Universidad Pablo de Olavide (España)
International Journal of Educational Research and Innovation,
número 20, 2023
ISSN: 2386-4303
DOI: 10.46661/ijeri.8449
Sección: Artículos
Recibido: 14-08-2023
Aceptado: 14-08-2023
Publicado: 15-12-2023
Páginas: 1-21

IJERI
INTERNATIONAL JOURNAL OF EDUCATIONAL RESEARCH AND INNOVATION
REVISTA INTERNACIONAL DE INVESTIGACIÓN E INNOVACIÓN EDUCATIVA

La tecnología central detrás y más allá de ChatGPT: Una revisión exhaustiva de los modelos de lenguaje en la investigación educativa

The core technology behind and beyond ChatGPT: A comprehensive review of language models in educational research

Kelvin Leong
University of Chester, United Kingdom
k.leong@chester.ac.uk

Anna Sung
University of Chester, United Kingdom
a.sung@chester.ac.uk

Lewis Jones
University of Chester, United Kingdom
lewis.jones@chester.ac.uk

RESUMEN

ChatGPT ha atraído una gran atención en el sector educativo. Dado que la tecnología central detrás de ChatGPT es el modelo de lenguaje, este estudio tiene como objetivo revisar críticamente publicaciones relacionadas y sugerir la dirección futura del modelo de lenguaje en la investigación educativa. Nuestro objetivo es abordar tres preguntas: i) cuál es la tecnología central detrás de ChatGPT, ii) cuál es el nivel de conocimiento de la investigación relacionada y iii) la dirección del potencial de investigación. Se llevó a cabo una revisión crítica de publicaciones relacionadas con el fin de evaluar el estado actual del conocimiento del modelo lingüístico en la investigación educativa. Además, sugerimos un marco rector para futuras investigaciones sobre modelos lingüísticos en educación. Nuestro estudio respondió rápidamente a las preocupaciones planteadas por el uso de ChatGPT en la industria educativa y proporciona a la industria una descripción general completa y sistemática de las tecnologías relacionadas. Creemos que esta es la primera vez que se realiza un estudio para revisar sistemáticamente el nivel de conocimiento del modelo lingüístico en la investigación educative.

PALABRAS CLAVE

ChatGPT; modelo de lenguaje; tecnología educativa; IA.

Cómo citar: Leon, K., Sung, A. & Jones, L. . (2023). La tecnología central detrás y más allá de ChatGPT: Una revisión exhaustiva de los modelos de lenguaje en la investigación educativa. *IJERI: International Journal of Educational Research and Innovation*, (20), 1-21. 10.46661/ijeri.8449

ABSTRACT

ChatGPT has garnered significant attention within the education industry. Given the core technology behind ChatGPT is language model, this study aims to critically review related publications and suggest future direction of language model in educational research. We aim to address three questions: i) what is the core technology behind ChatGPT, ii) what is the state of knowledge of related research and iii) the potential research direction. A critical review of related publications was conducted in order to evaluate the current state of knowledge of language model in educational research. In addition, we further suggest a purpose oriented guiding framework for future research of language model in education. Our study promptly responded to the concerns raised by ChatGPT from the education industry and offers the industry with a comprehensive and systematic overview of related technologies. We believe this is the first time that a study has been conducted to systematically review the state of knowledge of language model in educational research.

KEYWORDS

ChatGPT; Language model; EdTech; AI.

1. INTRODUCTION

Since its introduction, ChatGPT has garnered significant attention within the education industry. Many educators were shocked by its ability on generating human-like responses, answer questions, and perform other language tasks.

In fact, ChatGPT's capability has caused a stir in the education industry. There is a fear that students may use the tool for cheating (Hess, 2023; Westfall, 2023) and simply generate answers without fully understanding the underlying concepts, leading to a lack of critical thinking and problem-solving skills (Fitzpatrick, 2023; Meisner, 2023). Additionally, there is a worry that the use of such tools may lead to a decrease in the quality of education and a devaluation of academic degrees (Dempsey, 2023). Moreover, use of ChatGPT for completing assignments can also lead to academic dishonesty and compromise the integrity of academic institutions. If the assignments and assessments are not authentic, the value of the education is diminished and it becomes difficult to accurately evaluate student performance (Gift & Norman, 2023). This can also lead to a decrease in the overall quality of education and limit the potential for future employment and career opportunities (Greenhouse, 2023; Salim, 2023). Furthermore, the widespread use of language models like ChatGPT for completing assignments can contribute to a culture of shortcuts and laziness (Baron, 2023; Murray, 2023) where students prioritize the quickest and easiest solution over learning and hard work. This can have long-term implications for the development of the workforce and the economy as a whole.

The core technology behind the ChatGPT is language model. However, language models have demonstrated its ability to bring positive changes on education. A narrow focus on the potential negative impacts of ChatGPT will result in overlooking the potential benefits and opportunities that technology can bring to the learning process. In fact, while the industry is still worried about the negative impact of ChatGPT, many technology giants are ready to launch new products to compete with ChatGPT, such as Bing Chatbot, Google Bard, etc. (Q.ai, 2023).

In brief, this study aims to address three questions: i) what is the core technology that makes ChatGPT powerful, ii) what is the state of knowledge of related research and iii) the potential research direction. Building on this evidence-based research, the industry can make informed decisions and be better equipped to explore the opportunities of language model in education.

The rest of this paper is organised as follows: we first review the core technologies behind the ChatGPT (i.e. language model) in the next section. After that, we discuss the current state of knowledge and the advances of language model in educational research. Building on the understanding obtained from our review, we suggest the future research direction to the field. Finally, we summarise our study and contributions in the last section.

2. THE CORE TECHNOLOGY BEHIND THE CHATGPT: LANGUAGE MODEL AND ITS KEY DEVELOPMENT

The core technology behind the ChatGPT is language model. A language model can generate human-like text based on the learned patterns from large amounts of text and data it was trained on (van Dis et al., 2023).

Throughout the evolution of language models, two key phases can be identified: i) statistical-based model and ii) neural network-based model (Li, 2022).

2.1. Statistical-based model

As per Li (2022), statistical-based model was the language model used in the early days of language modelling. In brief, statistical-based model is the development of probabilistic model that can predict the next word in the sequence given the words that precede it. The most popular statistical-based models is the n-gram model (Rosenfeld, 2000) which has been widely used in various applications. Some examples include auto-completion of sentences, auto spell-check and semantic analysis. However, the n-gram model faced several limitations such as its inability to handle long-range dependencies between words and its inability to understand semantic information. To overcome these limitations, researchers started exploring new approaches to language modelling which eventually led to the development of neural-network based models.

1.2. Neural-network based model

Neural-network based models use artificial neural networks to model the complex relationships between words in a sentence. The underlying principle of neural-network based models is to utilize multi-layered neural architectures in order to grasp the hierarchical composition of language (Han et al., 2021). Some key breakthroughs in this area include the very first feed-forward neural network language model proposed by Bengio and Senecal (2008) and the word2vec model (Goldberg & Levy, 2014), etc. In recent years, there has been a significant increase in the development of transformer (Vaswani et al., 2017) and pre-trained models. The representative pre-trained models include ELMO (Peters et al., 2018) and BERT (Devlin et al., 2018). The language model behind the eye-catching ChatGPT is GPT-3. The first generation of it was GPT (Generative Pre-trained Transformer) model introduced by OpenAI. GPT's family uses the transformer architecture and is trained on a large corpus of text data. GPT-2 and GPT-3 (Han et al., 2021; Li, 2022) are follow-up models that are even larger and more powerful.

In conclusion, from the early days of statistical models to the most current developments in neural network-based models, the development of language models has come a long way. The shift from statistical-based models to neural-network-based models has been a major breakthrough in the field of NLP. The key language models and their major breakthroughs are summarized in table 1.

Table 1. SUMMARY of the representative language models and their major breakthrough.

Statistical based/Neural-network based	Representative Language Models	Major Breakthrough	Years
Statistical based	n-gram models, Hidden Markov Models (HMMs)	The development of these models marked the early stage of language modeling.	1960s-1980s

Statistical based/Neural-network based	Representative Language Models	Major Breakthrough	Years
Neural-network based	Word2Vec	Representation of words in continuous vectors, which can capture semantic relationships between words.	2013
Neural-network based	RNNs and LSTMs	Representation of sequential data, which allow the models to consider previous words to predict the next word.	1990s-2010s
Neural-network based	Transformers	Representation of multi-head self-attention mechanism, which allows the model to attend to different positions and information from the input sequence.	2017
Neural-network based	Pre-trained Models (BERT, GPT, etc.)	Pre-training on a large corpus and fine-tuning on specific tasks to achieve state-of-the-art results.	2018-Present

3. A REVIEW OF CURRENT STATE OF KNOWLEDGE OF LANGUAGE MODEL IN EDUCATIONAL RESEARCH

A critical review of related publications was conducted in order to evaluate the current state of knowledge of language model in educational research. The review's design and findings are detailed below.

3.1. The review design

The data source used in this review is the Web of Science database (*https://www.webofscience.com/wos/woscc/basic-search*) with the extraction date on the 21st January 2023. The Web of Science database is a comprehensive platform which provides access to high-quality and reliable research articles, conference proceedings and book chapters (López-Belmonte et al., 2021; Khushk et al., 2023).

Since the purpose of our research focuses on the language model in education, therefore we set the criteria of publication search from the database as below.

We searched publications containing the term "language model" in all selected fields under the research area "Education Educational Research" in Web of Science. In total, we obtained 111 related publications including the first publication found in 2012. The search criteria used in this study ensured that only relevant articles were considered and only those articles that were related to language model in education were included in the final data set.

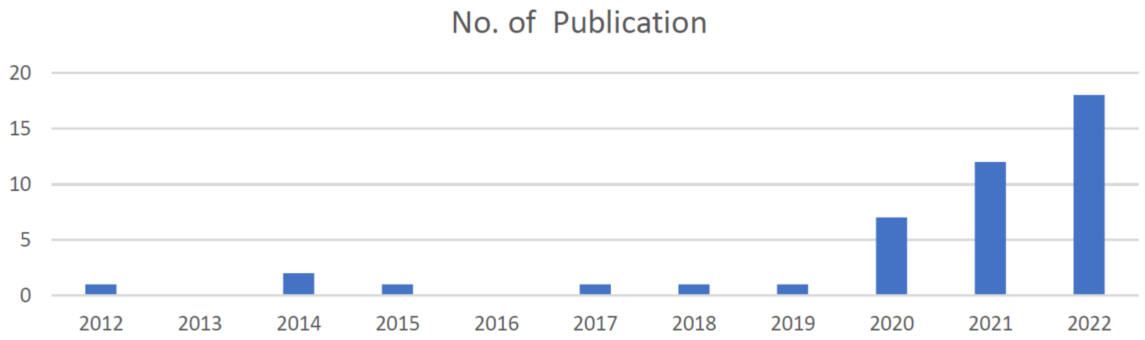
However, not all the 111 publications obtained are relevant to the purpose of our study. For example, some papers referred to the language model from linguistics point of view which is not relevant to computer technology. Therefore, we reviewed each of the publications and identified 44 publications as being relevant to this research. The process of reviewing each publication was conducted systematically and rigorously to ensure that only the most relevant publications were included. The process involved reading each publication in full, assessing its relevance to the research and evaluating its contribution. Publications that were not relevant to language models in education were discarded. The entire list of the 44 identified publications was discussed in section 3.2.3.

The final set of the 44 identified publications was then used for the analysis and synthesis of the research.

3.2. Overall trend of related research.

As per Figure 1, the trend of the identified publications appears to suggest a generally low level of output throughout the period from 2012 to 2019. However, starting in 2020 there is a significant increase in output with a value of 7 and a further increase to 12 in 2021, culminating in an even higher output of 18 in 2022. The trend shows a significant increase in recent years, suggesting rapid growth in the field. This is evident by the increase in the number of related publications in a short period of time.

Figure 1. The trend of the identified publications since the first output appeared in 2012.



In order to critically review the state of arts of the knowledge of the field, the 44 identified publications were evaluated through three lenses: i) sources of the identified publications, ii) who were the contributors and iii) what were the contributions.

3.2.1. Sources of the identified publications.

Table 2 summarizes the distribution of the identified publications by types. As shown in the table, 25% of previous outputs were published as journal articles while 75% of previous outputs were published as conference proceedings. This finding suggests that a significant proportion of the identified publications are being presented and disseminated at conference venues rather than being submitted to peer-reviewed journals for publication.

Table 2. The distribution of the 44 identified publications by types.

Sources	Total (counts)	Total In%
Journal articles		
IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES	2	5%
INTERNATIONAL JOURNAL OF EMERGING TECHNOLOGIES IN LEARNING	1	2%
JOURNAL OF COMPUTING IN HIGHER EDUCATION	1	2%
JOURNAL OF SCIENCE EDUCATION AND TECHNOLOGY	1	2%
ONLINE LEARNING	1	2%
VOPROSY OBRAZOVANIYA-EDUCATIONAL STUDIES MOSCOW	1	2%
BRITISH JOURNAL OF EDUCATIONAL TECHNOLOGY	1	2%
EDUCATION AND INFORMATION TECHNOLOGIES	1	2%
EDUCATIONAL MEASUREMENT-ISSUES AND PRACTICE	1	2%
INTERACTIVE LEARNING ENVIRONMENTS	1	2%

Sources	Total (counts)	Total In %
Conference proceedings		
13th International Conference on Computer Supported Education (CSEDU)	1	2 %
13th International Conference on Emerging eLearning Technologies and Applications (ICETA)	1	2 %
14th Annual Conference on Learning Ideas - Innovations in Learning and Technology for the Workplace and Higher Education	1	2 %
17th European Conference on Technology Enhanced Learning (EC-TEL)	1	2 %
17th IEEE International Conference on Machine Learning and Applications (IEEE ICMLA)	1	2 %
17th International Conference on Intelligent Tutoring Systems (ITS)	2	5 %
20th IEEE International Conference on Advanced Learning Technologies (ICALT)	1	2 %
20th International Conference on Artificial Intelligence in Education (AIED)	1	2 %
21st International Conference on Artificial Intelligence in Education (AIED)	3	7 %
22nd International Conference on Artificial Intelligence in Education (AIED) - Mind the Gap - AIED for Equity and Inclusion	4	9 %
23rd International Conference on Artificial Intelligence in Education (AIED)	9	20 %
34th AAAI Conference on Artificial Intelligence / 32nd Innovative Applications of Artificial Intelligence Conference / 10th AAAI Symposium on Educational Advances in Artificial Intelligence	2	5 %
35th AAAI Conference on Artificial Intelligence / 33rd Conference on Innovative Applications of Artificial Intelligence / 11th Symposium on Educational Advances in Artificial Intelligence	1	2 %
4th World Conference on Learning, Teaching and Educational Leadership (WCLTA)	1	2 %
6th International Conference on Education and New Learning Technologies (EDULEARN)	1	2 %
7th International Learning Analytics and Knowledge Conference (LAK)	1	2 %
IEEE 4th International Conference on Technology for Education (T4E)	1	2 %
IEEE International Conference on Teaching, Assessment, and Learning for Engineering (IEEE TALE)	1	2 %
Total	44	100 %
Breakdown:		
Journal articles	11	25 %
Conference proceedings	33	75 %

Conference proceeding is an important outlet for research dissemination. Conference provides a forum for researchers to present their work in a timely manner which is especially important in fields where interdisciplinary work is prevalent and where there is a need for quick dissemination of information.

In sum, the distribution pattern found according to the 44 identified publications by types reflects the language model related research in the education field is a fast-changing interdisciplinary topic. Consequently, timely dissemination of results is preferred and the community within this field values the opportunity for discussion and feedback that conference provides.

3.2.2. Who were the contributors

In this study, we focus on two types of contributors: affiliation and funding agency.

Figures 2 and 3 show the distributions of publication trends by affiliations and by funding agencies respectively in three periods: i) 2012 to 2015, ii) 2016 to 2019 and iii) 2020 to 2022. Their trends are discussed as follows.

- Publication trends by affiliations

During the first period from 2012 to 2015, there were only 4 papers published, involving authors from 8 different affiliations from around the world. This low level of publication may reflect the early stages of research in this field or limited awareness of the potential of language models for education.

From 2016 to 2019, the number of publications decreased to 3, with authors from 5 different affiliations involved in the research. The relatively low number of publications suggests that the field was still in its early stages of development.

During the most recent period from 2020 to 2022, the number of publications skyrocketed to 37, with authors from 74 different affiliations involved in relevant research. This substantial increase in publication activity suggests that the field of language models in education has rapidly grown and matured over the past few years. The large number of different affiliations involved in this research indicated the increasing recognition of the potential of language models to impact education as well as a growing interdisciplinary interest in the topic. In addition, the significant rise in affiliations involved in the research suggests a higher level of collaboration approach in this field.

- Publication trends by funding agencies.

From 2012 to 2015, only 4 papers were published and involved 2 funding agencies. This represents a relatively small amount of research in this area and highlights the limited investment in this area of study.

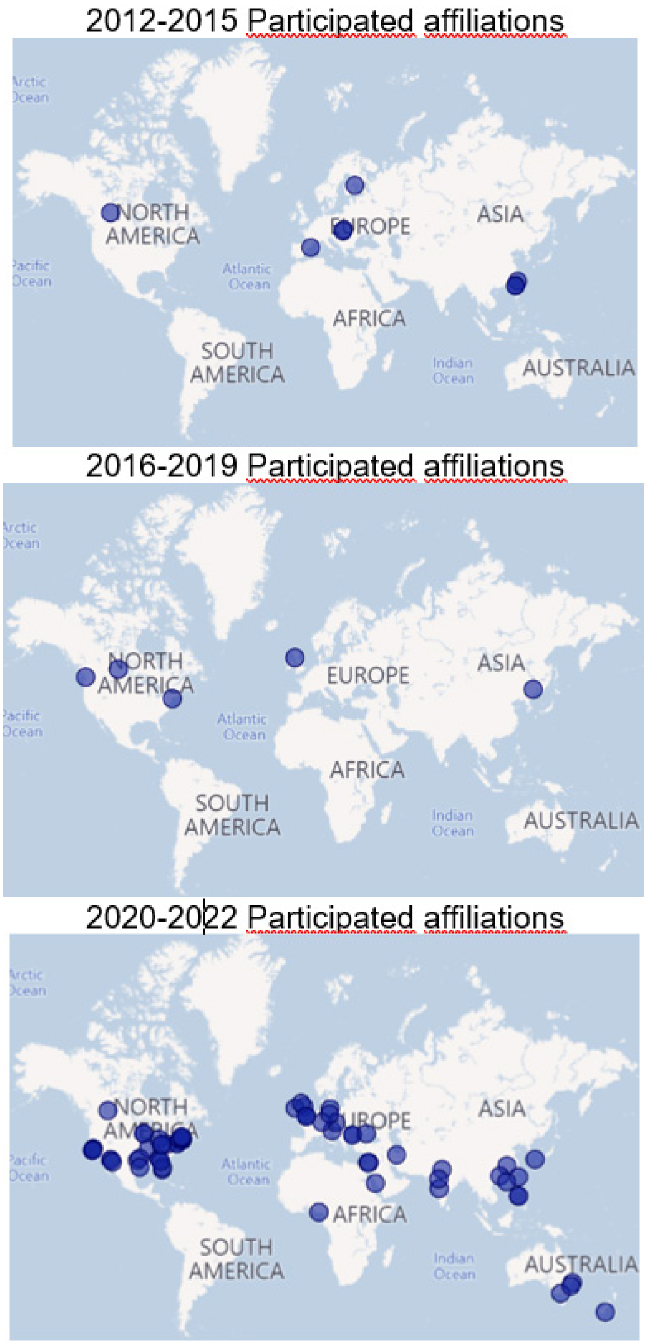
During the period of 2016 to 2019, a decrease in the number of related language model in education papers was observed with only 3 papers being published. The absence of funding agencies in this period could be attributed to various reasons, and the trend have been temporary.

In fact, from 2020 to 2022, the trend shifted dramatically with the publication of 37 papers and involvement of 34 funding agencies. This surge in research activity highlights the increasing importance of language model research in education with funding agencies investing in this area to better understand its applications and benefits. The trend of increased investment in language model research in education by funding agencies may be due to a growing recognition of the potential benefits of technology in education. The increasing involvement of funding agencies in this area of study may also be a result of interdisciplinary collaboration, as researchers from a range of fields work together to understand the impact of language models on education.

- Involvement of commercial organizations in related outputs.

We also observed the involvement of commercial organizations in related outputs. We consider the involvement with commercial organizations can bring many positive implications. Firstly, the involvement can lead to increased applied research with practical, market-driven solutions. Secondly, the increase in affiliations can foster collaboration and knowledge sharing across different organizations and disciplines. This could lead to a more interdisciplinary and integrated approach to research, resulting in better and more comprehensive solutions to complex problems. Additionally, commercial organizations can bring technical expertise, resources and market knowledge to academic research, leading to more efficient and effective research outcomes. In fact, the increasing number of research that involves authors or funding agencies from commercial organizations is a reflection of the growing importance of industry-academia collaborations in driving innovation and generating economic impact. Studies have shown that industry-academia collaborations can lead to more impactful research outcomes, increased innovation and acceleration of the commercialization (Leong et al., 2020; Leydesdorff & Etzkowitz, 2003; Perkmann et al., 2013).

Figure 2. Publication trends by affiliations.

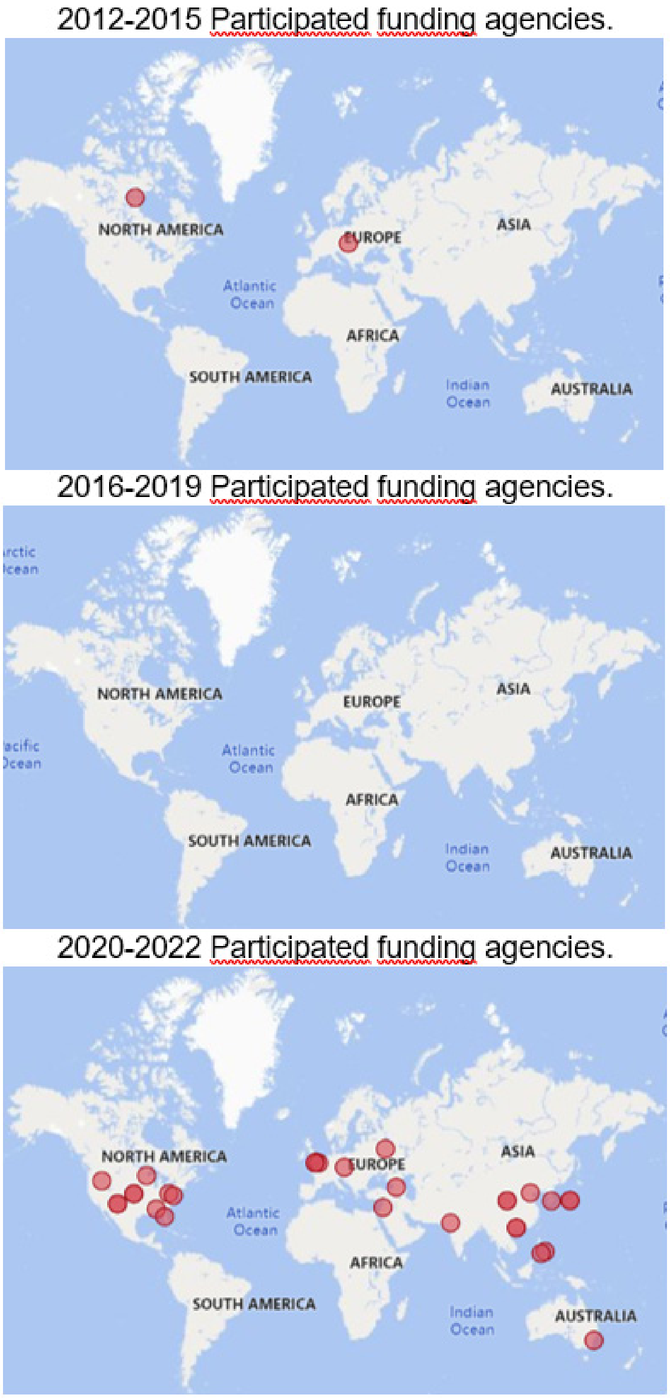


2012-2015 Participated affiliations

2016-2019 Participated affiliations

2020-2022 Participated affiliations

Figure 3. Publication trends by funding agencies.



3.2.3. What were the contributions

By analysing the 44 identified publications, the contributions of related research have been categorized into five distinct categories based on their focuses. These categories are i) assessment-related research, ii) learning support-related research, iii) instructional material preparation-related research, iv) student feedback-related research, and v) new language model-related research.

Table 3. The distribution pattern of the 44 identified publications by categories of focuses.

Categories of focuses	2012	2014	2015	2017	2018	2019	2020	2021	2022	Total
Assessment		1				1	4	5	7	18
learning support		1	1	1			2	6	6	17
Instructional Material preparation	1				1		1		2	5
New domain-specific Language model									1	1
Student Feedback								1	2	3
Total	**1**	**2**	**1**	**1**	**1**	**1**	**7**	**12**	**18**	**44**

As per table 3, assessment and Learning support related research have seen an increase in recent years with the highest count being in year 2022 with 7 articles and 6 articles respectively. A comprehensive analysis of these five categories of research focus is detailed in the subsequent sections of this study.

- Assessment.

The use of language models to support assessments can be broadly classified into two categories: assessment preparation and assessment scoring/grading. Assessment preparation-related studies focus on using language models to improve the quality of assessment questions, revise grammatical errors or generate new questions for students. Assessment scoring/grading-related studies aim to use language models to grade student's answers accurately and provide constructive feedback.

In the area of assessment preparation, Li et al. (2020) presented results indicating the strong potential of using BERT in the grammatical error correction task. More specifically, the authors demonstrated that BERT outperformed traditional grammatical error correction models, demonstrating the effectiveness of deep learning models in solving such problems. On the other hand, Liu et al. (2021) took this a step further by proposing a BART-based neural framework to solve K-12 ESL Sentence Completion questions. The proposed work provides feasible solutions that can help teachers revise and improve the overall quality of Sentence Completion questions. The results demonstrate the superiority of the proposed work in terms of prediction accuracy.

Another interesting development in this area is the use of language models to solve complex problems in academic courses. For example, Tang et al. (2022) introduced a model that can be used to solve problems in MIT's Probability and Statistics course. This work has the potential to contribute to both assessment preparation and checking, providing a valuable tool for students and teachers alike. Moore et al. (2022) proposed a work that can generate and evaluate short answer questions for assessment. The authors found that students are relatively capable of generating short answer questions that can be leveraged in their online courses, providing an opportunity for students to take a more active role in their own learning. Similarly, Wang et al. (2022) used a large pre-trained language model (PLMs) for automatic educational assessment question generation. The experiment results showed that subject matter experts could not effectively distinguish between generated and human-authored questions, suggesting that PLMs have the potential to create high-quality assessment questions.

Overall, the trend and development in the use of language models for assessment preparation suggest a strong interest in leveraging relevant technologies to improve the quality of assessments and make the assessment process more efficient and effective.

On the other hand, assessment scoring and grading related studies focus on using language models to grade or score written responses. One of the earliest related studies in this area was published by Lee et al. (2014) who proposed a part-of-speech (POS) based language model to detect grammatical errors committed by ESL/EFL learners. This study demonstrated the feasibility of using language models to automate the grading of writing assignments and laid the foundation for further research in this area.

Sung et al. (2019) focused on short answer grading on a dialogue-based tutoring platform and reported up to 10% absolute improvement in macro-average-F1 over state-of-the-art results. Condor (2020) presented an automatic short answer grading model, which could serve as a second opinion to confirm raters' correctness and improve consistency of judgement.

The use of language models for automated grading has been further improved by the introduction of novel approaches or new datasets. For examples, Ndukwe et al. (2020) introduced a grading system that can support the marking of text-based questions, and the results showed good inter-rater agreement while Khot et al. (2020) proposed a dataset that can help with question preparation and the results showed an improvement of 11% (absolute) over current state-of-the-art language models.

In addition to grading text-based questions, researchers have also explored the potential of language models to grade essay writing assignments. Xu et al. (2021) proposed a framework that can detect whether an essay responds to a requirement question and clearly mark where the essay answers the question. Beseiso et al. (2021) demonstrated the applicability of their model in automated essay scoring at the higher education level. Makhlouf & Mine (2021) introduced an automated feedback system that can generate appropriate feedback with 81% accuracy. Botarleanu et al. (2021) proposed a text summarization approach for automated scoring which showed accurate and robust results while ensuring generalizability to a wide array of topics. Zhu et al. (2022) proposed a work on automatic short answer grading that outperformed existing methods. Fernandez et al. (2022) presented an automated scoring approach that significantly reduced human grader effort and outperformed existing methods. Rakovic et al. (2022) proposed a work for evaluating and providing formative feedback to first year law students on case note writing. The use of language models for automated grading has also been combined with other methods to improve its performance. For instance, Firoozi et al. (2022) integrated active learning methods to support automated essay scoring.

- Learning support.

Previous studies have demonstrated the potential of using language models to provide students with personalized feedback, emotional support and a more engaging learning experience.

One common thread among the studies is related to language usage and linguistics in education. For example, Lopez-Ferrero et al. (2014) presented experiments in automatic correction of spelling and grammar errors in Spanish academic texts with the goal of developing a tool to assist university students in writing academic texts. Ondas et al. (2015) focused on the linguistics analysis of written or spoken Slovak texts which can contribute to learning activities in the local context. Dimzon and Pascual (2020) developed a phoneme-based forced aligner and recognizer for children's Filipino read speech. Nicula et al. (2021) developed a method to automatically assess the quality of paraphrases which can be very useful in facilitating literacy skills and providing timely feedback to learners.

Language models can also be used to identify and address specific personal challenges faced by students. For examples, Du et al. (2021) presented an approach to generate text automatically for emotional and community support targeting a massive online learning community. Geller et al. (2021) developed a model to detect confusion expressed among students' comments in course forums. Dyulicheva (2021) studied the emotional states of learners associated with math anxiety on MOOCs. Lu et al. (2021) provided support in generating personalized responses that correspond to the context of a conversation.

Online learning is another focus of previous studies. Wise et al. (2017) presented an approach for identifying and classifying learners on MOOC forums based on their posts while Xu et al. (2020) presented a study related to online one-on-one class and automatic dialogic instruction detection. The experiments demonstrated that the approach achieved satisfactory performance based on the real-world educational dataset. Hao et al. (2021) presented an approach to detect online dialogic instructions to help students understand learning materials. The researchers conducted experiments on a real-world online educational data set and the results showed that the proposed approach achieved superior performance compared to baseline methods. Hsu & Huang (2022) proposed an intelligent question answering bot for Chinese MOOCs while Lee et al. (2022) built a model to help online students to develop higher-order thinking and to measure and enhance the quality of interactions in discussion forums at scale. The results showed that the approach achieved 92.5% accuracy on the classification task. Ba et al. (2023) presented an approach for providing coaching in inquiry-based online discussions which contributed to cognitive presence development and higher-order thinking.

Other interesting works include: Wulff et al. (2022) assessed preservice physics teachers' attention to classroom events as elicited through open-ended written descriptions. Nehyba & Stefanik (2022) suggested a semi-automated approach in providing student feedback and measuring the effect of systematic changes in the educational process via students' responses. Jayaraman & Black (2022) built an intelligent question-answering system based on BERT to help students learn personal finance.

In overall, these studies have also demonstrated the ability of language models to detect and measure various aspects of learning, such as confusion, cognitive presence and higher-order thinking.

- Instructional Material preparation.

The field of language models has also seen several applications in the realm of instructional material preparation.

One of the related research focuses was on supporting contents selection in the age of big data. With millions of text contents and multimedia being published on the internet, converting these into suitable learning materials requires a significant effort from teachers and instructional designers. To address this issue, Yang et al. (2012) presented a personalized text-based content summarizer to help learners to retrieve and process information more efficiently based on their interests and preferences while Lin (2020) aimed to overcome the challenge of finding suitable information from a large repository of knowledge and to satisfy users' personalized online learning requirements. The proposed approach filtered out unsuitable learning materials and ranked the recommended learning materials more effectively.

Another related research focus was on contents generation. Pan (2018) proposed automatic generation of children's songs with the aim of making the process more efficient and effective. The experiment confirmed the effectiveness of the proposed method while Parasa et al. (2022) aimed to deploy a method for automatically generation of conceptual riddles in online learning environments. The riddles generated by the model were evaluated by human evaluators and the results were encouraging. Moreover, in the field of education, automatic readability assessment is a method of evaluating the difficulty level of texts used in the classroom. Ibanez et al. (2022) presented an approach that supports automatic readability assessment with results showing that it performs significantly better than conventional transfer learning approaches.

In overall, these studies not only improved the efficiency and effectiveness of instructional material preparation, but also enabled more personalized and effective learning experience for learners.

- Student feedback.

The trend in the use of language models in student feedback can be seen as a development towards more efficient and accurate ways of measuring students' opinions and perceptions. Previous studies (Esmaeilzadeh et al., 2022; Masala et al., 2021; Xiao et al., 2022) show the advancements in this area of research.

Masala et al. (2021) presents a method of reducing the volume of text while predicting course ratings from student open-text feedback with a mean average error increase of only 0.06. This is a significant improvement in terms of efficiency as it reduces the amount of data to be analyzed and allowing for faster and more accurate analysis of student feedback. The work of Xiao et al. (2022) focuses on getting insights into students' perceptions about course quality in MOOCs. Their method provides high and stable coverage of students' ideas, allowing for a more comprehensive understanding of students' experiences. In addition, Esmaeilzadeh et al. (2022) presents a framework for encoding textual data and analyzing surveys with open-ended responses. This framework reduces the cost of analyzing large amounts of survey data by automating the process of extracting the most insightful information pieces. This is a significant advancement as it allows for faster and more efficient analysis of student feedback.

Overall, these studies show the development and trend towards more efficient and accurate methods of measuring students' opinions and perceptions through the use of language models. By reducing the volume of text to be analyzed, increasing the coverage of students' ideas and automating the analysis process, the accuracy and efficiency of student feedback analysis are improved. These lead to a better understanding of students' experiences.

- New domain-specific Language Model.

Domain-specific language models refer to machine learning models that are trained on a specific domain of data, such as finance education, work-based learning, etc. These models are designed to capture the specific patterns, vocabulary and linguistic structures. Doman-specific models are able to provide more accurate and relevant results compared to generic models that are trained on a large corpus of text data. This is because domain-specific models are trained on data that is directly relevant to a specific field.

According to our review, there was only one study (Goel et al., 2022) related to new domain-specific language model development. In the study, the researchers proposed a new language model for K-12 education. Given the potential of domain-specific language model is huge, more efforts could be given as future research direction in this area on exploring more various types of domain specific language model for education purpose.

Table 4 summarise the contributions by categories of focuses.

Table 4. Categories of focuses and corresponding contributions.

Categories of focus	What language model contributed
Assessment	• Improve the quality of assessment questions, revise grammatical errors, or generate new questions for students. • Grade student's answers accurately and provide constructive feedback
Learning support	• Detect and measure various aspects of learning, such as confusion, cognitive presence, and higher-order thinking
Instructional Material preparation	• Improve the efficiency and effectiveness of instructional material preparation, also enable more personalized and effective learning experience for learners
Student feedback	• Offer more efficient and accurate methods of measuring students' opinions and perceptions.
New domain-specific Language Model	• Provide more accurate and relevant results compared to generic models

4. BEYOND CHATGPT: A GUIDING FRAMEWORK FOR FUTURE RESEARCH

In this section, we introduce a guiding framework for future research. The guiding framework is purpose oriented. From left to right, the framework started with the purposes and then corresponding example use cases are provided. On the right-hand side, examples of language model are provided as references. In addition, we further suggest a list of evaluation measures for reference.

Table 5. The guiding framework for future research of language model in education.

Purposes	Example use cases	Examples of Language Model
Text Classification	• automatically categorizing learning materials, such as articles, videos, and books, into relevant subjects or topics. • automatically grading student essays based on their content and writing style. • classifying questions asked by students and automatically providing appropriate answers. • automatically classifying student behaviors and learning activities, such as their study habits and engagement levels, to help educators better understand their students and provide targeted support.	BERT, ELMo, ULMFit, XLNet, RoBERTa, ALBERT, T5, etc.
Sentiment Analysis	• automatically process and categorize student feedback from surveys, online forums, or other sources. • monitor and analyze student opinions and attitudes towards education, institutions, and learning programs on social media. • Analysis can be used to evaluate the effectiveness of instructional materials, such as textbooks, videos, or online courses. • monitor and analyze students' engagement levels during online or in-person classes.	
Question Answering (QA)	• automatically create a knowledge base of frequently asked questions and their answers • QA system can be trained to provide personalized feedback to students based on their performance. • QA system can be used to get instant answers to their questions,	
Named Entity Recognition	• identify named entities in student essays, such as proper nouns and dates, and use this information to grade the essays based on criteria such as accuracy, completeness, and relevance. • identify named entities in student profiles, such as their interests, background, and prior knowledge, to personalize the content they receive.	
Text Summarization	• summarize lengthy classroom lectures or presentations to make the content more digestible and easier to understand. • generate summaries of online discussions or forums to summarize key takeaways and important topics discussed. • provide an overview of assessment results, such as test scores or exam reports. • summarize evaluations, feedback or surveys from students.	

Purposes	Example use cases	Examples of Language Model
Text Generation	• assist students in writing essays by providing suggestions. • generate exam questions based on specific topics or learning objectives. • provide customized feedback and guidance to students based on their learning style and pace.	GPT-2, GPT-3, XLNet, CTRL, T5, Seq2Seq, BERT.
Dialogue Generation	• create conversational interfaces for educational content. • generate prompts or questions to encourage students to engage with the material and deepen their understanding. • create conversational interfaces for educational content. • generate realistic scenarios for learning in fields such as medicine or law.	
Text Completion	• to generate responses to student questions, providing instant feedback and helping students learn more efficiently. • generate educational content such as summaries, outlines, and study materials.	

Evaluation

Suggested measure	Notes
Content Relevance	This measures the ability of the language model to understand the context of a given task and produce relevant content.
Coherence and Consistency	This measures the ability of the language model to generate coherent and consistent content.
User Experience	This measures the ability of the language model to provide a positive user experience.
Contextual Relevance	This measures the ability of the language model to generate content that is contextually relevant.
Style and Tone	This measures the ability of the language model to generate content in a style and tone that is appropriate for the target audience and context.
Error Analysis	This measure involves evaluating the types and frequency of errors made by the language model. This can include determining how often the model provides incorrect answers, how well it handles out-of-domain examples, and how well it generalizes to new data.
Model Evaluation Metrics	Related measures can be used to measure the performance of a language model on specific tasks. For example, in sentiment analysis, metrics such as accuracy, precision, recall, F1-score, and ROC curve can be used to evaluate the performance of the model.

5. CONCLUSION

ChatGPT has garnered significant attention within the education industry. This paper starts with discussing the major concerns of ChatGPT from participators in education industry. We then introduced the core technology behind ChatGPT - language model. The design and findings of our review based on language model related publications in educational research were reported. The overall trend of the outputs shows a low level from 2012 to 2019, but a significant increase starting in 2020 with a value of 18 in 2022. We specifically conducted the review through three lens: i) sources of research, ii) who were the contributors and iii) what were the contributions. We found the majority of publication types of related research is conference proceedings. We also observed that the involvement of commercial organizations in related outputs. Moreover, the contributions of related outputs have been categorized into five distinct categories based on their focus. Most of the related works were contributed to assessment and learning support during the period.

We further suggest a purpose oriented guiding framework for future research of language model in education and a list of evaluation measures for reference.

5.1. Limitations and potential future research

We identified three key limitations of this study. For each of identified limitation, we also propose potential future research to mitigate these limitations.

Foremost, it’s crucial to acknowledge that this study relied on literature sourced from the Web of Science database. Although the Web of Science is widely respected for hosting scholarly literature, it’s important to realize that its coverage might not be all-encompassing. This limitation could unintentionally omit valuable works from sources not included in the database, thus constraining the breadth of the analysis. In the future, researchers could enhance their approach by incorporating a more diverse range of sources. This would entail using multiple reputable databases and academic literature sources. By broadening the scope to encompass more than just the Web of Science, researchers can get a more complete view of the research landscape and reduce the potential bias that may come from relying on just one database.

Secondly, it’s important to take into account that the field of language model applications in education is constantly developing. While our current study has thoroughly included all relevant research studies available through the Web of Science, it’s possible that new breakthroughs and innovative application ideas have emerged after the timeframe of our study. Particularly, the rapid growth of new language model solutions following the introduction of advancements like GPT-4 might not have been fully covered in our research. Recognizing the ever-evolving nature of language model applications in education, future research should consider adopting a longitudinal design. This design would entail periodic reviews to capture the latest advancements and application ideas beyond the scope of the current study. By doing so, researchers can bridge the temporal gap that may lead to a lack of acknowledgment for recent innovations. This iterative approach would enable a more accurate reflection of the dynamic landscape and a deeper understanding of emerging trends.

The third limitation is that the study mainly focuses on research done primarily in English. Consequently, a substantial risk exists that significant contributions from non-English sources may have been inadvertently disregarded, leading to an incomplete assimilation of critical perspectives and insights. To address this limitation, future research endeavors could integrate a broader linguistic perspective. This entails a deliberate effort to access and incorporate research published in languages other than English. Researchers can collaborate with multilingual experts, employ translation services or collaborate with scholars proficient in different languages to ensure the inclusion of critical studies and viewpoints. Such an approach would contribute to a more comprehensive and inclusive understanding of the subject matter.

5.2. Contributions of this study

The research conducted on language models has contributed significantly to the educational community on multiple fronts. To begin with, this study’s prompt response to concerns about ChatGPT within the education industry is of paramount importance. By providing a comprehensive and systematic overview of the impacts and potential implications of related technologies, the research serves as a crucial resource. It helps the industry gain clarity, avoid misconceptions, and dispel unnecessary worries surrounding ChatGPT and similar technologies, fostering a more informed and confident adoption of AI-driven tools.

Furthermore, this study's novel approach of systematically reviewing the state of knowledge pertaining to language models in educational research marks a significant advancement. By offering a consolidated repository of insights, it not only aids in understanding the current landscape but also paves the way for more focused and meaningful research endeavors. As an inaugural exploration of this kind, the study acts as a foundational reference point for future investigations in this burgeoning field.

Equally noteworthy is the study's provision of a guiding framework for forthcoming research on language models in education. This framework not only informs the scholarly community but also extends its impact to policymakers, educators, and practitioners. By offering a structured roadmap, the research empowers decision-makers to harness the potential of language models effectively, facilitating evidence-based implementation that aligns with pedagogical goals and student needs.

In conclusion, this research's usefulness to the educational community is manifold. It dispels uncertainties, establishes a solid foundation for further studies, and equips stakeholders with the tools needed to integrate language models thoughtfully into the educational ecosystem. As technology continues to shape the future of education, the study's contributions hold the promise of fostering a more effective, efficient, and adaptive learning environment for students and educators alike.

Given language model in education is a topic of educational technology. The findings from our critical review contribute to educational technology publication. In fact, the market size of educational technology is huge and growing. As per Mckinsey (Sanghvi & Westhoff, 2022), the venture capitalists (VCs) invested US$20.8 billion in the edtech industry globally in 2021, more than 40 times the amount they invested in 2010. On the other hand, the value of technology products related to language models is growing. For instance, the valuation of the parent company of ChatGPT has skyrocketed to US$29 billion in 2023 (Rosen, 2023). It is foreseeable that more related technology products will be launched in the market.

We believe this study can shed some light on future research and development in this tremendous topic and help the industry to overcome the fear of the unknown by increasing awareness and understanding of this topic.

ACKNOWLEDGEMENT

We extend our heartfelt appreciation to our colleagues, Fran Shollo and Luisa Goncalves Rosado from the University of Chester for their invaluable contributions to this paper. Their efforts in checking readability and providing dedicated translations greatly enhanced the quality of our work. We extend our sincere appreciation to Professor Eloy López Meneses, the Editor, and the entire editorial team for their dedicated time and effort in organizing the review process for our manuscript. We are also grateful to the anonymous reviewers for their constructive feedback, which further strengthened the paper.

REFERENCES

Ba, S., Hu, X., Stein, D. & Liu, Q. (2023). Assessing cognitive presence in online inquiry-based discussion through text classification and epistemic network analysis. *British Journal of Educational Technology*, 54, 247-266. https://doi.org/10.1111/bjet.13285

Baron, N. (2023). Even kids are worried ChatGPT will make them lazy plagiarists, says a linguist who studies tech's effect on reading, writing and thinking. *Fortune*. https://fortune.com/2023/01/19/what-is-chatgpt-ai-effect-cheating-plagiarism-laziness-education-kids-students/

Bengio, Y. & Senecal, J.S. (2008). Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, *19*(4), 713–722. https://doi.org/10.1109/TNN.2007.912312

Beseiso, M., Alzubi, O.A. & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, *33*(3), 727-746. https://doi.org/10.1007/s12528-021-09283-1

Botarleanu, R.M., Dascalu, M., Allen, L.K., Crossley, S.A. & McNamara, D.S. (2021). Automated Summary Scoring with ReaderBench. In A. Cristea & C. Troussas (Eds.), *Intelligent Tutoring Systems (ITS 2021)*, 321-332. Springer. https://doi.org/10.1007/978-3-030-80421-3_35

Condor, A. (2020). Exploring automatic short answer grading as a tool to assist in human rating. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millan (Eds.), *Artificial Intelligence in Education (AIED 2020)*. Springer. https://doi.org/10.1007/978-3-030-52240-7_14

Dempsey, J. (2023). AI: Arguing its Place in Higher Education. *Higher Education Digest*. https://www.highereducationdigest.com/ai-arguing-its-place-in-higher-education/

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv,1810.04805v2 https://doi.org/10.48550/ARXIV.1810.04805

Dimzon, F.D. & Pascual, R.M. (2020). An automatic phoneme recognizer for children's filipino read speech. *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Takamatsu, Japan, 2020,1-5. https://doi.org/10.1109/TALE48869.2020.9368399

Van-Dis, E.A.M., Bollen, J., Zuidema, W., Van-Rooij, R. and Bockting, C.L. (2023). ChatGPT: five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

Du, H., Xing, W. & Pei, B. (2021). Automatic text generation using deep learning: providing large-scale support for online learning communities. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2021.1993932

Dyulicheva, Y.Y. (2021). Learning Analytics in MOOCS as an Instrument for Measuring Math Anxiety. *Voprosy Obrazovaniya-Educational Studies Moscow*. https://doi.org/10.17323/1814-9545-2021-4-243-265

Esmaeilzadeh, S., Williams, B., Shamsi, D. & Vikingstad, O. (2022). Providing insights for open-response surveys via end-to-end context-aware clustering. In M. Rodrigo, N. Matsuda, A. Cristea, & V. Dimitrova (Eds), *Artificial Intelligence in Education (AIED 2022)*. Springer. https://doi.org/10.1007/978-3-031-11644-5_44

Fernandez, N., Ghosh, A., Liu, N., Wang, Z., Choffin, B., Baraniuk, R. & Lan, A. (2022). Automated scoring for reading comprehension via in-context BERT tuning. In M. Rodrigo, N. Matsuda, A. Cristea, & V. Dimitrova (Eds). *Artificial Intelligence in Education (AIED 2022)*. Springer. https://doi.org/10.1007/978-3-031-11644-5_69

Firoozi, T., Mohammadi, H. & Gierl, M.J. (2022). Using active learning methods to strategically select essays for automated scoring. *Educational Measurement Issues and Practice*, 00, 1-10. https://doi.org/10.1111/emip.12537

Fitzpatrick, D. (2023). Overcoming ChatGPT fear in 3 steps. *FE News*. https://www.fenews.co.uk/exclusive/overcoming-chatgpt-fear-in-3-steps/

Geller, S.A., Gal, K., Segal, A., Sripathi, K., Kim, H.G., Facciotti, M.T., Igo, M., et al. (2021). New methods for confusion detection in course forums: student, teacher, and machine. *IEEE Transactions on Learning Technologies*, *14*(5), 665-679. https://doi.org/10.1109/TLT.2021.3123266

Gift, T. & Norman, J. (2023). AI makes university honour codes more necessary than ever. *Times Higher Education (THE)*. https://www.timeshighereducation.com/blog/ai-makes-university-honour-codes-more-necessary-ever

Goel, V., Sahnan, D., Venktesh, V., Sharma, G., Dwivedi, D. & Mohania, M. (2022). K-12BERT: BERT for K-12 education. In M. Rodrigo, N. Matsuda, A. Cristea, & V. Dimitrova (Eds), *Artificial Intelligence in Education: Posters and Late Breaking Results. Workshops and Tutorials, Industry and Innovation Tracks, Practitioners and Doctoral Consortium (AIED 2022)*. Springer. https://doi.org/10.1007/978-3-031-11647-6_123

Goldberg, Y. & Levy, O. (2014). Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXi. https://doi.org/10.48550/ARXIV.1402.3722

Greenhouse, S. (2023). US experts warn AI likely to kill off jobs – and widen wealth inequality. *The Guardian*. https://www.theguardian.com/technology/2023/feb/08/ai-chatgpt-jobs-economy-inequality

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., et al. (2021). Pre-trained models: Past, present and future. *AI Open*, 2, 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

Hao, Y., Li, H., Ding, W., Wu, Z., Tang, J., Luckin, R. & Liu, Z. (2021). Multi-task learning based online dialogic instruction detection with pre-trained language models. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin & V. Dimitrova (Eds.), *Artificial Intelligence in Education (AIED 2021)*. Springer. https://doi.org/10.1007/978-3-030-78270-2_33

Hess, F. (2023). Will ChatGPT Be A Blow To Learning, Or A Boon? We'll Decide. *Forbes*. https://www.forbes.com/sites/frederickhess/2023/02/08/will-chatgpt-be-a-blow-to-learning-or-a-boon-well-decide/

Hsu, H.H. & Huang, N.F. (2022). Xiao-Shih: a self-enriched question answering bot with machine learning on Chinese-based MOOCs. *IEEE Transactions on Learning Technologies*, *15*(2), 223-237. https://doi.org/10.1109/TLT.2022.3162572

Ibanez, M., Reyes, L.L.A., Sapinit, R., Hussien, M.A. & Imperial, J.M. (2022). On applicability of neural language models for readability assessment in Filipino. In M. Rodrigo, N. Matsuda, A. Cristea, & V. Dimitrova (Eds). *Artificial Intelligence in Education: Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners and Doctoral Consortium (AIED 2022)*. Springer. https://doi.org/10.1007/978-3-031-11647-6_118

Jayaraman, J.D. & Black, J. (2022). Effectiveness of an Intelligent Question Answering System for Teaching Financial Literacy: A Pilot Study. In D. Guralnick, M. Auer & A. Poce (Eds.), *Innovations in Learning and Technology for the Workplace and Higher Education (TLIC 2021)*. Springer. https://doi.org/10.1007/978-3-030-90677-1_13

Khot, T., Clark, P., Guerquin, M., Jansen, P. & Sabharwal, A. (2020). QASC: A Dataset for Question Answering via sentence composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(5). https://doi.org/10.1609/aaai.v34i05.6319

Khushk, A., Zhiying, L., Yi, X. & Zengtian, Z. (2023). Technology Innovation in STEM Education: A Review and Analysis. *International Journal of Educational Research and Innovation*, 19, 29–51. https://doi.org/10.46661/ijeri.7883

Lee, J., Soleimani, F., Irish, I., Hosmer, J., Soylu, M.Y., Finkelberg, R. & Chatterjee, S. (2022). Predicting cognitive presence in at-scale online learning: MOOC and for-credit online course environments. *Online Learning*, *26*(1). https://doi.org/10.24059/olj.v26i1.3060

Lee, M.C., Chang, J.W. & Chen, J.L. (2014). Detecting ESL/EFL grammatical errors based on n-grams and web resources. *Conference name: 6th International Conference on Education and New Learning Technologies (EDULEARN14 Proceedings)*, 345-351.

Leong, K., Sung, A., Au, D., & Blanchard, C. (2020). A review of the trend of microlearning. *Journal of Work-Applied Management*, *13*(1), 88-102. https://doi.org/10.1108/JWAM-10-2020-0044

Leydesdorff, L. & Etzkowitz, H. (2003). Conference report: Can 'the public' be considered as a fourth helix in university-industry-government relations? Report on the Fourth Triple Helix Conference, 2002. *Science and Public Policy*, *30*(1), 55–61. https://doi.org/10.3152/147154303781780678

Li, H. (2022). Language models: past, present, and future. *Communications of the ACM*, *65*(7), 56–63. https://doi.org/10.1145/3490443

Li, Y., Anastasopoulos, A. and Black, A.W. (2020). Towards Minimal Supervision BERT-Based Grammar Error Correction. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-20)*, *34*(10), 13859-13860. https://doi.org/10.1609/aaai.v34i10.7202

Lin, J. (2020). Hybrid translation and language model for micro learning material recommendation. *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT 2020)*, 384-386. https://doi.org/10.1109/ICALT49669.2020.00121

Liu, Q., Liu, T., Zhao, J., Fang, Q., Ding, W., Wu, Z., Xia, F., et al. (2021). Solving ESL sentence completion questions via pre-trained neural language models. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin & V. Dimitrova (Eds.), *Artificial Intelligence in Education (AIED 2021)*. Springer. https://doi.org/10.1007/978-3-030-78270-2_46

López-Belmonte, J., Segura-Robles, A., Cho, W. C., Parra-González, M.E. & Moreno-Guerrero, A. J. (2021). What does literature teach about digital pathology? A bibliometric study in Web of Science. *International Journal of Educational Research and Innovation*, (16), 106–121. https://doi.org/10.46661/ijeri.4918

Lopez-Ferrero, C., Renau, I., Nazar, R. & Torner, S. (2014). Computer-assisted revision in Spanish academic texts: Peer-assessment. *Procedia - Social and Behavioral Sciences*,141, 470-483. https://doi.org/10.1016/j.sbspro.2014.05.083

Lu, X., Sahay, S., Yu, Z. & Nachman, L. (2021). ACAT-G: An Interactive Learning Framework for Assisted Response Generation. *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, *35*(18), 16084-16086. https://doi.org/10.1609/aaai.v35i18.18019

Makhlouf, J. & Mine, T. (2021). Mining students' comments to build an automated feedback system. Proceedings of the 13th International Conference on Computer Supported Education (CSEDU),1. SciTePress. https://doi.org/10.5220/0010372200150025

Masala, M., Ruseti, S., Dascalu, M. & Dobre, C. (2021). Extracting and clustering main ideas from student feedback using language models. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin & V. Dimitrova (Eds.), *Artificial Intelligence in Education (AIED 2021)*. Springer. https://doi.org/10.1007/978-3-030-78292-4_23

Meisner, C. (2023). Baylor professors fear students will lose critical thinking skills with ChatGPT. *Baylot Lariat*. https://baylorlariat.com/2023/02/07/baylor-professors-fear-students-will-lose-critical-thinking-skills-with-chatgpt/

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 I. Hilliger, P. Munoz-Merino, T. DeLaet, A. Ortega-Arranz, & T. Farrell (Eds.), *Educating for a New Future*: *Making Sense of Technology-Enhanced Learning Adoption*. EC-TEL 2022. Lecture Notes in Computer Science, 13450. Springer. https://doi.org/10.1007/978-3-031-16290-9_18

Murray, B. (2023). ChatGPT forces us to rethink student effort and laziness. *Psychology Today*. https://www.psychologytoday.com/intl/blog/real-happiness-in-a-digital-world/202301/chatgpt-forces-us-to-rethink-student-effort-and

Ndukwe, I.G., Amadi, C.E., Nkomo, L.M. & Daniel, B.K. (2020). Automatic grading system using sentence-BERT network. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millan (Eds.), *Artificial Intelligence in Education (AIED 2020)*. Springer. https://doi.org/10.1007/978-3-030-52240-7_41

Nehyba, J. & Stefanik, M. (2022). Applications of deep language models for reflective writings. *Education and Information Technologies*, 28, 2961-2999. https://doi.org/10.1007/s10639-022-11254-7

Nicula, B., Dascalu, M., Newton, N., Orcutt, E. & McNamara, D.S. (2021). Automated paraphrase quality assessment using recurrent neural networks and language models. In A. Cristea & C. Troussas (Eds.), *Intelligent Tutoring Systems (ITS 2021)*. Springer. https://doi.org/10.1007/978-3-030-80421-3_36

Ondas, S., Hladek, D., Stas, J., Juhar, J., Kovacs, L. & Baksane, E.V. (2015). Semantic roles modeling using statistical language models. 2015 13th International Conference on Emerging Elearning Technologies and Applications (Iceta). IEEE. https://doi.org/10.1109/ICETA.2015.7558502

Pan, L. (2018). Automatic generation of children's songs based on machine statistic learning. *International Journal of Emerging Technologies in Learning*, *12*(3), 17-31. https://doi.org/10.3991/ijet.v13i03.8367

Parasa, N.S., Diwan, C. & Srinivasa, S. (2022). Automatic riddle generation for learning resources. In M., Rodrigo, N., Matsuda, A., Cristea, & V., Dimitrova (Eds). *Artificial Intelligence in Education*: *Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners and Doctoral Consortium (AIED 2022)*. Springer. https://doi.org/10.1007/978-3-031-11647-6_66

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., et al. (2013). Academic engagement and commercialisation: A review of the literature on university–industry relations. *Research Policy, 42*(2), 423-442. https://doi.org/10.1016/j.respol.2012.09.007

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Q.ai (2023). Here Comes the Bing Chatbot - Microsoft's ChatGPT For Search Has Arrived, Forcing Google's Hand. *Forbes*. https://www.forbes.com/sites/qai/2023/02/09/here-comes-the-bing-chatbotmicrosofts-chatgpt-for-search-has-arrived-forcing-googles-hand/?sh=6315ec6110fb

Rakovic, M., Sha, L., Nagtzaam, G., Young, N., Stratmann, P., Gasevic, D. & Chen, G. (2022). Towards the automated evaluation of legal casenote essays. In M., Rodrigo, N., Matsuda, A., Cristea, & V., Dimitrova (Eds), *Artificial Intelligence in Education (AIED 2022)*. Springer. https://doi.org/10.1007/978-3-031-11644-5_14

Rosen, P. (2023). ChatGPT's creator OpenAI has doubled in value since 2021 as the language bot goes viral and Microsoft pours in $10 billion. *Markets Insider*. https://markets.businessinsider.com/news/

stocks/chatgpt-openai-valuation-bot-microsoft-language-google-tech-stock-funding-2023-1#:~:text=OpenAI%2C%20the%20parent%20company%20of

Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, *88*(8), 1270–1278. https://doi.org/10.1109/5.880083

Salim, S. (2023). UAE jobs and ChatGPT: Over 70 % workers must learn new skills by 2025, says expert. *Khaleej Times*. https://www.khaleejtimes.com/jobs/uae-jobs-should-employees-worry-about-chatgpt-other-ai-tools-replacing-them

Sanghvi, S. & Westhoff, M. (2022). Education technology: Five trends to watch in the EdTech industry. *Mckinsey & Company*. https://www.mckinsey.com/industries/education/our-insights/five-trends-to-watch-in-the-edtech-industry

Sung, C., Dhamecha, T.I. & Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. In S. Isotani, E. Millan, A. Ogan, P. Hastings, B. McLaren, and R. Luckin (Eds.), *Artificial Intelligence in Education (AIED 2019)*. Springer. https://doi.org/10.1007/978-3-030-23204-7_39

Tang, L., Ke, E., Singh, N., Feng, B., Austin, D., Verma, N. & Drori, I. (2022). Solving probability and statistics problems by probabilistic program synthesis at human level and predicting solvability. In M., Rodrigo, N., Matsuda, A., Cristea, & V., Dimitrova (Eds.). *Artificial Intelligence in Education*: *Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners and Doctoral Consortium (AIED 2022)*. Spinger. https://doi.org/10.1007/978-3-031-11647-6_127

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., et al. (2017). Attention is all you need. *arXiv*. https://doi.org/10.48550/ARXIV.1706.03762

Wang, Z., Valdez, J., Mallick, D.B. & Baraniuk, R.G. (2022). Towards human-like educational question generation with large language models. In M., Rodrigo, N., Matsuda, A., Cristea, & V., Dimitrova (Eds). *Artificial Intelligence in Education (AIED 2022)*. Springer. https://doi.org/10.1007/978-3-031-11644-5_13

Westfall, C. (2023). Educators Battle Plagiarism As 89 % Of Students Admit To Using OpenAI's ChatGPT For Homework. *Forbes*. https://www.forbes.com/sites/chriswestfall/2023/01/28/educators-battle-plagiarism-as-89-of-students-admit-to-using-open-ais-chatgpt-for-homework/

Wise, A.F., Cui, Y. & Jin, W.Q. (2017). Honing in on social learning networks in MOOC forums: examining critical network definition decisions. Proceedings of the International Learning Analytics & Knowledge Conference (Lak'17), 383-392. https://doi.org/10.1145/3027385.3027446

Wulff, P., Buschhueter, D., Westphal, A., Mientus, L., Nowak, A. & Borowski, A. (2022). Bridging the gap between qualitative and quantitative assessment in science education research with machine learning - a case for pretrained language models-based clustering. *Journal of Science Education and Technology*, 31, 490-513. https://doi.org/10.1007/s10956-022-09969-w

Xiao, C., Shi, L., Cristea, A., Li, Z. & Pan, Z. (2022). Fine-grained Main Ideas Extraction and Clustering of Online Course Reviews. In M. Rodrigo, N. Matsuda, A. Cristea, & V. Dimitrova (Eds). *Artificial Intelligence in Education (AIED 2022)*. Springer. https://doi.org/10.1007/978-3-031-11644-5_24

Xu, S., Ding, W. & Liu, Z. (2020). Automatic Dialogic Instruction Detection for K-12 Online One-on-One Classes. In I. Bittencourt, M. Cukurova, K. MulZs://doi.org/10.1007/978-3-030-78292-4_36

Yang, G., Wen, D., Kinshuk, Chen, N.S. & Sutinen, E. (2012). Personalized Text Content Summarizer for Mobile Learning: An Automatic Text Summarization System with Relevance Based Language Model.2012 IEEE Fourth International Conference on Technology for Education. https://doi.org/10.1109/T4E.2012.23

Zhu, X., Wu, H. & Zhang, L. (2022). Automatic short-answer grading via BERT-based deep neural networks. *IEEE Transactions on Learning Technologies*, *15*(3), 364-375. https://doi.org/10.1109/TLT.2022.3175537