongzhi Wen, Yasemin Copur-Gencturk, Jiliang Tang, and Hui Liu. 2024. Content Knowledge Identification with Multi-Agent Large Language Models (LLMs). In *International Conference on Artificial Intelligence in Education*. Springer.

[196] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 180–189.

[197] Ke Yuan, Dafang He, Zhuoren Jiang, Liangcai Gao, Zhi Tang, and C Lee Giles. 2020. Automatic generation of headlines for online math questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9490–9497.

[198] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do Large Language Models perform in Arithmetic tasks? *arXiv preprint arXiv:2304.02015* (2023).

[199] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. 2023. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226* (2023).

[200] Jialu Zhang, José Cambronero, Sumit Gulwani, Vu Le, Ruzica Piskac, Gustavo Soares, and Gust Verbruggen. 2022. Repairing bugs in python assignments using large language models. *arXiv preprint arXiv:2209.14876* (2022).

[201] Peng Zhang and Gemma Tur. 2023. A systematic review of ChatGPT use in K-12 education. *European Journal of Education* (2023).

[202] Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[203] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the Performance of Large Language Models on GAOKAO Benchmark.

[204] Xiaowu Zhang, Xiaotian Zhang, Cheng Yang, Hang Yan, and Xipeng Qiu. 2023. Does Correction Remain An Problem For Large Language Models? *arXiv preprint arXiv:2308.01776* (2023).

[205] Honghong Zhao, Baoxin Wang, Dayong Wu, Wanxiang Che, Zhigang Chen, and Shijin Wang. 2022. Overview of ctc 2021: Chinese text correction for native speakers. *arXiv preprint arXiv:2208.05681* (2022).

[206] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv:2402.19473 [cs.CV]

[207] Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. Ape210k: A large-scale and template-rich dataset of math word problems. *arXiv preprint arXiv:2009.11506* (2020).

[208] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364* (2023).

[209] Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, et al. 2024. Solving Challenging Math Word Problems Using GPT-4 Code Interpreter with Code-based Self-Verification. In *The International Conference on Learning Representations*.

[210] Kyrie Zhixuan Zhou, Zachary Kilhoffer, Madelyn Rose Sanfilippo, Ted Underwood, Ece Gumusel, Mengyi Wei, Abhinav Choudhry, and Jinjun Xiong. 2024. "The teachers are confused as well": A Multiple-Stakeholder Ethics Discussion on Large Language Models in Computing Education. arXiv:2401.12453 [cs.CY]

[211] Zihao Zhou, Maizhen Ning, Qiufeng Wang, Jie Yao, Wei Wang, Xiaowei Huang, and Kaizhu Huang. 2023. Learning by Analogy: Diverse Questions Generation in Math Word Problem. In *Findings of the Association for Computational Linguistics: ACL 2023*. 11091–11104.

[212] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867* (2023).

[213] Niina Zuber and Jan Gogoll. 2023. Vox Populi, Vox ChatGPT: Large Language Models, Education and Democracy. arXiv:2311.06207 [cs.CY]

A APPENDIX

Table 1 summarizes the commonly used publicly available datasets and benchmarks for evaluating LLMs on education applications.

Shen Wang[1*], Tianlong Xu[1*], Hang Li[2*], Chaoli Zhang[3], Joleen Liang[4],
Jiliang Tang[2], Philip S. Yu[5], Qingsong Wen[1†]

Table 1: Summary of existing datasets and benchmarks in the area of LLMs for education.

Dataset&Benchmark	App	User	Subject	Level	Language	Modality	Amount	Source
GSM8K	QS	student	math	K-12	EN	text	8.5K	[28]
MATH	QS	student	math	K-12	EN	text	12.5K	[65]
Dolphin18K	QS	student	math	K-12	EN	text	18K	[67]
DRAW-1K	QS	student	math	comprehensive	EN	text	1K	[175]
Math23K	QS	student	math	K-12	ZH	text	23K	[177]
Ape210K	QS	student	math	K-12	EN, ZH	text	210K	[207]
MathQA	QS	student	math	K-12	EN	text	37K	[8]
ASDiv	QS	student	math	K-12	EN	text & image	2K	[122]
IconQA	QS	student	math	K-12	EN	text & table	107K	[116]
TQA	QS	student	science	K-12	EN	text & image	26K	[83]
Geometry3K	QS	student	geometry	K-12	EN	text & image	3K	[114]
AI2D	QS	student	science	K-12	EN	text & image	5K	[81]
SCIENCEQA	QS	student	science	K-12	EN	text & image	21K	[22]
MedQA	QS	student	medicine	professional	EN	text	40K	[72]
MedMCQA	QS	student	medicine	professional	EN	text	200K	[138]
TheoremQA	QS	student	science	college	EN	text	800	[22]
Math-StackExchange	QS	student	math	comprehensive	EN	text	310K	[197]
TABMWP	QS	student	math	K-12	EN	text	38K	[115]
ARC	QS	student	comprehensive	comprehensive	EN	text	7.7K	[27]
C-Eva	QS	student	comprehensive	comprehensive	ZH	text	13.9K	[69]
GAOKAO-bench	QS	student	comprehensive	comprehensive	ZH	text	2.8K	[203]
AGIEval	QS	student	comprehensive	comprehensive	EN, ZH	text	8k	[208]
MMLU	QS	student	comprehensive	comprehensive	EN	text	1.8K	[64]
CMMLU	QS	student	comprehensive	comprehensive	ZH	text	11.9K	[99]
SuperCLUE	QS	student	comprehensive	comprehensive	ZH	text	15.9K	[188]
LANG-8	EC	student	linguistic	language training	Multi	text	1M	[153]
CLANG-8	EC	student	linguistic	language training	Multi	text	2.6M	[153]
CoNLL-2014	EC	student	linguistic	language training	EN	text	58k	[131]
BEA-2019	EC	student	linguistic	language training	EN	text	686K	[17]
SIGHAN	EC	student	linguistic	language training	ZH	text	12K	[172]
CTC	EC	student	linguistic	language training	ZH	text	218K	[205]
FCGEC	EC	student	linguistic	language training	ZH	text	41K	[189]
FlaCGEC	EC	student	linguistic	language training	ZH	text	13K	[45]
GECCC	EC	student	linguistic	language training	CS	text	83K	[129]
RULEC-GEC	EC	student	linguistic	language training	RU	text	12K	[154]
Falko-MERLIN	EC	student	linguistic	language training	GE	text	24K	[61]
COWS-L2H	EC	student	linguistic	language training	ES	text	12K	[39]
UA-GEC	EC	student	linguistic	language training	UK	text	20K	[165]
RONACC	EC	student	linguistic	language training	RO	text	10K	[34]
Defects4J	EC	student	computer science	professional	EN & Java	text& code	357	[75]
ManyBugs	EC	student	computer science	professional	EN & C	text& code	185	[93]
IntroClass	EC	student	computer science	professional	EN & C	text& code	998	[93]
QuixBugs	EC	student	computer science	professional	EN & multi	text& code	40	[108]
Bugs2Fix	EC	student	computer science	professional	EN & Java	text& code	2.3M	[173]
CodeReview	EC	student	computer science	professional	EN & Multi	text& code	642	[105]
CodeReview-New	EC	student	computer science	professional	EN & Multi	text& code	15	[62]
SciQ	QG	teacher	science	MOOC	EN	text	13.7K	[180]
RACE	QG	teacher	linguistic	K-12	EN	text	100K	[90]
FairytaleQA	QG	teacher	literature	K-12	EN	text	10K	[190]
LearningQ	QG	teacher	comprehensive	MOOC	EN	text	231K	[19]
KHANQ	QG	teacher	science	MOOC	EN	text	1K	[57]
EduQG	QG	teacher	comprehensive	MOOC	EN	text	3K	[63]
MCQL	QG	teacher	comprehensive	MOOC	EN	text	7.1K	[106]
Televic	QG	teacher	comprehensive	MOOC	EN	text	62K	[13]
CLC-FCE	AG	teacher	linguistic	standardized test	EN	text	1K	[196]
ASAP	AG	teacher	linguistic	K-12	EN	text	17K	[170]
TOEFL11	AG	teacher	linguistic	standardized test	EN	text	1K	[14]
ICLE	AG	teacher	linguistic	standardized test	EN	text	4K	[160]
HSK	AG	teacher	linguistic	standardized test	ZH	text	10K	[36]