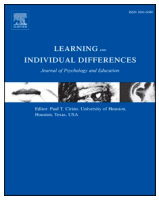
Learning and Individual Differences 103 (2023) 102274

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif



Commentary

ChatGPT for good? On opportunities and challenges of large language models for education

Enkelejda Kasneci [a,*], Kathrin Sessler [a], Stefan Küchemann [b], Maria Bannert [a], Daryna Dementieva [a], Frank Fischer [b], Urs Gasser [a], Georg Groh [a], Stephan Günnemann [a], Eyke Hüllermeier [b], Stephan Krusche [a], Gitta Kutyniok [b], Tilman Michaeli [a], Claudia Nerdel [a], Jürgen Pfeffer [a], Oleksandra Poquet [a], Michael Sailer [b], Albrecht Schmidt [b], Tina Seidel [a], Matthias Stadler [b], Jochen Weller [b], Jochen Kuhn [b], Gjergji Kasneci [c]

[a] Technical University of Munich, Germany
[b] Ludwig-Maximilians-Universität München, Germany
[c] University of Tübingen, Germany

ARTICLE INFO

Keywords
Large language models
Artificial intelligence
Education
Educational technologies

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 commentary 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.

1. Introduction

Large language models, such as the Generative Pre-trained Transformer (GPT-3) (Floridi & Chiriatti, 2020), 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 (Devlin et al., 2018, Tay et al., 2022) and the underlying attention mechanism (Vaswani et al., 2017), which have greatly improved the ability of language models to handle long-range dependencies in natural-language texts. More specifically, the transformer architecture, introduced in Vaswani et al. (2017), 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 language 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 (Min et al.,

* Corresponding author.
E-mail address: Enkelejda.Kasneci@tum.de (E. Kasneci).

https://doi.org/10.1016/j.lindif.2023.102274
Received 13 February 2023; Received in revised form 28 February 2023; Accepted 1 March 2023
Available online 9 March 2023

2021). For example, Bidirectional Encoder Representations from Transformers (or BERT for short) (Devlin et al., 2018) 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 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 (Brown et al., 2020), e.g., more recently in the context of generating synthetic and yet realistic heterogeneous tabular data (Borisov et al., 2022).

Recent advancements also include ChatGPT (Team, 2022), which was trained on a much larger dataset, i.e., texts from a very large web corpus, and has 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. Moreover, as each individual has unique learning preferences, abilities, and needs, large language models offer a unique opportunity to provide personalized and effective learning experiences.

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 (Magazine, 2022) 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 complete 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 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

In recent years, several large language models have been developed, including GPT (Radford et al., 2018), BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), T5 (Raffel et al., 2020), RoBERTa (Liu et al., 2019), and the most widely used GPT-3 (Floridi & Chiriatti, 2020; Scao et al., 2022). These models are based on transformer architecture and have been pre-trained on massive datasets of text to generate human-like text, answer questions, assist in translation and summarization, and perform many NLP tasks with a single pre-training and fine-tuning pipeline. BLOOM is the latest addition to this family, developed by the BigScience-community and released as an open-source project, providing a transparently trained multilingual language model explicitly designed to cover 46 natural languages and 13 programming languages (Scao et al., 2022). These developments mark significant milestones in the field of NLP and offer enormous opportunities for applications in research and industrial contexts. We anticipate that future advancements in NLP, and specifically large language models, will lead to even more improved capabilities of language models, thus highlighting the need to explore their potential applications 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. As the field continues to develop, there are many unknown unknowns that are yet to be explored, and can only be identified and addressed through systematic and rigorous empirical research and evaluations.

2.1. Research works addressing the 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 (Dijkstra et al., 2022; Gabajiwala et al., 2022). More specifically, in a recent work by Dijkstra et al. (2022), 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 (Dijkstra et al., 2022).

In another recent work, GPT-3 was employed as a pedagogical agent to stimulate the curiosity of children and enhance question-asking skills (Abdelghani et al., 2022). 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 (Abdelghani et al., 2022).

In computing education, a recent work by MacNeil et al. (2022) 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. (2022) 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 labeling 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 (Bhat et al., 2022).

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. (2021) 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 (Ji et al., 2022), 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 (El Shazly, 2021). Another application is to support students when they experience foreign language learning anxiety (Bao, 2019) or have a lower willingness to communicate (Tai & Chen, 2020). In Jeon (2021), the application of providing feedback, as a needs analyst, and evaluator when primary school students practice their vocabulary was explored. The authors of Lin and Mubarok (2021) 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. (2022) 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 (Kung et al., 2022).

2.2. Research works addressing the 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 (Salas-Pilco et al., 2022). 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 (Hwang & Chang, 2021). 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 (Polak et al., 2022). 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 (Ayanwale et al., 2022). 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 (Choi et al., 2023). 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 Chocarro et al. (2021): 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 (Fadel et al., 2019).

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. (2022) employed a fine-tuned GPT-3 model to evaluate student-generated answers in a learning environment for chemistry education (Moore et al., 2022). 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 (Moore et al., 2022). In addition, the following studies examined NLP-based models for generating automatic adaptive feedback: Zhu et al. (2020) 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. (2023) 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. (2022) 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, Sarsa et al. (2022) 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, Qu et al. (2021) 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 (Raina & Gales, 2022; Rodriguez-Torrealba et al., 2022). Bringing language models to mathematics education, several works discuss the automatic generation of math word problems (Shen et al., 2021; Wang et al., 2021; Yu et al., 2021), which combines the challenge of understanding equations and putting them into the appropriate context.

Finally, another recent work (Tack & Piech, 2022) 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 (Tack & Piech, 2022), thus emphasizing the need for further research.

2.3. Unknown unknowns

From an education perspective, there are still many knowledge gaps and uncertainties when it comes to the successful and responsible integration of large language models into learning and teaching processes. Specifically, customizing models to specific needs, addressing biases in specific use cases, dealing with ethical considerations and copyright issues requires multidisciplinary evidence-based research and evaluation. While large language models can generate multiple-choice

questions, produce text from bullet points, and scaffold learning, it is clear that they can only serve as assistive tools to human learners and educators and cannot replace the teacher.

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) (Ahuja et al., 2023; Gao et al., 2021; Rojas-Sánchez et al., 2022), 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 (Roussou, 2001) and is in general seen as a key factor for improving the interaction between humans and AI systems (Guzman & Lewis, 2020).

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 (Ahuja et al., 2023; Kerr & Lawson, 2020).

In general, the ability of these models to generate contextualized natural language texts, code for various implementation tasks (Becker et al., 2022) as well as various types of multimedia content (e.g., in combination with other AI systems, such as DALL-E (Ramesh et al., 2021)) 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

4.1. 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.

4.2. 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.

4.3. 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 (Pavlik, 2023), 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.

4.4. 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.

4.5. 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 (Redecker et al., 2017). 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., Salomon, 1993). 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.

4.6. 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 (Cotton et al., 2023; Elkins & Chun, 2020; Gao et al., 2022; Nassim, 2021). As a result, the New York City's Department of Education recently banned ChatGPT from schools' devices and networks (News, 2023).

Just recently, Cotton et al. (2023) 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 (Tian, 2023), 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 (Gu et al., 2022; Kirchenbauer et al., 2023), 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.

4.7. 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.

4.8. 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.

4.9. 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.

4.10. 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.

4.11. 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.

4.12. 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

5.1. 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 (Kuhlthau et al., 2015), 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.

5.2. 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 organizations 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 (UNESCO, 2023): "*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 commentary 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 commentary could serve as a starting point.

CRediT authorship contribution statement

Enkelejda Kasneci: Conceptualization, Methodology, Writing – original draft. **Kathrin Sessler:** Writing – review & editing. **Stefan Küchemann:** Writing – review & editing. **Maria Bannert:** Writing – review & editing. **Daryna Dementieva:** Writing – review & editing. **Frank Fischer:** Writing – review & editing. **Urs Gasser:** Writing – review & editing. **Georg Groh:** Writing – review & editing. **Stephan Günnemann:** Writing – review & editing. **Eyke Hüllermeier:** Writing – review & editing. **Stephan Krusche:** Writing – review & editing. **Gitta Kutyniok:** Writing – review & editing. **Tilman Michaeli:** Writing – review & editing. **Claudia Nerdel:** Writing – review & editing. **Jürgen Pfeffer:** Writing – review & editing. **Oleksandra Poquet:** Writing – review & editing. **Michael Sailer:** Writing – review & editing. **Albrecht Schmidt:** Writing – review & editing. **Tina Seidel:** Writing – review & editing. **Matthias Stadler:** Writing – review & editing. **Jochen Weller:** Writing – review & editing. **Jochen Kuhn:** Writing – review & editing. **Gjergji Kasneci:** Conceptualization, Methodology, Writing – original draft.

References

Abdelghani, R., Wang, Y.-H., Yuan, X., Wang, T., Sauzéon, H., & Oudeyer, P.-Y. (2022). *GPT-3-driven pedagogical agents for training children's curious question-asking skills*. arXiv. preprint arXiv:2211.14228.

Ahuja, A. S., Polascik, B. W., Doddapaneni, D., Byrnes, E. S., & Sridhar, J. (2023). The digital metaverse: Applications in artificial intelligence, medical education, and integrative health. *Integrative Medicine Research, 12*(1), Article 100917.

Ayanwale, M. A., Sanusi, I. T., Adelana, O. P., Aruleba, K. D., & Oyelere, S. S. (2022). Teachers' readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence, 3*, Article 100099.

Bao, M. (2019). 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.

Becker, B. A., Denny, P., Finnie-Ansley, J., Luxton-Reilly, A., Prather, J., & Santos, E. A. (2022). *Programming is hard–or at least it used to be: Educational opportunities and challenges of ai code generation*. arXiv. preprint arXiv:2212.01020.

Bernius, J. P., Krusche, S., & Bruegge, B. (2022). Machine learning based feedback on textual student answers in large courses. Computers and education. *Artificial Intelligence, 3*.

Bhat, S., Nguyen, H. A., Moore, S., Stamper, J., Sakr, M., & Nyberg, E. (2022). Towards automated generation and evaluation of questions in educational domains. In *Proceedings of the 15th international conference on educational data mining* (pp. 701–704). Durham, United Kingdom: International Educational Data Mining Society.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., & Kasneci, G. (2022). *Language models are realistic tabular data generators*. arXiv. preprint arXiv:2210.06280.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Chocarro, R., Cortiñas, M., & Marcos-Matás, G. (2021). Teachers' attitudes towards chatbots in education: A technology acceptance model approach considering the effect of social language, bot proactiveness, and users' characteristics. *Educational Studies*, 1–19.

Choi, S., Jang, Y., & Kim, H. (2023). 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.

Cotton, D. R., Cotton, P. A., & Shipway, J. (2023). Chatting and cheating. In *Ensuring academic integrity in the era of ChatGPT*. EdArXiv.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. preprint arXiv: 1810.04805.

Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). *Reading comprehension quiz generation using generative pre-trained transformers*. https://e.humanities.uva.nl/publications/2022/dijk_read22.pdf.

El Shazly, R. (2021). Effects of artificial intelligence on english speaking anxiety and speaking performance: A case study. *Expert Systems, 38*(3), Article e12667.

Elkins, K., & Chun, J. (2020). Can GPT-3 pass a writer's turing test? *Journal of Cultural Analytics, 5*, 17212.

Fadel, C., Holmes, W., & Bialik, M. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. The Center for Curriculum Redesign.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681–694.

Gabajiwala, E., Mehta, P., Singh, R., & Koshy, R. (2022). Quiz maker: Automatic quiz generation from text using NLP. In *Futuristic trends in networks and computing technologies* (pp. 523–533). Singapore: Springer.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). *Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers*. bioRxiv.

Gao, H., Bozkir, E., Hasenbein, L., Hahn, J.-U., Göllner, R., & Kasneci, E. (2021). Digital transformations of classrooms in virtual reality. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–10).

Gu, C., Huang, C., Zheng, X., Chang, K.-W., & Hsieh, C.-J. (2022). *Watermarking pre-trained language models with backdooring*. arXiv. preprint arXiv:2210.07543.

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society, 22*(1), 70–86.

Hwang, G.-J., & Chang, C.-Y. (2021). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, 1–14.

Jeon, J. (2021). Chatbot-assisted dynamic assessment (ca-da) for l2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, 1–27.

Ji, H., Han, I., & Ko, Y. (2022). A systematic review of conversational ai in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education*, 1–16.

Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., & Gehringer, E. F. (2021). *ALL-IN-ONE: Multi-task learning BERT models for evaluating peer assessments*. International Educational Data Mining Society.

Kerr, J., & Lawson, G. (2020). Augmented reality in design education: Landscape architecture studies as ar experience. *International Journal of Art & Design Education, 39*(1), 6–21.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). *A watermark for large language models*. arXiv. preprint arXiv:2301.10226v1.

Kuhlthau, C. C., Maniotes, L. K., & Caspari, A. K. (2015). *Guided inquiry: Learning in the 21st century: Learning in the 21st century*. Abc-Clio.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepano, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2022). *Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models*. medRxiv.

Lin, C.-J., & Mubarok, H. (2021). Learning analytics for investigating the mind map-guided AI chatbot approach in an efl flipped speaking classroom. *Educational Technology & Society, 24*(4), 16–35.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*. arXiv. preprint arXiv:1907.11692.

MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., & Huang, Z. (2022). 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*. Association for Computing Machinery.

Magazine, T. A. A. I. (2022). *Freaky ChatGPT fails that caught our eyes!* Accessed: 2023-01-22 https://analyticsindiamag.com/freaky-chatgpt-fails-that-caught-our-eyes/.

Mengxiao Zhu, O., & Lydia Liu, H.-S. L. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*, Article 103668.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2021). *Recent advances in natural language processing via large pre-trained language models: A survey*. arXiv. preprint arXiv:2111.01243.

Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). 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, EC-TEL 2022, Toulouse, France, September 12–16, 2022, Proceedings* (pp. 243–257). Springer.

Nassim, D. (2021). Plagiarism in the age of massive generative pre-trained transformers (GPT-3). *Ethics in Science and Environmental Politics, 21*, 17–23.

News, K. R. N. (2023). *ChatGPT banned from New York City public schools' devices and networks*. https://nbcnews.to/3iTE0t6. (Accessed 22 January 2023).

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator, 78*(1), Article 10776958221149577.

Polak, S., Schiavo, G., & Zancanaro, M. (2022). 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*. Association for Computing Machinery.

Qu, F., Jia, X., & Wu, Y. (2021). 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* (pp. 2583–2593).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). *Improving language understanding by generative pre-training* Accessed: 2023-01-22.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Raina, V., & Gales, M. (2022). *Multiple-choice question generation: towards an automated assessment framework*. arXiv. preprint arXiv:2209.11830.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., … Radford, A. Sutskever (2021). Zero-shot text-to-image generation. In , *139*. *Proceedings of the 38th International Conference on Machine Learning* (pp. 8821–8831). PMLR.

Redecker, C., et al. (2017). European framework for the digital competence of educators: DigCompEdu. In *Technical report*. Seville site: Joint Research Centre.

Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022). End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications, 208*, Article 118258.

Rojas-Sánchez, M. A., Palos-Sánchez, P. R., & Folgado-Fernández, J. A. (2022). Systematic literature review and bibliometric analysis on virtual reality and education. *Education and Information Technologies*, 1–38.

Roussou, M. (2001). Immersive interactive virtual reality in the museum. In *Proc. of TiLE (Trends in Leisure Entertainment)*.

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction, 83*, Article 101620.

Salas-Pilco, S. Z., Xiao, K., & Hu, X. (2022). Artificial intelligence and learning analytics in teacher education: A systematic review. *Education Sciences, 12*(8).

Salomon, G. (1993). On the nature of pedagogic computer tools: The case of the writing partner. *Computers as Cognitive Tools, 179*, 196.

Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large language models. In , *1*. *Proceedings of the 2022 ACM conference on international computing education research* (pp. 27–43) Accessed: 2023-01-22.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). *BLOOM: A 176B-parameter open-access multilingual language model*. arXiv. preprint arXiv:2211.05100.

Shen, J., Yin, Y., Li, L., Shang, L., Jiang, X., Zhang, M., & Liu, Q. (2021). Generate & Rank: A multi-task framework for math word problems. In , *2021*. *Findings of the Association for Computational Linguistics: EMNLP* (pp. 2269–2279).

Tack, A., & Piech, C. (2022). 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* (pp. 522–529). Durham, United Kingdom: International Educational Data Mining Society.

Tai, T.-Y., & Chen, H. H.-J. (2020). The impact of google assistant on adolescent efl learners' willingness to communicate. *Interactive Learning Environments*, 1–18.

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A survey. *ACM Computing Surveys, 55*(6), 1–28.

Team, O. (2022). *ChatGPT: Optimizing language models for dialogue.* . (Accessed 19 January 2023).

Tian, E. (2023). GPTZero. https://gptzero.me/ Accessed: 22.01.2023.

UNESCO. (2023). *Education 2030 agenda*. https://www.unesco.org/en/digital-education/artificial-intelligence. (Accessed 22 January 2023).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. preprint arXiv:2212.01020.

Wang, Z., Lan, A., & Baraniuk, R. (2021). Math word problem generation with mathematical consistency and problem context constraints. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 5986–5999).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in neural information processing systems*, 32. preprint arXiv:1810.04805.

Yu, W., Wen, Y., Zheng, F., & Xiao, N. (2021). Improving math word problems with pre-trained knowledge and hierarchical reasoning. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 3384–3394).