OmniaScience

Journal of Technology and Science Education

JOTSE, 2024 – 14(2): 633-647 – Online ISSN: 2013-6374 – Print ISSN: 2014-5349

https://doi.org/10.3926/jotse.2111

QUICK REVIEW OF PEDAGOGICAL EXPERIENCES USING GPT-3 IN EDUCATION

Joel Manuel Prieto-Andreu[1*], Antonio Labisa-Palmeira[2]

[1]Universidad Internacional de La Rioja (Spain)

[2]Universidade Lusófona de Humanidades e Tecnologias (Portugal)

Corresponding author: jmprietoandreu@gmail.com

p126@ulusofona.pt

Received February 2023

Accepted September 2023

Abstract

GPT-3 is a neuronal language model that performs tasks such as classification, question-answering and text summarization. Although chatbots like BlenderBot-3 work well in a conversational sense, and GPT-3 can assist experts in evaluating questions, they are quantifiably worse than real teachers in several pedagogical dimensions. We present the first systematic literature review that analyzes the main contributions and uses of GPT-3 in the field of education. The protocols suggested in the PRISMA 2020 statement were followed for the drafting of the review. According to the results, 34 significant productions were identified through a systematic search in ISI Web of Science, SCOPUS and Google Scholar. GPT-3 has been considered in the academic, ethical and medical fields, in humanities and in computer science, in the formulation of questions and answers, and through cooperative educational dialogs. GPT-3 has been proven to have valuable applications in education, such as the automation of routine tasks, in making quick diagnoses of the students' weaknesses and in the automatic generation of questions, but it still faces challenges and limitations that require additional investigation. We discuss the educational possibilities and the limitations to the use of GPT-3.

Keywords – IA, GPT-3, ChatGPT, Artificial intelligence.

To cite this article:

Prieto-Andreu, J.M., & Labisa-Palmeira, A. (2024). Quick review of pedagogical experiences using GPT-3 in education. *Journal of Technology and Science Education,* 14(2), 633-647. https://doi.org/10.3926/jotse.2111

---

1. Introduction

In 2020, OpenAI developed Generative Pre-trained Transformer (GPT-3), a neuronal language model with sophisticated natural language generation and task completion, such as classification, question-answering and summarization. According to Nath, Marie, Ellershaw, Korot and Keane (2022), natural language processing (NLP) is a subfield of artificial intelligence focused on the interaction of human language with computer systems. NLP has recently been discussed in the mass media and in the literature with the advent of GPT-3, a language model capable of producing text similar to that of a human. ChatGPT-3 is part of Artificial Intelligence (IA) itself, but it can also generate natural human conversations.

Figure 1 shows the search in Google Trends for "GPT-3" and "ChatGPT" from early 2020, when the GPT architecture was designed, until now.

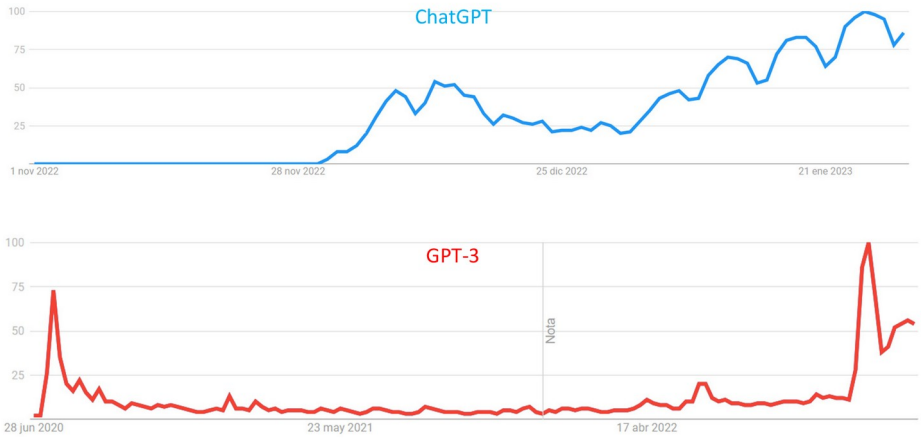

Figure 1. Google Trends for "ChatGPT" and "GPT-3"

Figure 1 shows that searches on GPT-3 interested Internet surfers when this architecture was created in mid-2020 in a first wave, and interest was aroused once again in late 2022 in a second wave. On the other hand, searches on ChatGPT began together with the second wave of the GPT-3 searches in late 2022.

This AI is such a powerful tool that its creators constantly state that users have not yet discovered all of its full functions and that its path is promising and difficult to explore (Zapata & Hoyos, 2023). The advantages are obvious; it saves time by generating different types of multilingual content precisely and with multiple variations. It could become a gift to students who cheat, a powerful teaching assistant or a tool to boost creativity. Even though conversational agents (chatbots) like Blender work well in conversational acquisition, they are quantifiably worse than real teachers in several pedagogical dimensions, especially with regard to assistance (Tack & Piech, 2022). This is a difficult problem because there is no standard or universal solution to measure the capacity and efficacy of teachers (Goe, Bell & Little, 2008).

According to MacNeil, Tran, Mogil, Bernstein, Ross and Huang (2022), Large Language Models (LLM) are changing the way in which students interact with code. For example, Github's Copilot can generate code for programmers, which leads researchers to ponder concerns about possible cheating. According to Villarroel (2021), Copilot constitutes a surprising service capable of assisting developers, writing portions of code and making suggestions that help them in their daily activities. It is an assistant to streamline routine tasks and to propose predictive contexts. The system is based on Codex, a new system of artificial intelligence created by OpenAI, which will undoubtedly provide significant support in education. According to Gaikwad, Rambhia and Pawar (2022), the increase in human interactions with AI has given rise to the demand for interactive interfaces, such as chatbots, text translators, text predictors and text generators that use pre-trained language models to carry out their own specific tasks. According to Sarsa, Denny, Hellas and Leinonen (2022), the performance of the deep learning models is based on both a large number of parameters (175 billion, in the case of GPT-3) and an extensive corpus of text used for training (570 GB of text for GPT-3). Codex, also from OpenAI, is a model of GPT similar to GPT-3, but it has been perfected using code that is publicly available on GitHub, with the aim of translating natural language into source code. Besides Codex, other models of generative automatic learning have been developed that are capable of generating natural language from source code and/or vice versa. One of the first models of this type is CodeBERT from Microsoft or AlphaCode from DeepMind.

GPT-3 can help experts in the evaluation of their questions. Moore, Nguyen, Bier, Domadia and Stamper (2022) demonstrate one way to help escalate on-line learning and improve educational resources, opening

up more opportunities to involve students in order to aid in the evaluation process of question generation, thus taking advantage of human and linguistic models. This language model has recent applications for classifying email messages (Thiergart, Huber & Übellacker, 2021), determining whether news items were real or fake (Chan, 2022), generating and evaluating questions based on text-based learning materials (Bhat, Nguyen, Moore, Stamper, Sakr & Nyberg, 2022; Moore et al., 2022), evaluating text-based cooperative discourse (Phillips, Saleh, Glazewski, Hmelo-Silver, Mott & Lester, 2022), generating summaries of scientific articles (Alvarez-Carmona, Aranda, Diaz-Pacheco & Ceballos-Mejıa, 2022), determining the most appropriate thesis supervisor in order to lead a project (Agustin, 2022), helping parents and children to re-write stories together (Lee, Kim, Chang & Kim, 2022), calculating students' grades through a similarity index of their responses (Ramnarain-Seetohul, Bassoo & Rosunally, 2022), learning new languages (Franganillo, 2022), solving math problems (Zong & Krishnamachari, 2022), analyzing the emotions of children in interviews (Lammerse, Hassan, Sabet, Riegler & Halvorsen, 2022), creating content on the social media (Aljanabi, Ghazi, Ali & Abed, 2023), and in particular, in its potential for automating learning (Alawi, 2023).

The future is rapidly approaching, and with the recent launch of the open code AI software ChatGPT3, some could argue that it is already here in education. According to Aydin and Erdem (2022), in the world of digitalization, AI paves the way for the automation of routine work performed by humans and generally facilitates life. How do educators use GPT-3? Few researchers have tested the usability of these tools in education, and it is the intention of this work to conduct the first literature review that analyzes the main contributions made by GPT-3 in the field of education.

2. Methodology

The review methodology used is an adaptation of Cochrane's quick review methodology (Garritty, Stevens, Gartlehner, King & Kamel, 2016). According to Hamel, Michaud, Thuku, Skidmore, Stevens, Nussbaumer-Streit et al. (2021), a quick review is one way of synthesizing knowledge that accelerates the process of conducting a traditional systematic review by streamlining or omitting a series of methods. A quick review must be rigorous and must ask a very focused question. In the case of this review, it was: How do educators use GPT-3? To select the studies, the review has been conducted by following the protocols set out in the PRISMA 2020 (Page, McKenzie, Bossuyt, Boutron, Hoffmann, Mulrow et al., 2021) statement. The following inclusion criteria were taken into account: studies written in Spanish and in English; peer-reviewed manuscripts; works with full text access. On the other hand, the exclusion criteria dictated that articles from unreliable sources were to be rejected. The sample of the systematic review consists of studies that have investigated the use of GPT-3 in an educational setting. These studies were located in the most important on-line computerized databases in the areas of Health and Social sciences: ISI Web Of Science, SCOPUS, and Google Scholar, in order to review the grey literature.

An ad-hoc quantitative table was created, organized by year, filtering the results for the last 3 years since 2020, as this was the time when this architecture began to be used (said time fluctuation can be seen in Figure 2), and assigning a record number to the studies that allowed them to be selected later in the classification phase (Table 1).

The search expression [GPT-3 OR ChatGPT AND education] was used. The search in the title, abstract and keywords was filtered in ISI Web Of Science (TOPIC) for the years 2020, 2021, 2022 and 2023. Altogether, in the first identification phase, a total of 3,729 records were identified since the first record in 2020 until January 2023 (145 in ISI, 144 in SCOPUS and 3,440 in Google Scholar). The records were distributed following the random selection in a stratified manner, with a 95% level of confidence and a standard error of +/- 4.3 in the 3 databases since 2020. In the second classification phase, the search filtered the articles by thematic category. Table 1 shows the details of the searches in ISI Web of Science and SCOPUS, obtaining 116 results for the search expression used in ISI Web of Science for the category "Education Educational Research", 143 results in SCOPUS for the categories Computer Science, Social Sciences, Engineering and Mathematics, which were the categories where the most results were found, and 3,410 in Google Scholar, not including citations, for a total of 3,669 results.

Year	ISI Web of Science	SCOPUS	SCHOLAR
2023	0	8	271
2022	65	93	1914
2021	46	36	1015
2020	5	6	210

Table 1. Records in ISI, SCOPUS and Google Scholar in the classification phase

In the eligibility phase, after reading the title, 3,435 studies were ruled out, resulting in 12 results in ISI Web of Science, 37 results in SCOPUS, and 185 results in Google Scholar, for a total of 234 studies.

Finally, in the inclusion phase of articles for the review, 34 works were selected (33% proceedings), excluding the rest after having read the abstract and the full text, as they did not meet the inclusion criteria or did not follow the thematic line of the systematic review. Figure 2 shows a flow chart of the article selection process.

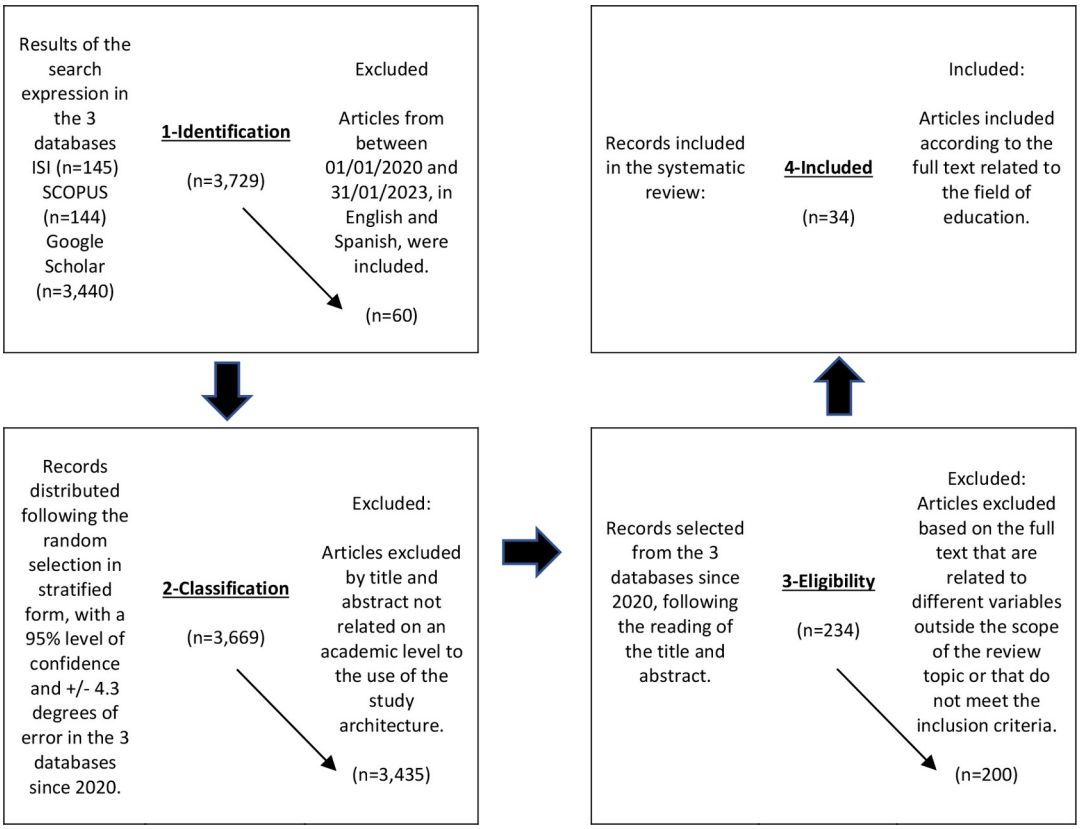

Figure 2. Flow chart of the article selection process

3. Results

Of the 34 articles analyzed, 7 are focused on the field of academics (intellectual property, article abstracts and titles, academic writing and selection of thesis supervisor), 6 consider instructional proposals based on formulating questions and answers in order to encourage curiosity and autonomy by the students, 6 have to do with the humanities from a creative perspective (representations of real people, roles of characters in narrative texts and re-writing stories), 5 are focused on mathematics, computer science and programming (creation of code, animations and creative graphic expressions and mathematical problem solving), 4 deal with ethical matters (the political situation surrounding the use of AI, biases and social consequences and dangers derived from improper use), 4 are focused on the medical field (digital clinical simulations, instructional clinical resources, medical support and diagnostic assistance), and 3 are studies

related to cooperative educational dialogs. Table 2 provides details of the contribution of each of the 34 studies analyzed, indicating the objective and the main contributions:

Title	Aim	Main contributions
Academic Field		
Plagiarism in the Age of Massive Generative Pre-trained Transformers (GPT-3).	To instill a sense of urgency, as well as a sense of the current delay by the academic community in considering the potential use of GPT-3 to facilitate plagiarism, undermining intellectual property.	GPT-3 was used through AI Dungeon to generate 3 types of text contents (academic essays, oral presentations and opinion articles). Each text sample was sent to a plagiarism detection service (https://plagiarismdetector.net) and it was determined to be an original. NLP technology is used to prevent the publication of false, plagiarized or fraudulent findings. If the definition itself of these concepts changes, it would also be necessary to reconsider the objective of the peer review and the possible role of AI in scientific writing.
News summarization and evaluation in the era of GPT-3.	To compare the results of 3 representative models for creating academic abstracts with GPT-3, and gathering the human preferences in terms of quality.	The abstracts generated by adjusted models emulate standard abstracts in their training data sets. However, most of the research on abstracts encountered a lack of consensus about which parameters or dimensions should be evaluated, task design and other factors. In contrast, the GPT3-D2 models based on indications generate abstracts considering how the task description influences the behavior learned during the previous training or the adaptation of instructions. This leads to a preference for GPT-3 abstracts by humans and prevents common problems specifically related to data sets, such as being outdated.
Ghost in the machine or monkey with a typewriter–generating titles for Christmas research articles in The BMJ using artificial intelligence: observational study.	To check the use of GPT-3 to generate titles for research articles, and to determine the attractiveness of the titles of these articles for potential readers.	In the context of extravagant titles, such as those that appear in the Christmas editions of The BMJ, GPT-3 has the potential to generate plausible results that are attractive and could attract potential readers. However, it is only possible to capture the interest of the reader with the guidance of an expert, since some of the titles of the articles in this study were irrelevant or even offensive.
Implications of artificial intelligence in virtual classrooms for higher education.	To discuss the influence of AI in virtual classrooms and the synergies between artificial intelligence and the gamification of learning.	In an academic setting, GPT-3 is capable of writing any article, be it journalistic, academic, etc., and it is also capable of learning other skills for which it has never been trained, for example, translating text from one language to another. On the other hand, gamification and artificial intelligence will be strategies with multiple benefits, and thus they will be implemented in virtual learning environments, increasing motivation and commitment and reducing student drop out rates.
ChatGPT: Open Possibilities.	To comment on the uses that ChatGPT-3 might have in academic writing, such as a search engine, in coding tasks, to detect security vulnerabilities or to create content on social networks.	ChatGPT-3 can be used to power chatbots, virtual assistants and other conversational functions. On an academic level, it can generate article abstracts, extract key points and even provide citations. This can save researchers a significant amount of time and effort, allowing them to concentrate on other tasks. It can generate text for various types of academic documents, provide feedback on grammar, style and coherence, and help students understand and summarize difficult texts. It can also be used as a search engine, receiving precise, relevant information in return. On tasks that involve coding, ChatGPT-3 provides information that is directly related to the code snippet or command that is being used. It is useful to detect security vulnerabilities. It is capable of understanding the intention behind a query. It can also create content on social networks in a way that is intuitive and easy to use.

Title	Aim	Main contributions
Automatic generator of scientific abstracts in tourism research.	To analyze the use of GPT-3 and other similar architectures for the analysis and generation of abstracts in 227 scientific articles on natural language processing applied to the tourism sector.	Based on all the data collected and the use of different architectures on learning models, it is indicated that the best result is the 0.21 obtained by GPT-3, according to the Jaccard coefficient. The Jaccard coefficient is a measure of similarity between sets, which can be used to compare the similarity between two sets of words, phrases or texts. In this case, the value 0.21 indicates that a certain measure of similarity has been reached by the GPT-3 model, as compared to a set of reference data or a set of desired responses. The closer to 1 a Jaccard coefficient is, the greater the similarity will be between the sets.
Implementation of Generative Pre-Trained Transformer 3 Classify-Text in Determining Thesis Supervisor.	To verify the potential of GPT-3 to select the most appropriate thesis supervisor based on 823 titles presented by students.	The resulting accuracy of the processing of the data set is capable of producing a precision value of 98% by implementing the GPT-3 Classify-Text algorithm in the recommendation of thesis supervisors. It is therefore adequate for implementation in a system supporting the decision on thesis supervisors.
Questions and Answers		
GPT-3-driven pedagogical agents for training children's curious question-asking skills.	To observe the ability of 75 students between 9 and 10 years of age to ask curious, relevant and profound questions, using GPT-3 to give them specific hints to spark their curiosity and autonomy while learning.	Two types of clues were presented: "closed", which guided them to specific questions and "open", which allowed them to explore different questions, in both cases, as suggested by GPT-3. The results showed a similar performance in the formulation of questions among the children, although this was significantly better for the participants that interacted with the clues provided by GPT-3 in the "open" mode, which boosted their curiosity, divergence and autonomy of expression to a greater extent. This method is still limited, since it is based on the manual generation of said indicators for each educational resource, which can be a very long and costly process.
Towards Generalized Methods for Automatic Question Generation in Educational Domains.	To create a model to generate and evaluate questions stemming from text-based learning materials in an introductory course on data science, and to rate the questions generated according to their relevance for the key concepts extracted, according to 3 methods.	The generated questions were rated favorably by the three evaluation methods (information score, automated GPT-3 classification and human judgment), for capturing important concepts or being instructionally solid. GPT-3 was able to replicate 66.50% of the consensus from the two expert evaluators, which is far above random level. The model seemed to learn that the longer questions were probably solid, which is a reasonable supposition, since these questions can contain more relevant contextual information.
Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3.	To evaluate the short answer questions generated by 143 university level on-line chemistry students through the classification made by GPT-3.	Generating short answer questions is a popular way of learning with benefits for both higher order thinking by the students and for the instructors, by collecting evaluation items. GPT-3 was used for the binary classification in order to see if it was possible to classify the questions generated by the students as good or poor quality. 32% of the questions generated by students were evaluated by experts as being of high quality, identifying that 23% of the questions generated by the students evaluated higher cognitive processes according to the Bloom Taxonomy. It was concluded that 91% of the short-answer questions generated by the students were classified by GPT-3 as being comprehensible. The classification kept the students more involved in the learning process, by allowing them to create questions in a more natural context as they worked through the GPT-3 evaluations.

Title	Aim	Main contributions
Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI.	To construct a synthetic data set in German, using GPT-3 based on human instructions.	The questions were consciously selected to be used as instructions, in such a way that the labels on the naturalistic data set were applied to the synthetic results of GPT-3. It was concluded that GPT-3 trained with synthetic data works much better in situations where there are very few data items, and this is preferable than the collection and annotation of naturalistic data. The reasons include a difference in the conversation style between the user and the synthetic data and a lack of content and semantic variability in the samples that were generated.
An Extensive Analysis Between Different Language Models: GPT-3, BERT and MACAW.	To experiment with the GPT-3, BERT and Macaw language models with different categorical questions in order to understand their architecture and behavior under different circumstances.	GPT-3 is pre-trained on a robust data set and provides very elaborate responses similar to those of humans, while the results produced by BERT can be customized, providing a personalized context; on the other hand, Macaw shows greater precision when responding to general questions.
Towards Automated Generation and Evaluation of Questions in Educational Domains.	To evaluate the relevance of the key concepts reflected in the development of questions asked by humans and by GPT-3, based on learning materials stemming from a text in an introductory course in data science.	The questions were classified as useful or not useful for learning, with two different focuses: automatic labeling by GPT-3, and manual labeling by expert human judges. The results showed that the questions generated were favorably rated by all the evaluation methods.
Humanities		
Breakfast with Confucius, Dinner with Lem. Linguistic Avatars of GPT-3.	To use GPT-3 to investigate the conversational aspects of the ability of linguistic AI to construct believable representations of real people.	The GPT-3 conversational simulations seem to have a form of intentionality (the desire to be understood). They discuss the social, philosophical and ethical implications of this new generation of conversational simulations from eastern mythology to the post-human visions in contemporary critical discourse. The focus on the possible impact of the mimetic representations of AI on the existing notions of identity, bringing AI closer to the Humanities, establishing conversations with Plato, Queen Min and Princess Diana.
Heroes, Villains, and Victims, and GPT-3: Automated Extraction of Character Roles Without Training Data.	To demonstrate the usefulness of GPT-3 to extract character roles in narrative texts, with the ability to identify the hero, villain and victim in different media types: newspaper articles, movie plot summaries and political speeches.	Introducing the problem to the students as an automatic reading comprehension task, they provided an input document and asked GPT-3 who is the hero (or the villain or the victim), successfully extracting character roles from unformatted text documents. The results show that GPT-3 is more effective than other methods in labeling these character roles in a diverse set of narrative domains.
Black Box Karl Marx: What do large language models have to say about Das Kapital? A Comparison of GPT-2 and GPT-3 Outputs.	To compare the results of GPT-2 and GPT-3, using "Das Kapital" by Karl Marx as a text corpus.	There is a real potential for creativity by using GPT-3 for theoretical or philosophical work. GPT-3 can offer an impartial historical synopsis of Karl Marx and his theories, with an enormous potential for developing lesson plans for students to filter our errors and misinformation.

Title	Aim	Main contributions
Interaction and text: Resource cross-over in electronic art.	To comment on the artistic synergies in the text creation process through GPT-3, and to discuss the human or non-human nature of the creative process.	Artificial intelligence is opening doors of communication to the future that give us a glimpse at the work of future human and non-human artists. There are numerous works of electronic art that incorporate text as a base of the production and creative process, as well as works in which text forms part of the final result. In both cases, the text in its written version is one of the fundamental aspects of the creative process, and also the core part of the communication processes, in this case, between the artist, their work and the audience. GPT-3 is so powerful and sophisticated that it can emulate the information collection and processing process of the brain in a way that is so similar that the moral and ethical implications of its use are currently being studied.
Interactive Children's Story Rewriting Through Parent-Children Interaction.	To design an interface based on GPT-3 to support children and parents in order to re-write stories, together with the help of AI techniques.	The interface consists of three main components: the original story component, the questions and answers component and the re-written story components. Through the original story component, the parent can see the original story, as well as the phrases that can be changed while the story is read. The questions and answers component presents a set of suggested questions generated by AI that the parent can ask their child, referring to the preferences, feelings and/or daily life of the child. Once the answer has been submitted, the parent can then see how the story has been re-written. The interface presents various biases: gender, race and culture, which could trigger negative results, such as reinforcing gender stereotypes or building limited comprehension of normative behavior.
Programming		
Generating diverse code explanations using the GPT-3 large language model.	To analyze the different explanations in natural language that GPT-3 can automatically generate for a given code snippet.	GPT-3 can automatically create a checklist of common errors that students can make with regard to a certain code snippet. It is possible to trace the code execution, correct errors and explain how they were solved, generate analogies with real-world configurations and name relevant programming concepts.
Automatic generation of programming exercises and code explanations using large language models.	To create programming exercises using Codex (which include sample solutions and test cases) and code explanations, evaluating them qualitatively and quantitatively in order to determine if they are sensible, novel and easy to apply.	There is a significant value in the mass generative automatic learning models as a tool for instructors, although there is still the need for certain supervision in order to ensure the quality of the content that is generated before it is handed over to the students. Most of the programming exercises created by Codex were sensible, novel and included a solution with an appropriate sample, even though the exercises created could be easily influenced, since many exercises lacked evidence or had defective evidence. The explanations created by Codex cover most (90%) of the code, although they are inaccurate in some instances (67.2% of the explanation lines were correct).
Cracking the code: Co-coding with AI in creative programming education.	To demonstrate that Codex can coexist and be used in a higher education programming course for designers in order to create animations and creative graphic expressions.	Codex can be asked in natural language to produce programming code that performs specific functions, such as simple cases of drawing a circle, as well as more advanced programming expressions. There was a certain problem with the syntax and coding errors in relation to the Codex tool, in many cases it also provides a different outcome if it is run twice with the same instruction. This type of inconsistencies and unpredictability was experienced as frustrating and confusing by the participants, but it also led them to explore creative ways of interacting with the system, where they would repeat the instructions until they obtained a satisfactory result.

Title	Aim	Main contributions
The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming.	To explore how Codex works on typical introductory programming problems, generating code solutions as an output, explaining in English the input code and translating the code between programming languages.	They compared the answers from Codex to real questions taken from programming exams with the students results, demonstrating that Codex performed better than most students. On the other hand, they explored how much variation there is in the solutions generated by Codex, and they observed that an identical input indicator often leads to very different solutions in terms of the algorithmic focus and code length.
A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level.	To demonstrate that a previously text-trained neuronal network (OpenAI Codex), with code adjustment, solves math problems, explains solutions and generates questions on a human level.	Codex automatically solves questions from university level math courses on a human level (single and multivariate calculation, differential equations, introduction to probability and statistics, linear algebra and math for computer science) and explains and generates questions for university-level math courses on a scale with an automatic precision of 81%, a benchmark for higher education. GPT-3 automatically solves only 18.8% of these university questions through learning without any attempts, and 30.8% through learning with only a few attempts.
Ethical Questions		
The radicalization risks of GPT-3 and advanced neural language models.	To analyze the current political situation regarding the use of AI, making the suggestion to invest as soon as possible in the development of social guidelines, public policies and educational initiatives to avoid the affluence of misinformation and propaganda generated by AI.	GPT-3 shows strengths in the generation of text that precisely emulates interactive, informative and influential content that could be used to radicalize people towards far right-wing ideologies and violent behaviors. Researchers and service providers might consider contributing and investing in educational programs on the use of AI targeting mass audiences.
Gender and representation bias in GPT-3 generated stories.	To propose solutions to prevent undesirable social biases when GPT-3 is used for storytelling.	The language models generate different occupations and levels of respect for different genders, races and sexual orientations. GPT-3 associates women with verbs related to the family and appearance, and they are described as being less powerful than male characters.
Playing Games with AIs: The Limits of GPT-3 and Similar Large Language Models.	To evaluate how well GPT-3 completes the Turing test, determining its limits, the tendency to generate falsehoods and their social consequences.	GPT-3 completely lacks any semantic capacity, although it can competently participate in various semantic tasks. It has three limitations: regularity, prioritization and modal limitations. The conclusion is that it will not be able to show a constant fidelity to the real world, and while GPT-3 is very good at generating plausible text, it is a poor narrator of the truth.
Content generated by artificial intelligence: opportunities and threats.	To comment on the potential benefits and the main hazards derived from the incorrect use of AI in the contents industry.	GPT-3 not only produces texts, it is also capable of summarizing them and translating them, based on the context study, which has a large number of practical applications in the field of marketing, literature and journalism. Other web applications such as Anyword, Copy.ai, Copymatic, Copysmith, Jasper, Peppertype, StoryLab.ai, WordAI, Wordtune and Writesonic produce quick suggestions for headlines, commercial slogans, tweets, blog articles and product descriptions, among other texts. There are questions about how AI could shake up the contents industry. The organizations involved in the development of AI applications must act transparently, ensuring the ethical, responsible use of AI, being very vigilant in order to mitigate any possible negative effects of its use.

Title	Aim	Main contributions
Medical Issues		
Comparing Few-Shot Learning with GPT-3 to Traditional Machine Learning Approaches for Classifying Teacher Simulation Responses.	To experiment with GPT-3 and traditional automatic learning models (ML) to examine whether it could be used with open text answers on digital clinical simulations (DCS) to provide opportunities for improvisation in low-risk classroom settings.	GPT-3 performed substantially worse than the traditional models of automatic learning (ML). However, the performance of GPT-3 only marginally decreased as compared to traditional ML models with a 20-item training set (−0.06). Traditional ML models generally worked well, and in some cases showed a performance similar to that of the human base line.
New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology.	To verify the usefulness of GPT-3 in specific applications in ophthalmology to improve the use of valuable ophthalmic clinical resources, improve the flow of the clinic and bring the development of AI solutions to the end users.	This work represents a particularly challenging use of autoregressive language models and introduces a highly complex and multifaceted clinical problem. The careful implementation of GPT-3 and the combination of text image models, such as DALL-E 2, have the potential to transform patient care and revolutionize modern ophthalmology.
Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery.	To determine the potential of GPT-3 as a clinical and medical care aid to improve patient care.	In order for GPT-3 to be considered for use in health-related contexts, there are both possibilities and risks. Patients must be informed that the interaction is with a computer-based text generator, since the trend is to anthropomorphize the technological applications with human traits, assuming humanity and attributing empathetic emotional responses when there are none.
A Research On The New Generation Artificial Intelligence Technology Generative Pre-training Transformer 3.	To analyze the use of GPT-3 in health services through 476 samples, its ability to provide diagnostic help and support to doctors, being used to detect diseases, describe the patient profile based on their symptoms and suggest what might be the best diagnosis; and to describe the resources available in the interface for GPT-3 developers based on the Playground website.	The diagnostic values for accuracy and score were obtained at a level of 77 %. The API parameters must be correctly selected in order for the applications to function effectively and efficiently. These parameters analyzed on the Playground website are: "Execution engine" (Ada, Babbage, Curie and Davinci); "Response length" (amount of text); "Text randomness"; "Frequency and presence penalties" (rating of the model in order to identify new topics and repeat itself); "Best of" (number of completions (n) that the server will generate); "Stop sequence" (character string); "Inject starting text and inject restart text" (to continue with the desired data pattern); and "Show probabilities" (debugging input texts).
Communication		
Exploring the use of GPT-3 as a tool for evaluating text-based collaborative discourse.	To verify the ability of GPT-3 to summarize the students' chat in a learning environment based on a cooperative computer-assisted format.	With only one single sentence that explains the learning context, GPT***T***-3 is able to extract and summarize the students' conversations (correctly attributing emotional states such as frustration and confusion), and synthesize in a reliable manner statements that are not present in the source text, although it does not discriminate among certain topics and does not always recognize hyperbolic statements. Being able to see the summaries of the students' conversations in real time allows teachers to better assign their attention to frustrated and confused students and obtain an idea of the students' progress, although GPT***T***-3 cannot always discern between the content that is considered important in the specific context from a cooperative educational game.

Title	Aim	Main contributions
The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues.	To determine whether the latest generation of generative models, such as Blender and GPT-3, are good AI teachers, capable of responding to a student in an educational dialog.	Conversational agents were run in parallel to human teachers in real-world dialogs, simulating how a student would respond and comparing these answers to Blender and GPT-3 in terms of three skills: talking like a teacher, understanding a student and helping a student. Although the conversational agents (particularly Blender) performed well in capturing conversation, they are worse than real teachers in several instructional aspects. Blender surpassed the real teacher and GPT-3. These findings could be attributed to the particular training Blender received, in which it learned how to be more empathetic and understanding. On the contrary, both Blender and GPT-3 fall far behind human interaction when it comes to helping students.
Understanding emails and drafting responses–an approach using GPT-3.	To explore the possibility of rationalizing email communication, using GPT-3 to reduce its inefficiency through the automated classification of email messages.	GPT-3 can cut costs as compared to the manual writing and classification of email messages. A great demand has been indicated in this regard, in fields ranging from customer service in the insurance industry and public service to the public administrations.

Table 2. Main contributions of the 34 articles included in the review

4. Educational Possibilities and Limitations to the Use of GPT-3

All routine tasks that occupy the teacher's time can be addressed by the chatbot, freeing up the teacher to attend to more important actions related to the learning context and the student's personality. According to Villarroel (2021), GPT-3 makes it possible to quickly diagnose the weaknesses of each student, without the need for the teacher to review a large quantity of questionnaires and tasks, giving them more time for individualized tasks that are aimed at sets of specific skills. AI can be used for automatic learning in which each student is specifically tracked, producing detailed progress reports that specify their comprehension of the different skills being evaluated. The teacher's work thus becomes the personalizing of learning programs according to the strengths and weaknesses of each student. On the other hand, the research by Nguyen, Bhat, Moore, Bier and Stamper (2022) is prompted by the overall lack of practical opportunities in online higher education, as well as by the high labor costs when manually generating questions. Furthermore, the capacity to generate customized questions can greatly aid adaptable and personalized learning technologies, especially in the learning context in which students are asked to continue practicing until they achieve mastery. According to Nguyen et al. (2022), questions that start with "what" are aimed primarily at remembering information. The incorporation of other types of questions on the generation channel could activate more cognitive processes on Bloom's taxonomy. For example, "how" questions can promote comprehension and "why" questions are designed to analyze, which in turn contributes to better learning by students. This diversifying path is also an area of active investigation.

In the work by Jonsson and Tholander (2022), the importance of presenting activities that involve generative AI is emphasized for co-creation. By designing environments that involve post-human design perspective, generativity and unpredictability is placed in the center of learning, replacing a well-defined question-answer system.

The launching of GPT-3 has also awakened a renewed interest in the applicability of NLP to problems associated with contemporary medical care. GPT-3 offers convincing solutions to modern clinical problems (Korngiebel & Mooney, 2021), can process large volumes of text on a scale that is unattainable for human evaluators and requires little or no visual context in order to extract the requested data. It is even able to identify potential participants in clinical trials through the extraction of inclusion and exclusion criteria (Naseri, Kafi, Skamene, Tolba, Faye, Ramia et al., 2022).

With regard to the limitations of GPT-3, according to Nath et al. (2022), GPT-3 incorporates biases related to gender, race and geopolitical condition, and it also has difficulties with the 'common sense physics' and lacks context about the world, since it is not based on domains other than language, such as video and interaction with the real world. On the other hand, Abdelghani, Wang, Yuan, Wang Sauzéon and Oudeyer (2022) observe that the LLMs suffer from a "currency bias", i.e., they place too much trust in the examples that are closer to the end of the message and, therefore, tend to distort the output toward a copy of the most recent examples. On the other hand, following the lines of Goyal, Li and Durrett (2022), adapting GPT-3 to documents longer than the permitted context, or structured inputs, such as tables, poses challenges for research that go beyond the current capabilities of GPT-3, which would be interesting to study. According to Nath et al. (2022), like all the autoregressive language models, GPT-3 also has the limitation that it cannot correct itself once it starts to make mistakes. When writing prose, for example, it cannot go back and edit, and one error often leads to many more, since it uses the previous words to predict its next result. According to Stammbach, Antoniak and Ash (2022), another important limitation for the use of GPT-3 is the cost of the queries. The queries cost nine dollars using the 13B parameter of the GPT-3 model. Extending this to larger corpora containing thousands of documents, it would be prohibitively expensive. According to Aydin and Erdem (2022), the price of the GPT-3 execution engines for every 1000 tokens is $0.0004 in Ada, $0.0005 in Babbage, $0.0020 in Curie and $0.0200 in Davinci. Therefore, the use of even larger pre-trained models would probably not be profitable for most academic research.

As a limitation to this review, there are other aspects that could have been analyzed in the included studies in order to classify them, such as the content type (articles that verify whether AI can be used for a certain task, articles that compare it to human performance, analysis and opinion articles on the consequences of its use, etc.). Additional trends or characteristics could also have been analyzed. Furthermore, of the 34 articles analyzed, 5 were taken from ISI, 4 from SCOPUS and the remaining 25 from Google Scholar, and thus 73% of the literature reviewed has been obtained by means of this database. It is suggested that scientific studies be conducted along these thematic lines and published in indexed peer-reviewed journals, with the aim of providing greater reliability to the scientific contribution that could be derived from it. The following possible research lines can be considered, which could be materialized in future investigations through the use of GPT-3 in the field of education: incorporation of different types of questions in the content generation channel in different areas of educational knowledge, reflective dialogs between students and AI, and the creation of learning models based on AI with cooperative formats. The future of language models like GPT-3 seems promising, but it is also important to continue researching and improving in order to take full advantage of its potential.

5. Conclusions

The review provides information on the use and applications of GPT-3 in the field of education, as well as certain limitations and challenges associated with the language model. GPT-3 has been shown to have valuable applications in education, but it still faces challenges and limitations that require additional research. It is crucial to address the biases (gender, currency, racial, geopolitical, etc.), to improve the performance in broader and structured contexts, and to find more affordable solutions for the use of pre-trained models in different applications. Based on these results, it is possible to draw the following conclusions:

- GPT-3 has proven to be useful in the field of education, especially for automating routine tasks and providing quick diagnoses of the students' weaknesses.
- The automatic generation of questions can improve the opportunities for practice in on-line education and personalize the learning according to the strengths and weaknesses of each student.
- Even though GPT-3 has proven capacities in education, there are still active research areas, such as the diversification of questions and the co-creation of learning environments.

- The presence of biases in GPT-3 is a matter of concern and may require additional investigation in order to properly address this problem.
- It is still not fully understood how to improve the performance of GPT-3 in situations where a broader or more structured context is required, such as large documents or entries with tables.
- It is important to continue to investigate and develop language models like GPT-3, in order to overcome its current limitations, such as the incapacity to correct itself once it starts to make mistakes and the high cost of its queries.
- More work is necessary to improve the capacity of GPT-3 to manage domains that go beyond language, such as video or interaction with the real world, and to reduce biases in its results.
- Researchers can search for more profitable and accessible solutions in order to use larger pre-trained models, especially for academic projects and large-scale applications. For example, models such as BlenderBot-3, BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (A Robustly Optimized BERT Pre-training Approach) have been widely adopted in the research community and can be a more affordable option.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest concerning to the research, authorship, and/or publication of this article. The authors also declared that there was no funding for the research.

Funding

This work has been financed by the 2022/2023 Call for Grants for Research Stays Abroad at La Rioja International University (UNIR).

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. *International Journal of Artificial Intelligence in Education.* https://doi.org/10.1007/s40593-023-00340-7

Agustin, Y.H. (2022). Implementation of Generative Pre-Trained Transformer 3 Classify-Text in Determining Thesis Supervisor. *Sinkron: Jurnal dan Penelitian Teknik Informatika,* 7(4), 2415-2420. https://doi.org/10.33395/sinkron.v7i4.11757

Alawi, F. (2023). Artificial intelligence: The future might already be here. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology,* 135(3), 313-315. https://doi.org/10.1016/j.oooo.2023.01.002

Aljanabi, M., Ghazi, M., Ali, A.H., & Abed, S.A. (2023). ChatGpt: Open Possibilities. *Iraqi Journal for Computer Science and Mathematics,* 4(1), 62-64. https://doi.org/10.52866/20ijcsm.2023.01.01.0018

Alvarez-Carmona, M.A., Aranda, R., Diaz-Pacheco, A., & Ceballos-Mejıa, J.J. (2022). Generador automático de resúmenes científicos en investigación turística. *SciELO Preprints.* https://doi.org/10.1590/scielopreprints.4194

Aydın, N., & Erdem, O.A. (2022). A Research On The New Generation Artificial Intelligence Technology Generative Pretraining Transformer 3. In *3rd International Informatics and Software Engineering Conference (IISEC)* (1-6). IEEE. https://doi.org/10.1109/iisec56263.2022.9998298

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* (701-704). Durham, United Kingdom.

Chan, A. (2022). GPT-3 and Instruct GPT: Technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI Ethics*, 3, 53-64. https://doi.org/10.1007/s43681-022-00148-6

Franganillo, J. (2022). Contenido generado por inteligencia artificial: oportunidades y amenazas. *Anuario ThinkEPI,* 16. https://doi.org/10.3145/thinkepi.2022.e16a24

Gaikwad, A., Rambhia, P., & Pawar, S. (2022). An Extensive Analysis Between Different Language Models: GPT-3, BERT and MACAW. *Research Square.* https://doi.org/10.21203/rs.3.rs-2155616/v1

Garritty, C., Stevens, A., Gartlehner, G., King, V., & Kamel, C. (2016). Cochrane Rapid Reviews Methods Group to play a leading role in guiding the production of informed high-quality, timely research evidence syntheses. *Systematic Reviews,* 5(1),184. https://doi.org/s13643-016-0360-z

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* National Comprehensive Center for Teacher Quality.

Goyal, T., Li, J.J., & Durrett, G. (2022). News Summarization and Evaluation in the Era of GPT-3. *arXiv,* arXiv:2209.12356. https://doi.org/10.48550/arXiv.2209.12356

Hamel, C., Michaud, A., Thuku, M., Skidmore, B., Stevens, A., Nussbaumer-Streit, B. et al. (2021). Defining rapid reviews: a systematic scoping review and thematic analysis of definitions and defining characteristics of rapid reviews. *Journal of Clinical Epidemiology,* 129, 74-85. https://doi.org/10.1016/j.jclinepi.2020.09.041

Jonsson, M., & Tholander, J. (2022). Cracking the code: Co-coding with AI in creative programming education. In *Proceedings of the 14th Conference on Creativity and Cognition* (5-14). https://doi.org/10.1145/3527927.3532801

Korngiebel, D.M., & Mooney, S.D. (2021). Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digital Medicine,* 4(1), 1-3. https://doi.org/10.1038/s41746-021-00464-x

Lammerse, M., Hassan, S.Z., Sabet, S.S., Riegler, M.A., & Halvorsen, P. (2022). Human vs. GPT-3: The challenges of extracting emotions from child responses. In *14th International Conference on Quality of Multimedia Experience (QoMEX)* (1-4). IEEE. https://doi.org/10.1109/qomex55416.2022.9900885

Lee, Y., Kim, T.S., Chang, M., & Kim, J. (2022). Interactive Children's Story Rewriting Through Parent-Children Interaction. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing)* (62-71). https://doi.org/10.18653/v1/2022.in2writing-1.9

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* (2, 37-39). https://doi.org/10.1145/3501709.3544280

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 *European Conference on Technology Enhanced Learning* (243-257). Springer, Cham. https://doi.org/10.1007/978-3-031-16290-9_18

Naseri, H., Kafi, K., Skamene, S., Tolba, M., Faye, M.D., Ramia, P. et al. (2021). Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases. *Journal of Biomedical Informatics,* 120, 103864. https://doi.org/10.1016/j.jbi.2021.103864

Nath, S., Marie, A., Ellershaw, S., Korot, E., & Keane, P.A. (2022). New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. *British Journal of Ophthalmology,* 106(7). https://doi.org/10.1136/bjophthalmol-2022-321141

Nguyen, H.A., Bhat, S., Moore, S., Bier, N., & Stamper, J. (2022). Towards Generalized Methods for Automatic Question Generation in Educational Domains. In *European Conference on Technology Enhanced Learning* (272-284). Springer, Cham. https://doi.org/10.1007/978-3-031-16290-9_20

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D. et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International Journal of Surgery,* 88, 105906. https://doi.org/10.31222/osf.io/jb4dx

Phillips, T., Saleh, A., Glazewski, K.D., Hmelo-Silver, C.E., Mott, B., & Lester, J.C. (2022). Exploring the use of GPT-3 as a tool for evaluating text-based collaborative discourse. *Companion Proceedings of the 12th International Conference on Learning Analytics & Knowledge (LAK22)*, (1-3).

Ramnarain-Seetohul, V., Bassoo, V., & Rosunally, Y. (2022). Work-in-Progress: Computing Sentence Similarity for Short Texts using Transformer models. In *IEEE Global Engineering Education Conference (EDUCON)* (1765-1768). IEEE. https://doi.org/10.1109/educon52537.2022.9766649

Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research* (1, 27-43). https://doi.org/10.1145/3501385.3543957

Stammbach, D., Antoniak, M., & Ash, E. (2022). Heroes, Villains, and Victims, and GPT-3: Automated Extraction of Character Roles Without Training Data. *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)* (47-56). https://doi.org/10.18653/v1/2022.wnu-1.6

Tack, A., & Piech, C. (2022). The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. *arXiv,* arXiv:2205.07540. https://doi.org/10.48550/arXiv.2205.07540

Thiergart, J., Huber, S., & Übellacker, T. (2021). Understanding emails and drafting responses–an approach using GPT-3. *ArXiv,* arXiv:2102.03062. https://doi.org/10.48550/arXiv.2102.03062

Villarroel, J.J.G. (2021). Implicancia de la inteligencia artificial en las aulas virtuales para la educación superior. *Orbis Tertius-UPAL,* 5(10), 31-52.

Zapata, V.J.V., & Hoyos, D.M.R. (2023). Correspondencia artificial: exploraciones del ChatGPT y sus implicaciones en el quehacer académico. *Folios, revista de la Facultad de Comunicaciones y Filología,* 49, 36-41.

Zong, M., & Krishnamachari, B. (2022). Solving math word problems concerning systems of equations with GPT-3. *Proceedings of the AAAI Conference on Artificial Intelligence,* 37(13), 15972-15979. https://doi.org/10.1609/aaai.v37i13.26896

Published by OmniaScience (www.omniascience.com)

Journal of Technology and Science Education, 2024 (www.jotse.org)