

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/367541637>

ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education

Article · January 2023

DOI: 10.1016/j.lindif.2023.102274

CITATIONS

312

READS

11,238

23 authors, including:



Kathrin Seßler

Technische Universität München

6 PUBLICATIONS 416 CITATIONS

[SEE PROFILE](#)



Stefan Kuechemann

Ludwig-Maximilians-University of Munich

82 PUBLICATIONS 1,558 CITATIONS

[SEE PROFILE](#)



Maria Bannert

Technische Universität München

117 PUBLICATIONS 5,060 CITATIONS

[SEE PROFILE](#)



Daryna Dementieva

Technische Universität München

23 PUBLICATIONS 364 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



FAMULUS: Fostering diagnostic competence in medical and teacher education via adaptive online-case-simulations [View project](#)



Teaching the rectal examination [View project](#)

ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education

Enkelejda Kasneci^{1*}, Kathrin Sessler¹, Stefan Küchemann², Maria Bannert¹, Daryna Dementieva¹, Frank Fischer², Urs Gasser¹, Georg Groh¹, Stephan Günemann¹, Eyke Hüllermeier², Stephan Krusche¹, Gitta Kutyniok², Tilman Michaeli¹, Claudia Nerdel¹, Jürgen Pfeffer¹, Oleksandra Poquet¹, Michael Sailer², Albrecht Schmidt², Tina Seidel¹, Matthias Stadler², Jochen Weller², Jochen Kuhn², Gjergji Kasneci³

Abstract

Large language models represent a significant advancement in the field of AI. The underlying technology is key to further innovations and, despite critical views and even bans within communities and regions, large language models are here to stay. This position paper presents the potential benefits and challenges of educational applications of large language models, from student and teacher perspectives. We briefly discuss the current state of large language models and their applications. We then highlight how these models can be used to create educational content, improve student engagement and interaction, and personalize learning experiences. With regard to challenges, we argue that large language models in education require teachers and learners to develop sets of competencies and literacies necessary to both understand the technology as well as their limitations and unexpected brittleness of such systems. In addition, a clear strategy within educational systems and a clear pedagogical approach with a strong focus on critical thinking and strategies for fact checking are required to integrate and take full advantage of large language models in learning settings and teaching curricula. Other challenges such as the potential bias in the output, the need for continuous human oversight, and the potential for misuse are not unique to the application of AI in education. But we believe that, if handled sensibly, these challenges can offer insights and opportunities in education scenarios to acquaint students early on with potential societal biases, criticalities, and risks of AI applications. We conclude with recommendations for how to address these challenges and ensure that such models are used in a responsible and ethical manner in education.

Keywords

Large language models — Artificial Intelligence — Education — Educational Technologies

¹Technical University of Munich, Germany

²Ludwig-Maximilians-Universität München, Germany

³University of Tübingen, Germany

*Corresponding author: Enkelejda.Kasneci@tum.de

1. Introduction

Large language models, such as GPT-3 [1], have made significant advancements in natural language processing (NLP) in recent years. These models are trained on massive amounts of text data and are able to generate human-like text, answer questions, and complete other language-related tasks with high accuracy.

One key development in the area is the use of transformer architectures [2, 3] and the underlying attention mechanism [4], which have greatly improved the ability of auto-regressive¹ self-supervised² language models to handle long-

range dependencies in natural-language texts. The transformer architecture, introduced in [4], uses the self-attention mechanism to determine the relevance of different parts of the input when generating predictions. This allows the model to better understand the relationships between words in a sentence, regardless of their position.

Another important development is the use of pre-training, where a model is first trained on a large dataset before being fine-tuned on a specific task. This has proven to be an effective technique for improving performance on a wide range of language tasks [5]. For example, BERT [2] is a pre-trained transformer-based encoder model that can be fine-tuned on various NLP tasks, such as sentence classification, question answering and named entity recognition. In fact, the so-called few-shot learning capability of large language models to be

¹Auto-regressive because the model uses its previous predictions as input for new predictions.

²Self-supervised because they learn from the data itself, rather than being explicitly provided with correct answers as in supervised learning.

efficiently adapted to down-stream tasks or even other seemingly unrelated tasks (e.g., as in transfer learning) has been empirically observed and studied for various natural-language tasks [6], e.g., more recently in the context of generating synthetic and yet realistic heterogeneous tabular data [7].

Recent advancements also include GPT-3 [1] and ChatGPT [8], which were trained on a much larger datasets, i.e., texts from a very large web corpus, and have demonstrated state-of-the-art performance on a wide range of natural-language tasks ranging from translation to question answering, writing coherent essays, and computer programs. Additionally, extensive research has been conducted on fine-tuning these models on smaller datasets and applying transfer learning to new problems. This allows for improved performance on specific tasks with smaller amount of data.

While large language models have made great strides in recent years, there are still many limitations that need to be addressed. One major limitation is the lack of interpretability, as it is difficult to understand the reasoning behind the model's predictions. There are ethical considerations, such as concerns about bias and the impact of these models, e.g., on employment, risks of misuse and inadequate or unethical deployment, loss of integrity, and many more. Overall, large language models will continue to push the boundaries of what is possible in natural language processing. However, there is still much work to be done in terms of addressing their limitations and the related ethical considerations.

1.1 Opportunities for Learning

The use of large language models in education has been identified as a potential area of interest due to the diverse range of applications they offer. Through the utilization of these models, opportunities for enhancement of learning and teaching experiences may be possible for individuals at all levels of education, including primary, secondary, tertiary and professional development.

For elementary school students, large language models can assist in the development of reading and writing skills (e.g., by suggesting syntactic and grammatical corrections), as well as in the development of writing style and critical thinking skills. These models can be used to generate questions and prompts that encourage students to think critically about what they are reading and writing, and to analyze and interpret the information presented to them. Additionally, large language models can also assist in the development of reading comprehension skills by providing students with summaries and explanations of complex texts, which can make reading and understanding the material easier.

For middle and high school students, large language models can assist in the learning of a language and of writing styles for various subjects and topics, e.g., mathematics, physics, language and literature, and other subjects. These models can be used to generate practice problems and quizzes, which can help students to better understand, contextualize and retain the material they are learning. Additionally,

large language models can also assist in the development of problem-solving skills by providing students with explanations, step-by-step solutions, and interesting related questions to problems, which can help them to understand the reasoning behind the solutions and develop analytical and out-of-the-box thinking.

For university students, large language models can assist in the research and writing tasks, as well as in the development of critical thinking and problem-solving skills. These models can be used to generate summaries and outlines of texts, which can help students to quickly understand the main points of a text and to organize their thoughts for writing. Additionally, large language models can also assist in the development of research skills by providing students with information and resources on a particular topic and hinting at unexplored aspects and current research topics, which can help them to better understand and analyze the material.

For group & remote learning, large language models can be used to facilitate group discussions and debates by providing a discussion structure, real-time feedback and personalized guidance to students during the discussion. This can help to improve student engagement and participation. In collaborative writing activities, where multiple students work together to write a document or a project, language models can assist by providing style and editing suggestions as well as other integrative co-writing features. For research purposes, such models can be used to span the range of open research questions in relation to already researched topics and to automatically assign the questions and topics to the involved team members. For remote tutoring purposes, they can be used to automatically generate questions and provide practice problems, explanations, and assessments that are tailored to the students' level of knowledge so that they can learn at their own pace.

To empower learners with disabilities, large language models can be used in combination with speech-to-text or text-to-speech solutions to help people with visual impairment. In combination with the previously mentioned group and remote tutoring opportunities, language models can be used to develop inclusive learning strategies with adequate support in tasks such as adaptive writing, translating, and highlighting of important content in various formats. However, it is important to note that the use of large language models should be accompanied by the help of professionals such as speech therapists, educators, and other specialists that can adapt the technology to the specific needs of the learner's disabilities.

For professional training, large language models can assist in the development of language skills that are specific to a particular field of work. They can also assist in the development of skills such as programming, report writing, project management, decision making and problem-solving. For example, large language models can be fine-tuned on a domain-specific corpus (e.g. legal, medical, IT) in order to generate domain-specific language and assist learners in writing technical reports, legal documents, medical records etc.

They can also generate questions and prompts that encourage learners to think critically about their work and to analyze and interpret the information presented to them.

In conclusion, large language models have the potential to provide a wide range of benefits and opportunities for students and professionals at all stages of education. They can assist in the development of reading, writing, math, science, and language skills, as well as providing students with personalized practice materials, summaries and explanations, which can help to improve student performance and contribute to enhanced learning experiences. Additionally, large language models can also assist in research, writing, and problem-solving tasks, and provide domain-specific language skills and other skills for professional training. However, as previously mentioned, the use of these models should be done with caution, as they also have limitations such as lack of interpretability and potential for bias, unexpected brittleness in relatively simple tasks [9] which need to be addressed.

1.2 Opportunities for Teaching

Large language models, such as ChatGPT, have the potential to revolutionize teaching and assist in teaching processes. Below we provide only a few examples of how these models can benefit teachers:

For personalized learning, teachers can use large language models to create personalized learning experiences for their students. These models can analyze student's writing and responses, and provide tailored feedback and suggest materials that align with the student's specific learning needs. Such support can save teachers' time and effort in creating personalized materials and feedback, and also allow them to focus on other aspects of teaching, such as creating engaging and interactive lessons.

For lesson planning, large language models can also assist teachers in the creation of (inclusive) lesson plans and activities. Teachers can input to the models the corpus of document based on which they want to build a course. The output can be a course syllabus with short description of each topic. Language models can also generate questions and prompts that encourage the participation of people at different knowledge and ability levels, and elicit critical thinking and problem-solving. Moreover, they can be used to generate targeted and personalized practice problems and quizzes, which can help to ensure that students are mastering the material.

For language learning, teachers of language classes can use large language models in an assistive way, e.g., to highlight important phrases, generate summaries and translations, provide explanations of grammar and vocabulary, suggest grammatical or style improvements and assist in conversation practice. Language models can also provide teachers with adaptive and personalized means to assist students in their language learning journey, which can make language learning more engaging and effective for students.

For research and writing, large language models can assist teachers of university and high school classes to com-

plete research and writing tasks (e.g., in seminar works, paper writing, and feedback to students) more efficiently and effectively. The most basic help can happen at a syntactic level, i.e., identifying and correcting typos. At a semantic level, large language models can be used to highlight (potential) grammatical inconsistencies and suggest adequate and personalized improvement strategies. Going further, these models can be used to identify possibilities for topic-specific style improvement. They can also be used to generate summaries and outlines of challenging texts, which can help teachers and researchers to highlight the main points of a text in a way that is helpful for further deep dive and understanding of the content in question.

For professional development, large language models can also assist teachers by providing them with resources, summaries, and explanations of new teaching methodologies, technologies, and materials. This can help teachers stay up-to-date with the latest developments and techniques in education, and contribute to the effectiveness of their teaching. They can be used to improve the clarity of the teaching materials, locate information or resources that professionals may be in need for as they learn on the job, as well as used for on-the-job training modules that require presentation and communication skills.

For assessment and evaluation, teachers can use large language models to semi-automate the grading of student work by highlighting potential strengths and weakness of the work in question, e.g., essays, research papers, and other writing assignments. This can save teachers a significant amount of time for tasks related to individualized feedback to students. Furthermore, large language models can also be used to check for plagiarism, which can help to prevent cheating. Hence, large language models can help teachers to identify areas where students are struggling, which adds to a more accurate assessments of student learning development and challenges. Targeted instruction provided by the models can be used to help students excel and to provide opportunities for further development.

The acquaintance of students with AI challenges related to the potential bias in the output, the need for continuous human oversight, and the potential for misuse of large language models are not unique to education. In fact, these challenges are inherent to transformative digital technologies. Thus, we believe that, if handled sensibly by the teacher, these challenges can be insightful in learning and education scenarios to acquaint students early on with potential societal biases, and risks of AI application.

In conclusion, large language models have the potential to revolutionize teaching from a teacher's perspective by providing teachers with a wide range of tools and resources that can assist with lesson planning, personalized content creation, differentiation and personalized instruction, assessment, and professional development. Overall, large language models have the potential to be a powerful tool in education, and there are a number of ongoing research efforts exploring its potential applications in this area.

2. Current Research and Applications of Language Models in Education

2.1 Overview of Current Large Language Models

The GPT (Generative Pre-trained Transformer) [10] model developed by OpenAI was the first large language model that was publicly released in 2018. GPT was able to generate human-like text, answer questions, and assist in tasks, such as translation and summarization, through human-like completion. Based on this initial model, OpenAI later released the GPT-2 and GPT-3 models with more advanced capabilities. It can be argued that the release of GPT marked a significant milestone in the field of NLP and has opened up many ways for dissemination, both in research and industrial applications.

Another model that was released by Google Research in 2018 is BERT (Bidirectional Encoder Representations from Transformers [2]), which is also based on a transformer architecture and is pre-trained on a massive dataset of text data on two unsupervised tasks, namely masked language modeling (to predict missing parts in a sentence and learn their context) and next sentence prediction (to learn plausible subsequent sentences of a given sentence) with the aim to learn the broader context of words in various topics.

One year later in 2019, Google AI released XLNet [11], which is trained using a process called permutation language modeling and enables XLNet to cope with tasks that involve understanding the dependencies between words in a sentence, e.g., natural language inference and question answering. Another model developed by Google Research was T5 (Text-to-Text Transfer Transformer) [12], which was released in 2020. Like the predecessor models, T5 is also transformer-based model trained on a massive text dataset with the key feature being its ability to perform many NLP tasks with a single pre-training and fine-tuning pipeline.

In parallel with Open AI and Google, Facebook AI developed a large language model called RoBERTa (Robustly Optimized BERT Pre-training) [13], which was released in 2019. RoBERTa is a variant of the BERT model which uses dynamic masking instead of static masking during pre-training. Additionally, RoBERTa is trained on a much larger dataset, and hence clearly outperformed BERT and other models such as GPT-2 and XLNet by the time of its release.

Currently, the most widely used and the largest available language model is GPT-3, which was also pre-trained on a massive text dataset (including books, articles, and websites, among other sources) and has 175 billion parameters. As all other previously described language models, GPT-3 uses a transformer architecture, which allows it to efficiently process sequential data and generate more coherent and contextually adjusted text. Indeed, text generated by GPT-3 is almost indistinguishable from human-written text [14]. With the ability to perform zero-shot learning, GPT-3 can cope with tasks it has not been specifically trained on, providing hence enormous opportunities for applications, from automation (summarizing, completing texts from bullet points), to dialog

systems, chatbots, and creative writing.

Just recently, the BigScience-community developed and released the large language model BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) [15] as an open-source joint project by HuggingFace, GENCI and IDRIS³. The aim of this project was to provide a transparently trained multi-lingual language model for the academic and non-profit community. BLOOM is based on the same transformer architecture as the models from the GPT-family with only minor structural changes, but the training data was explicitly chosen to cover 46 natural languages and 13 programming languages, resulting in a data volume of 1.6TB.

2.2 Review of Research Applying Large Language Models in Education

In the following, we provide an overview of research works employing large language models in education that were published since the release of the first large language model in 2018. These studies have been discussed in the following according to their target groups, i.e., learners or teachers.

Learners' perspective. From a student's perspective, large language models can be used in multiple ways to assist the learning process. One example is in the creation and design of educational content. For example, researchers have used large language models to generate interactive educational materials such as quizzes and flashcards, which can be used to improve student learning and engagement [16, 17]. More specifically, in a recent work by Dijkstra et al. [17], researchers have used GPT-3 to generate multiple-choice questions and answers for a reading comprehension task and argue that automated generation of quizzes not only reduces the burden of manual quiz design for teachers but, above all, provides a helpful tool for students to train and test their knowledge while learning from textbooks and during exam preparation [17].

In another recent work, GPT-3 was employed as a pedagogical agent to stimulate the curiosity of children and enhance question-asking skills [18]. More specifically, the authors automated the generation of curiosity-prompting cues as an incentive for asking more and deeper questions. According to their results, large language models not only bear the potential to significantly facilitate the implementation of curiosity-stimulating learning but can also serve as an efficient tool towards an increased curiosity expression [18].

In computing education, a recent work by MacNeil et al. [19] has employed GPT-3 to generate code explanations. Despite several open research and pedagogical questions that need to be further explored, this work has successfully demonstrated the potential of GPT-3 to support learning by explaining aspects of a given code snippet.

For a data science course, Bhat et al. [20] proposed a pipeline for generating assessment questions based on a fine-tuned GPT3 model on text-based learning materials. The generated questions were further evaluated with regard to their usefulness to the learning outcome based on automated label-

³<https://huggingface.co/bigscience/bloom>

ing by a trained GPT-3 model and manual reviews by human experts. The authors reported that the generated questions were rated favorably by human experts, promoting thus the usage of large language models in data science education [20].

Students can learn from each other by peer-reviewing and assessing each other's solutions. This, of course, has the best effect when the given feedback is comprehensive and of high quality. For example, Jia et al. [21] showed how BERT can be used to evaluate the peer assessments so that students can learn to improve their feedback.

In a recent review on conversational AI in language education, the authors found that there are five main applications of conversational AI during teaching [22], the most common one being the use of large language models as a conversational partner in a written or oral form, e.g., in the context of a task-oriented dialogue that provides language practice opportunities such as pronunciation [23]. Another application is to support students when they experience foreign language learning anxiety [24] or have a lower willingness to communicate [25]. In [26], the application of providing feedback, as a needs analyst, and evaluator when primary school students practice their vocabulary was explored. The authors of [27] found that a chatbot that is guided by a mind map is more successful in supporting students by providing scaffolds during language learning than a conventional AI chatbot.

A recent work in the area of medical education by Kung et al. [28] explored the performance of ChatGPT on the United States Medical Licensing Exam. According to the evaluation results, the performance of ChatGPT on this test was at or near the passing threshold without any domain fine-tuning. Based on these results, the authors argue that large language models might be a powerful tool to assist medical education and eventually clinical decision-making processes [28].

Teachers' perspective. As the rate of adoption of AI in education is still slow compared to other fields, such as industrial applications (e.g., finance, e-commerce, automotive) or medicine, there are less studies considering the use of large language models in education [29]. A recent review of opportunities and challenges of chatbots in education pointed out that the studies related to chatbots in education are still in an early stage, with few empirical studies investigating the use of effective learning designs or learning strategies [30]. Therefore, we discuss first the teachers' perspectives concerning AI and Learning Analytics in education and transfer these on the much newer field of large language models.

In this view, a pilot study with European teachers indicates a positive attitude towards AI for education and a high motivation to introduce AI-related content at school. Overall, the teachers from the study seemed to have a basic level of digital skills but low AI-related skills [31]. Another study with Nigerian teachers emphasized that the willingness and readiness of teachers to promote AI are key prerequisites for the integration of AI-based technologies in education [32]. Along the same lines, the results of a study with teachers from South Korea indicate that teachers with constructivist beliefs

are more likely to integrate educational AI-based tools than teachers with transmissive orientations [33]. Furthermore, perceived usefulness, perceived ease of use, and perceived trust in these AI-based tools are determinants to be considered when predicting their acceptance by the teachers. Similar results concerning teachers attitudes towards chatbots in education were reported in [34]: perceiving the AI chatbot as easy-to-use and useful leads to greater acceptance of the chatbot. As for the chatbots' features, formal language by a chatbot leads to a higher intention of using it.

As it seems that teachers' perspectives on the general use of AI in education have a lot in common with the mentioned attitude towards chatbots in particular, a responsible integration of AI into education by involving the expertise of different communities is crucial [35].

Recent works addressing the use of large language models from the teacher's perspective have focused on the automated assessment of student answers, adaptive feedback, and the generation of teaching content.

For example, a recent work by Moore et al. [36] employed a fine-tuned GPT-3 model to evaluate student-generated answers in a learning environment for chemistry education [36]. The authors argue that large language models might (especially when fine-tuned to the specific domain) be a powerful tool to assist teachers in the quality and pedagogical evaluation of student answers [36]. In addition, the following studies examined NLP-based models for generating automatic adaptive feedback: Zhu et al. [37] examined an AI-based feedback system incorporating automated scoring technologies in the context of a high school climate activity task. The results show that the feedback helped students revise their scientific arguments. Sailer et al. [38] used NLP-based adaptive feedback in the context of diagnosing students' learning difficulties in teacher education. In their experimental study, they found that pre-service teachers who received adaptive feedback were better able to justify their diagnoses than prospective teachers who received static feedback. Bernius et al. [39] used NLP-based models to generate feedback for textual student answers in large courses, where grading effort could be reduced by up to 85% with a high precision and an improved quality perceived by the students.

Large language models can not only support the assessment of student's solutions but also assist in the automatic generation of exercises. Using few-shot learning, [40] showed that the OpenAI Codex model is able to provide a variety of programming tasks together with the correct solution, automated tests to verify the student's solutions, and additional code explanations. With regard to testing factual knowledge in general, [41] proposed a framework to automatically generate question-answer pairs. This can be used in the creation of teaching materials, e.g., for reading comprehension tasks. Beyond the generation of the correct answer, transformer models are also able to create distractor answers, as needed for the generation of multiple choice questionnaires [42, 43]. Bringing language models to mathematics education, sev-

eral works discuss the automatic generation of math word problems [44, 45, 46], which combines the challenge of understanding equations and putting them into the appropriate context.

Finally, another recent work [47] investigated the capability of state-of-the-art conversational agents to adequately reply to a student in an educational dialogue. Both models used in this work (Blender and GPT-3) were capable of replying to a student adequately and generated conversational dialogues that conveyed the impression that these models understand the learner (in particular Blender). They are however well behind human performance when it comes to helping the student [47], thus emphasizing the need for further research.

3. Opportunities for Innovative Educational Technologies

Looking forward, large language models bear the potential to considerably improve digital ecosystems for education, such as environments based on Augmented Reality (AR), Virtual Reality (VR) [48, 49], and other related digital experiences. Specifically, they can be used to amplify several key factors, which are crucial for the immersive interaction of users with digital content. For example, large language models can considerably improve the natural language processing and understanding capabilities of an AR/VR system to enable an effective natural communication and interaction between users and the system (e.g., virtual teacher or virtual peers). The latter has been identified early on as a key usability aspect for immersive educational technologies [50] and is in general seen as a key factor for improving the interaction between humans and AI systems [51].

Large language models can also be used to develop more natural and sophisticated user interfaces by exploiting their ability to generate contextualized, personalized, and diverse responses to natural language questions asked by users. Furthermore, their ability to answer natural language questions across various domains can facilitate the integration of diverse digital applications into a unified framework or application, which is also critical for expanding the bounds of educational possibilities and experiences [52, 49].

In general, the ability of these models to generate contextualized natural language texts, code for various implementation tasks [53] as well as various types of multimedia content (e.g., in combination with other AI systems, such as DALL-E [54]) can enable and scale the creation of compelling and immersive digital (e.g., AR/VR) experiences. From gamification to detailed simulations for immersive learning in digital environments, large language models are a key enabling technology. To fully realize this potential, however, it is important to consider not only technical aspects but also ethical, legal, ecological and social implications.

In the following section, we take a brief look at the risks related to the application on large language models in education and provide corresponding mitigation strategies.

4. Key Challenges and Risks Related to the Application of Large Language Models in Education

Copyright Issues. When we train large language models on a task to produce education-related content – course syllabus, quizzes, scientific paper – the mode should be trained on examples of such texts. During the generation for a new prompt, the answer may contain a full sentence or even a paragraph seen in the training set, leading to copyright and plagiarism issues.

Important steps to responsibly mitigate such an issue can be the following:

- Asking the authors of the original documents transparently (i.e., purpose and policy of data usage) for permission to use their content for training the model
- Compliance with copyright terms for open-source content
- Inheritance and detailed terms of use for the content generated by the model
- Informing and raising awareness of the users about these policies

Bias and fairness. Large language models can perpetuate and amplify existing biases and unfairness in society, which can negatively impact teaching and learning processes and outcomes. For example, if a model is trained on data that is biased towards certain groups of people, it may produce results that are unfair or discriminatory towards those groups (e.g., local knowledge about minorities such as small ethnic groups or cultures can fade into the background). Thus, it is important to ensure that the training data or the data used for fine-tuning on down-stream tasks for the model is diverse and representative of different groups of people. Regular monitoring and testing of the model's performance on different groups of people can help identify and address any biases early on. Hence, human oversight in the process is indispensable and critical for the mitigation of bias and beneficial application of large language models in education.

More specifically, a responsible mitigation strategy would focus on the following key aspects:

- A diverse set of data to train or fine-tune the model, to ensure that it is not biased towards any particular group
- Regular monitoring and evaluation of the model's performance (on diverse groups of people) to identify and address any biases that may arise
- Fairness measures and bias-correction techniques, such as pre-processing or post-processing methods
- Transparency mechanisms that enable users to comprehend the model's output, and the data and assumptions that were used to generate it

- Professional training and resources to educators on how to recognize and address potential biases and other failures in the model's output
- Continuous updates of the model with diverse, unbiased data, and supervision of human experts to review the results

Learners may rely too heavily on the model. The effortlessly generated information could negatively impact their critical thinking and problem-solving skills. This is because the model simplifies the acquisition of answers or information, which can amplify laziness and counteract the learners' interest to conduct their own investigations and come to their own conclusions or solutions.

To encounter this risk, it is important to be aware of the limitations of large language models and use them only as a tool to support and enhance learning [55], rather than as a replacement for human authorities and other authoritative sources. Thus a responsible mitigation strategy would focus on the following key aspects:

- Raising awareness of the limitations and unexpected brittleness of large language models and AI systems in general (i.e., experimenting with the model to build an own understanding of the workings and limitations)
- Using language models to generate hypotheses and explore different perspectives, rather than just to generate answers
- Strategies to use other educational resources (e.g., books, articles) and other authoritative sources to evaluate and corroborate the factual correctness of the information provided by the model (i.e., encouraging learners to question the generated content)
- Incorporating critical thinking and problem-solving activities into the curriculum, to help students develop these skills
- Incorporating human expertise and teachers to review, validate and explain the information provided by the model

It is important to note that the use of large language models should be integrated into the curriculum in a way that complements and enhances the learning experience, rather than replacing it.

Teachers may become too reliant on the models. Using large language models can provide accurate and relevant information, but they cannot replace the creativity, critical thinking, and problem-solving skills that are developed through human instruction. It is therefore important for teachers to use these models as a supplement to their instruction, rather than a replacement. Thus, crucial aspects to mitigate the risk of becoming too reliant on large language models are:

- The use of language models only as a complementary supplement to the generation of instructions
- Ongoing training and professional development for teachers, enabling them to stay up-to-date on the best-practice use of language models in the classroom to elicit and promote creativity and critical thinking
- Critical thinking and problem-solving activities through the assistance of digital technologies as an integral part of the curriculum to ensure that students are developing these skills
- Engagement of students in creative and independent projects that allow them to develop their own ideas and solutions
- Monitoring and evaluating the use of language models in the classroom to ensure that they are being used effectively and not negatively impacting student learning
- Incentives for teachers and schools to develop (inclusive, collaborative, and personalized) teaching strategies based on large language models and engage students in problem-solving processes such as retrieving and evaluating course/assignment-relevant information using the models and other sources

Lack of understanding and expertise. Many educators and educational institutions may not have the knowledge or expertise to effectively integrate new technologies in their teaching [56]. This particularly applies to the use and integration of large language models into teaching practice. Educational theory has long since suggested ways of integrating novel tools into educational practice (e.g., [57]). As with any other technological innovation, integrating large language models into effective teaching practice requires understanding their capabilities and limitations, as well as how to effectively use them to supplement or enhance specific learning processes.

There are several ways to address these challenges and encounter this risk:

- Research on the challenges of large language models in education by investigating existing educational models of technology integration, students' learning processes and transfer them to the context of large language models, as well as developing a new educational theory specifically for the context of large language models
- Assessing the needs of the educators and students and provide case-based guidance (e.g., for the secure ethical use of large language models in education scenarios)
- Demand-oriented Training and professional development opportunities for educators and institutions to learn about the capabilities and potential uses of large language models in education, as well as providing best practices for integrating them into their teaching methods

- Open educational resources (e.g., tutorials, studies, use cases, etc.) and Guidelines for educators and institutions to access and learn about the use of language models in education
- Incentives for collaboration and community building (e.g., professional learning communities) among educators and institutions that are already using language models in their teaching practice, so they can share their knowledge and experience with others
- Regular analysis and feedback on the use of language models to ensure their effective use and make adjustments as necessary

Difficulty to distinguish model-generated from student-generated answers. It is becoming increasingly difficult to distinguish whether a text is machine- or human-generated, presenting an additional major challenge to teachers and educators [58, 59, 60, 61]. As a result, the New York City’s Department of Education recently banned ChatGPT from schools’ devices and networks [62].

Just recently, Cotton et al. [60] proposed several strategies to detect work that has been generated by large language models, and specifically ChatGPT. In addition, tools, such as the recently released GPTZero [63], which uses perplexity, as a measure that hints at generalization capabilities (of the agent by which the text was written), to detect AI involvement in text writing, are expected to provide additional support. More advanced techniques aim at watermarking the content generated by language models [64, 65], e.g., by biasing the content generation towards terms, which are rather unlikely to be jointly used by humans in a text passage. In the long run, however, we believe that developing curricula and instructions that encourage the creative and evidence-based use of large language models will be the key to solving this problem. Hence, a reasonable mitigation strategy for this risk should focus on:

- Research on transparency, explanation and analysis techniques and measures to distinguish machine- from human-generated text
- Incentives and support to develop curricula and instructions that require the creative and complementary use of large language models

Cost of training and maintenance. The maintenance of large language models could be a financial burden for schools and educational institutions, especially those with limited budgets. To address this challenge, the use of pre-trained models and cloud technology in combination with cooperative schemes for usage in partnership with institutions and companies can serve as a starting point. Specifically, a mitigation strategy for this risk should focus on the following aspects:

- Use of pre-trained open-source models, which can be fine-tuned for specific tasks

- Development and exploration of partnerships with private companies, research institutions as well as governmental and non-profit organizations that can provide financial support, resources and expertise to support the use of large language models in education
- Shared costs and cooperative use of scalable (e.g., cloud) computing services that provide access to powerful computational resources at a low cost
- Use of the model primarily for high-value educational tasks, such as providing personalized and targeted learning experiences for students (i.e., assignment of lower priority to low-value tasks)
- Research and development of compression, distillation, and pruning techniques to reduce the size of the model, the data, and the computational resources required

Data privacy and security. The use of large language models in education raises concerns about data privacy and security, as student data is often sensitive and personal. This can include concerns about data breaches, unauthorized access to student data, and the use of student data for purposes other than education.

Some specific focus areas to mitigate privacy and security concerns when using large language models in education are:

- Development and implementation of robust data privacy and security policies that clearly outline the collection, storage, and use of student data in compliance with regulation (e.g., GDPR, HIPAA, FERPA) and ethical standards
- Transparency towards students and their families about the data collection, storage, and use practices, with obligatory consent before data collection and use
- Modern technologies and measures to protect the collected data from unauthorized access, breaches, or unethical use (e.g., anonymized data and secure infrastructures with modern means for encryption, federation, privacy-preserving analytics, etc.)
- Regular audits of the data privacy and security measures in place to identify and address any potential vulnerabilities or areas for improvement
- Incident response plan to quickly respond and mitigate any data breaches or unauthorized access to data
- Education and awareness of the staff, i.e., educators and students about the data privacy and security policies, regulations, ethical concerns and best practices to handle and report related risks

Sustainable usage. Large language models have high computational demands, which can result in high energy consumption. Hence, energy-efficient hardware and shared (e.g., cloud) infrastructure based on renewable energy are crucial for their environmentally sustainable operation and scaling needed in the context of education.

For model training and updates, only data that has been collected and annotated in a regulatory compliant and ethical way should be considered. Therefore, governance frameworks that include policies, procedures, and controls to ensure such appropriate use of such models are key to their successful adoption.

Likewise, for the long-term trustworthy and responsible use of the models, transparency, bias mitigation, and ongoing monitoring are indispensable.

In summary, the mitigation strategy for this risk would include:

- Energy-efficient hardware and shared infrastructure based on renewable energy as well as research on reducing the cost of training and maintenance (i.e., efficient algorithms, representation, and storage)
- Collection, annotation, storage, and processing of data in a regulatory compliant and ethical way
- Transparency and explanation techniques to identify and mitigate biases and prevent unfairness
- Governance frameworks that include policies, procedures, and controls to ensure the above points and the appropriate use in education

Cost to verify information and maintain integrity. It is important to verify the information provided by the model by consulting external authoritative sources to ensure accuracy and integrity. Additionally, there may be financial costs associated with maintaining and updating the model to ensure it is providing accurate up-to-date information.

A responsible mitigation strategy for this risk would consider the following key aspects:

- Regularly updates of the model with new and accurate information to ensure it is providing up-to-date and accurate information
- Use of multiple authoritative sources to verify the information provided by the model to ensure correctness and integrity
- Use of the model in conjunction with human expertise, e.g., teachers or subject matter experts, who review and validate the information provided by the model
- Development of protocol and standards for fact-checking and corroborating information provided by the model
- Provide clear and transparent information on the model's performance, what it is or is not capable of, and the conditions under which it operates.

- Training and resources for educators and learners on how to use the model, interpret its results and evaluate the information provided
- Regular review and evaluation of the model with transparent reporting on the model's performance, i.e., what it is or is not capable of and the identification of conditions under which inaccuracies or other issues may arise

Difficulty to distinguish between real knowledge and convincingly written but unverified model output.

The ability of large language models to generate human-like text can make it difficult for students to distinguish between real knowledge and unverified information. This can lead to students accepting false or misleading information as true, without questioning its validity.

To mitigate this risk, in addition to the above verification- and integrity-related mitigation strategy, it is important to provide education on how information can be evaluated critically and teach students exploration, investigation, verification, and corroboration strategies.

Lack of adaptability. Large language models are not able to adapt to the diverse needs of students and teachers, and may not be able to provide the level of personalization required for effective learning. This is a limitation of the current technology, but it is conceivable that with more advanced models, the adaptability will increase.

More specifically, a sensible mitigation strategy would be comprised of:

- Use of adaptive learning technologies to personalize the output of the model to the needs of individual students by using student data (e.g., about learning style, prior knowledge, and performance, etc.)
- Customization of the language model's output to align with the teaching style and curriculum (by using data provided by the teacher)
- Use of multi-modal learning and teaching approaches, which combine text, audio, video, and experimentation to provide a more engaging and personalized experience for students and teachers
- Use of hybrid approaches, which combine the strengths of both human teachers and language models to generate targeted and personalized learning materials (based on feedback, guidance, and support provided by the teachers)
- Regular review of the model and continual improvement for curriculum-related uses cases to ensure adequate and accurate functioning for education purposes
- Research and development to create more advanced models that can better adapt to the diverse needs of students and teachers

5. Further Issues Related to User Interfaces and Fair Access

Appropriate user interfaces. For the integration of large language models into educational workflows, further research on Human-Computer Interaction and User Interface Design is necessary.

In this work, we have discussed several potential use cases for learners of different age – from children to adults. While creating such AI-based assistants, we should take into account the degree of psychological maturity, fine motor skills, and technical abilities of the potential users. Thus, the user interface should be appropriate for the task, but may also have varying degrees of human imitation – for instance, for children it might be better to hide machinery artifacts in generated text and use gamified interaction and learning approaches as much as possible so as to enable a smooth and engaging interaction with such technologies, whereas for older learners the machine-based content could be exploited to promote problem-solving, critical thinking and fact-checking abilities.

In general, the design of user interfaces for AI-based assistance and learning tools should promote the development of 21st century learning and problem-solving skills [66], especially, critical thinking, creativity, communication, and collaboration, for which further evidence-based research is needed. In this context, a crucial aspect is the appropriate age- and background-related integration of AI-based assistance to maximize its benefits and minimize any potential drawbacks.

Multilingualism and fair access. While the majority of the research in large language models is done for the English language, there is still a gap of research in this field for other languages. This can potentially make education for English-speaking users easier and more efficient than for other users, causing unfair access to such education technologies for non-English speaking users. Despite the efforts of various research communities to address multilingualism fairness for AI technologies, there is still much room for improvement.

Lastly, the unfairness related to financial means for accessing, training and maintaining large language models may need to be regulated by governmental organisations with the aim to provide equity-oriented means to all educational entities interested in using these modern technologies. Without fair access, this AI technology may seriously widen the education gap like no other technology before it.

We therefore conclude with UNESCO’s call to ensure that AI does not widen the technological and educational divides within and between countries, and recommended important strategies for the use of AI in a responsible and fair way to reduce this existing gap instead. According to the UNESCO education 2030 Agenda [67]: “*UNESCO’s mandate calls inherently for a human-centred approach to AI. It aims to shift the conversation to include AI’s role in addressing current inequalities regarding access to knowledge, research and the diversity of cultural expressions and to ensure AI does not widen the technological divides within and between countries.*

The promise of “AI for all” must be that everyone can take advantage of the technological revolution under way and access its fruits, notably in terms of innovation and knowledge.”

6. Concluding Remarks

The use of large language models in education is a promising area of research that offers many opportunities to enhance the learning experience for students and support the work of teachers. However, to unleash their full potential for education, it is crucial to approach the use of these models with caution and to critically evaluate their limitations and potential biases. Integrating large language models into education must therefore meet stringent privacy, security, and - for sustainable scaling - environmental, regulatory and ethical requirements, and must be done in conjunction with ongoing human monitoring, guidance, and critical thinking.

While this position paper reflects the optimism of the authors about the opportunities of large language models as a transformative technology in education, it also underscores the need for further research to explore best practices for integrating large language models into education and to mitigate the risks identified.

We believe that despite many difficulties and challenges, the discussed risks are manageable and should be addressed to provide trustworthy and fair access to large language models for education. Towards this goal, the mitigation strategies proposed in this position paper could serve as a starting point.

References

- [1] Luciano Floridi and Massimo Chiriatti. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [3] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [5] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*, 2021.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances*

- in *neural information processing systems*, 33:1877–1901, 2020.
- [7] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.
 - [8] OpenAI Team. ChatGPT: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt/>, November 2022. Accessed: 2023-01-19.
 - [9] Tasmia Ansari (Analytics India Magazine). Freaky ChatGPT fails that caught our eyes! <https://analyticsindiamag.com/freaky-chatgpt-fails-that-caught-our-eyes/>, Dec 2022. Accessed: 2023-01-22.
 - [10] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
 - [11] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russlan Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in neural information processing systems*, 32, 2019.
 - [12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
 - [13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
 - [14] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that’s ‘human’ is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online, August 2021. Association for Computational Linguistics.
 - [15] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*, 2022.
 - [16] Ebrahim Gabajiwala, Priyav Mehta, Ritik Singh, and Reeta Koshy. Quiz Maker: Automatic Quiz Generation from Text Using NLP. In *Futuristic Trends in Networks and Computing Technologies*, pages 523–533, Singapore, 2022. Springer.
 - [17] Ramon Dijkstra, Zülküf Genç, Subhradeep Kayal, and Jaap Kamps. Reading Comprehension Quiz Generation using Generative Pre-trained Transformers. https://e.humanities.uva.nl/publications/2022/dijk_read22.pdf, 2022.
 - [18] Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Hélène Sauzéon, and Pierre-Yves Oudeyer. GPT-3-driven pedagogical agents for training children’s curious question-asking skills. *arXiv preprint arXiv:2211.14228*, 2022.
 - [19] Stephen MacNeil, Andrew Tran, Dan Mogil, Seth Bernstein, Erin Ross, and Ziheng Huang. Generating Diverse Code Explanations Using the GPT-3 Large Language Model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 2*, ICER ’22, page 37–39, New York, NY, USA, 2022. Association for Computing Machinery.
 - [20] Shravya Bhat, Huy A. Nguyen, Steven Moore, John Stamper, Majd Sakr, and Eric Nyberg. Towards Automated Generation and Evaluation of Questions in Educational Domains. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 701–704, Durham, United Kingdom, 2022. International Educational Data Mining Society.
 - [21] Qinjin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, and Edward F. Gehring. ALL-IN-ONE: Multi-Task Learning BERT models for Evaluating Peer Assessments. *International Educational Data Mining Society*, 2021.
 - [22] Hyangeun Ji, Insook Han, and Yujung Ko. A systematic review of conversational ai in language education: focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, pages 1–16, 2022.
 - [23] Reham El Shazly. Effects of artificial intelligence on english speaking anxiety and speaking performance: A case study. *Expert Systems*, 38(3):e12667, 2021.
 - [24] Minhui Bao. Can home use of speech-enabled artificial intelligence mitigate foreign language anxiety—investigation of a concept. *Arab World English Journal (AWEJ) Special Issue on CALL*, (5), 2019.
 - [25] Tzu-Yu Tai and Howard Hao-Jan Chen. The impact of google assistant on adolescent efl learners’ willingness to communicate. *Interactive Learning Environments*, pages 1–18, 2020.
 - [26] Jaeho Jeon. Chatbot-assisted dynamic assessment (cada) for l2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, pages 1–27, 2021.
 - [27] Chi-Jen Lin and Husni Mubarak. Learning analytics for investigating the mind map-guided ai chatbot approach in an efl flipped speaking classroom. *Educational Technology & Society*, 24(4):16–35, 2021.

- [28] Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepano, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *medRxiv*, 2022.
- [29] Sdenka Zobeida Salas-Pilco, Kejiang Xiao, and Xinyun Hu. Artificial intelligence and learning analytics in teacher education: A systematic review. *Education Sciences*, 12(8), 2022.
- [30] Gwo-Jen Hwang and Ching-Yi Chang. A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, pages 1–14, 2021.
- [31] Sara Polak, Gianluca Schiavo, and Massimo Zancanaro. Teachers’ Perspective on Artificial Intelligence Education: An Initial Investigation. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA ’22, New York, NY, USA, 2022. Association for Computing Machinery.
- [32] Musa Adekunle Ayanwale, Ismaila Temitayo Sanusi, Owolabi Paul Adelana, Kehinde D. Aruleba, and Solomon Sunday Oyeler. Teachers’ readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence*, 3:100099, 2022.
- [33] Seongyune Choi, Yeonju Jang, and Hyeoncheol Kim. Influence of Pedagogical Beliefs and Perceived Trust on Teachers’ Acceptance of Educational Artificial Intelligence Tools. *International Journal of Human–Computer Interaction*, 39(4):910–922, 2023.
- [34] Raquel Chocarro, Mónica Cortiñas, and Gustavo Marcos-Matás. Teachers’ attitudes towards chatbots in education: a technology acceptance model approach considering the effect of social language, bot proactiveness, and users’ characteristics. *Educational Studies*, pages 1–19, 2021.
- [35] Charles Fadel, Wayne Holmes, and Maya Bialik. Artificial intelligence in education: Promises and implications for teaching and learning. *The Center for Curriculum Redesign*, 2019.
- [36] Steven Moore, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3. In *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, ECTEL 2022, Toulouse, France, September 12–16, 2022, Proceedings*, pages 243–257. Springer, 2022.
- [37] Hee-Sun Lee Mengxiao Zhu, Ou Lydia Liu. The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, 143:103668, 2020.
- [38] Michael Sailer, Elisabeth Bauer, Riikka Hofmann, Jan Kiesewetter, Julia Glas, Iryna Gurevych, and Frank Fischer. Adaptive feedback from artificial neural networks facilitates pre-service teachers’ diagnostic reasoning in simulation-based learning. *Learning and Instruction*, 83:101620, 2023.
- [39] Jan Philip Bernius, Stephan Krusche, and Bernd Bruegge. Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence*, 3, 2022.
- [40] Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43, 2022.
- [41] Fanyi Qu, Xin Jia, and Yunfang Wu. Asking Questions Like Educational Experts: Automatically Generating Question-Answer Pairs on Real-World Examination Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2583–2593, 2021.
- [42] Ricardo Rodriguez-Torrealba, Eva Garcia-Lopez, and Antonio Garcia-Cabot. End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, 208:118258, 2022.
- [43] Vatsal Raina and Mark Gales. Multiple-Choice Question Generation: Towards an Automated Assessment Framework. *arXiv preprint arXiv:2209.11830*, 2022.
- [44] Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. Generate & Rank: A Multi-task Framework for Math Word Problems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2269–2279, 2021.
- [45] Weijiang Yu, Yingpeng Wen, Fudan Zheng, and Nong Xiao. Improving math word problems with pre-trained knowledge and hierarchical reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3384–3394, 2021.
- [46] Zichao Wang, Andrew Lan, and Richard Baraniuk. Math word problem generation with mathematical consistency and problem context constraints. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5986–5999, 2021.
- [47] Anaïs Tack and Chris Piech. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 522–529, Durham, United Kingdom, 2022. International Educational Data Mining Society.
- [48] Mario A Rojas-Sánchez, Pedro R Palos-Sánchez, and José A Folgado-Fernández. Systematic literature review and bibliometric analysis on virtual reality and education. *Education and Information Technologies*, pages 1–38, 2022.

- [49] Abhimanyu S Ahuja, Bryce W Polascik, Divyesh Dodapaneni, Eamonn S Byrnes, and Jayanth Sridhar. The digital metaverse: Applications in artificial intelligence, medical education, and integrative health. *Integrative Medicine Research*, 12(1):100917, 2023.
- [50] Maria Roussou. Immersive interactive virtual reality in the museum. *Proc. of TiLE (Trends in Leisure Entertainment)*, 2001.
- [51] Andrea L Guzman and Seth C Lewis. Artificial intelligence and communication: A human-machine communication research agenda. *New Media & Society*, 22(1):70–86, 2020.
- [52] Jeremy Kerr and Gillian Lawson. Augmented reality in design education: Landscape architecture studies as ar experience. *International Journal of Art & Design Education*, 39(1):6–21, 2020.
- [53] Brett A Becker, Paul Denny, James Finnie-Ansley, Andrew Luxton-Reilly, James Prather, and Eddie Antonio Santos. Programming is hard—or at least it used to be: Educational opportunities and challenges of ai code generation. *arXiv preprint arXiv:2212.01020*, 2022.
- [54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021.
- [55] John V. Pavlik. Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator*, page 10776958221149577, 2023.
- [56] Christine Redecker et al. European framework for the digital competence of educators: DigCompEdu. Technical report, Joint Research Centre (Seville site), 2017.
- [57] Gavriel Salomon. On the nature of pedagogic computer tools: The case of the writing partner. *Computers as cognitive tools*, 179:196, 1993.
- [58] Katherine Elkins and Jon Chun. Can GPT-3 pass a Writer’s turing test? *Journal of Cultural Analytics*, 5:17212, 2020.
- [59] Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*, 2022.
- [60] Debby R.E. Cotton, Peter A. Cotton, and J.Reuben Shipway. Chatting and Cheating. Ensuring academic integrity in the era of ChatGPT. *EdArXiv*, 2023.
- [61] Dehouche Nassim. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, 21:17–23, 2021.
- [62] Kalhan Rosenblatt (NBC News). ChatGPT banned from New York City public schools’ devices and networks. <https://nbcnews.to/3iTE0t6>, January 2023. Accessed: 22.01.2023.
- [63] Edward Tian. GPTZero. <https://gptzero.me/>, 2023. Accessed: 22.01.2023.
- [64] Chenxi Gu, Chengsong Huang, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. Watermarking Pre-trained Language Models with Backdooring. *arXiv preprint arXiv:2210.07543*, 2022.
- [65] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models. *arXiv preprint arXiv:2301.10226v1*, 2023.
- [66] Carol C Kuhlthau, Leslie K Maniotes, and Ann K Caspari. *Guided inquiry: Learning in the 21st century: Learning in the 21st century*. Abc-Clio, 2015.
- [67] UNESCO. Education 2030 Agenda. <https://www.unesco.org/en/digital-education/artificial-intelligence>, 2023. Accessed: 22.01.2023.