


International Journal of  
Interactive Multimedia  
and Artificial Intelligence

March 2024, Vol. VIII, Number 5  
ISSN: 1989-1660

**unir** LA UNIVERSIDAD  
EN INTERNET



*“But what you learn, as you get older, is that there are a few billion other people in the world all trying to be clever at the same time, and whatever you do with your life will certainly be lost—swallowed up in the ocean—unless you are doing it along with like-minded people who will remember your contributions and carry them forward.”*

*Neal Stephenson, quote from The Diamond Age*

Special Issue on Generative Artificial Intelligence in Education

## **EDITORIAL TEAM**

### **Editor-in-Chief**

Dr. Elena Verdú, Universidad Internacional de La Rioja (UNIR), Spain

Dr. Rubén González Crespo, Universidad Internacional de La Rioja (UNIR), Spain

### **Managing Editors**

Dr. Xiomara Patricia Blanco Valencia, Universidad Internacional de La Rioja (UNIR), Spain

Dr. Paulo Alonso Gaona-García, Universidad Distrital Francisco José de Caldas, Colombia

### **Office of Publications**

#### **Editorial Coordination**

Dr. Pedro Hípola, Universidad Internacional de La Rioja (UNIR), Spain

Lic. Blanca Albarracín, Universidad Internacional de La Rioja (UNIR), Spain

#### **Indexing and Metrics**

Dr. Álvaro Cabezas Clavijo, Universidad Internacional de La Rioja (UNIR), Spain

Lic. Mercedes Contreras, Universidad Internacional de La Rioja (UNIR), Spain

#### **Layout and Graphic Edition**

Lic. Ainhoa Puente, Universidad Internacional de La Rioja (UNIR), Spain

### **Associate Editors**

Dr. Enrique Herrera-Viedma, University of Granada, Spain

Dr. Witold Perdrycz, University of Alberta, Canada

Dr. Kuan-Ching Li, Providence University, Taiwan

Dr. Robertas Damaševičius, Kaunas University of Technology, Lithuania

Dr. Javier Martínez Torres, Universidad de Vigo, Spain

Dr. Miroslav Hudec, University of Economics of Bratislava, Slovakia

Dr. Seifedine Kadry, Noroff University College, Norway

Dr. Nilanjan Dey, Techno International New Town, India

Dr. Mahdi Khosravy, Cross Labs, Cross Compass Ltd., Tokyo, Japan

Dr. Jörg Thomaschewski, Hochschule Emden/Leer, Emden, Germany

Dr. Mu-Yen Chen, National Cheng Kung University, Taiwan

Dr. Francisco Mochón Morcillo, National Distance Education University, Spain

Dr. Manju Khari, Jawaharlal Nehru University, New Delhi, India

Dr. Carlos Enrique Montenegro Marín, Francisco José de Caldas District University, Colombia

Dr. Yaping Mao, Qinghai Normal University, China

Dr. Juan Manuel Corchado, University of Salamanca, Spain

Dr. Giuseppe Fenza, University of Salerno, Italy

Dr. S.P. Raja, Vellore Institute of Technology, Vellore, India

Dr. Jerry Chun-Wei Lin, Western Norway University of Applied Sciences, Norway

Dr. Juan Antonio Morente, University of Granada, Spain

Dr. Abbas Mardani, The University of South Florida, USA

Dr. Amrit Mukherjee, University of South Bohemia, Czech Republic

Dr. Suyel Namasudra, National Institute of Technology Agartala, India

Dr. Ricardo S. Alonso, AIR Institute, Spain

### **Editorial Board Members**

Dr. Rory McGreal, Athabasca University, Canada

Dr. Óscar Sanjuán Martínez, Lumen Technologies, USA

Dr. Anis Yazidi, Oslo Metropolitan University, Norway

Dr. Juan Pavón Mestras, Complutense University of Madrid, Spain

Dr. Smriti Srivastava, Netaji Subhas University of Technology, New Delhi, India

Dr. Lei Shu, Nanjing Agricultural University, China/University of Lincoln, UK  
Dr. Ali Selamat, Malaysia Japan International Institute of Technology, Malaysia  
Dr. Hamido Fujita, Iwate Prefectural University, Japan  
Dr. Francisco García Peñalvo, University of Salamanca, Spain  
Dr. Francisco Chiclana, De Montfort University, United Kingdom  
Dr. S. Vimal, Ramco Institute of Technology, Tamil Nadu, India  
Dr. Jordán Pascual Espada, Oviedo University, Spain  
Dr. Ioannis Konstantinos Argyros, Cameron University, USA  
Dr. Ligang Zhou, Macau University of Science and Technology, Macau, China  
Dr. Palanichamy Naveen, KPR Institute of Engineering and Technology, India  
Dr. Juan Manuel Cueva Lovelle, University of Oviedo, Spain  
Dr. Pekka Siirtola, University of Oulu, Finland  
Dr. Peter A. Henning, Karlsruhe University of Applied Sciences, Germany  
Dr. Yago Saez, Universidad Carlos III de Madrid, Spain  
Dr. Vijay Bhaskar Semwal, National Institute of Technology, Bhopal, India  
Dr. Anand Paul, Kyungpook National Univeristy, South Korea  
Dr. Javier Bajo Pérez, Polytechnic University of Madrid, Spain  
Dr. Jinlei Jiang, Dept. of Computer Science & Technology, Tsinghua University, China  
Dr. B. Cristina Pelayo G. Bustelo, University of Oviedo, Spain  
Dr. Masao Mori, Tokyo Institue of Technology, Japan  
Dr. Rafael Bello, Universidad Central Marta Abreu de Las Villas, Cuba  
Dr. Daniel Burgos, Universidad Internacional de La Rioja - UNIR, Spain  
Dr. JianQiang Li, Beijing University of Technology, China  
Dr. Rebecca Steinert, RISE Research Institutes of Sweden, Sweden  
Dr. Monique Janneck, Lübeck University of Applied Sciences, Germany  
Dr. Carina González, La Laguna University, Spain  
Dr. Mohammad S Khan, East Tennessee State University, USA  
Dr. David L. La Red Martínez, National University of North East, Argentina  
Dr. Juan Francisco de Paz Santana, University of Salamanca, Spain  
Dr. José Estrada Jiménez, Escuela Politécnica Nacional, Ecuador  
Dr. Octavio Loyola-González, Stratesys, Spain  
Dr. Guillermo E. Calderón Ruiz, Universidad Católica de Santa María, Peru  
Dr. Moamin A Mahmoud, Universiti Tenaga Nasional, Malaysia  
Dr. Madalena Riberio, Polytechnic Institute of Castelo Branco, Portugal  
Dr. Manik Sharma, DAV University Jalandhar, India  
Dr. Edward Rolando Núñez Valdez, University of Oviedo, Spain  
Dr. Juha Röning, University of Oulu, Finland  
Dr. Paulo Novais, University of Minho, Portugal  
Dr. Sergio Ríos Aguilar, Technical University of Madrid, Spain  
Dr. Hongyang Chen, Fujitsu Laboratories Limited, Japan  
Dr. Fernando López, Universidad Complutense de Madrid, Spain  
Dr. Runmin Cong, Beijing Jiaotong University, China  
Dr. Manuel Perez Cota, Universidad de Vigo, Spain  
Dr. Abel Gomes, University of Beira Interior, Portugal  
Dr. Víctor Padilla, Universidad Internacional de La Rioja - UNIR, Spain  
Dr. Mohammad Javad Ebadi, Chabahar Maritime University, Iran  
Dr. Andreas Hinderks, University of Sevilla, Spain  
Dr. Brij B. Gupta, National Institute of Technology Kurukshetra, India  
Dr. Alejandro Baldominos, Universidad Carlos III de Madrid, Spain



## TABLE OF CONTENTS

GENERATIVE ARTIFICIAL INTELLIGENCE IN EDUCATION: FROM DECEPTIVE TO DISRUPTIVE .....	5
A CYBERNETIC PERSPECTIVE ON GENERATIVE AI IN EDUCATION: FROM TRANSMISSION TO COORDINATION .....	15
ETHICAL IMPLICATIONS AND PRINCIPLES OF USING ARTIFICIAL INTELLIGENCE MODELS IN THE CLASSROOM: A SYSTEMATIC LITERATURE REVIEW .....	25
A TRUSTWORTHY AUTOMATED SHORT-ANSWER SCORING SYSTEM USING A NEW DATASET AND HYBRID TRANSFER LEARNING METHOD.....	37
VIRTUAL REALITY AND LANGUAGE MODELS, A NEW FRONTIER IN LEARNING.....	46
GENERATIVE ARTIFICIAL INTELLIGENCE IN PRODUCT DESIGN EDUCATION: NAVIGATING CONCERNS OF ORIGINALITY AND ETHICS .....	55
CAN GENERATIVE AI SOLVE GEOMETRY PROBLEMS? STRENGTHS AND WEAKNESSES OF LLMS FOR GEOMETRIC REASONING IN SPANISH.....	65
EVALUATING CHATGPT-GENERATED LINEAR ALGEBRA FORMATIVE ASSESSMENTS .....	75

### OPEN ACCESS JOURNAL

ISSN: 1989-1660

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded*, *Journal Citation Reports/ Science Edition*, *Current Contents®/Engineering Computing and Technology*.

### COPYRIGHT NOTICE

Copyright © 2024 UNIR. This work is licensed under a Creative Commons Attribution 3.0 unported License. You are free to make digital or hard copies of part or all of this work, share, link, distribute, remix, transform, and build upon this work, giving the appropriate credit to the Authors and IJIMAI, providing a link to the license and indicating if changes were made. Request permission for any other issue from [journal@ijimai.org](mailto:journal@ijimai.org).

<http://creativecommons.org/licenses/by/3.0/>

# Generative Artificial Intelligence in Education: From Deceptive to Disruptive

Marc Alier<sup>1</sup>, Francisco José García-Peñalvo<sup>2</sup>, Jorge D. Camba<sup>3</sup> †\*

† Guest editors of the Special Issue on Generative Artificial Intelligence in Education

<sup>1</sup> Polytechnic University of Catalonia (Spain)

<sup>2</sup> Research Institute for Educational Sciences, Universidad de Salamanca (Spain)

<sup>3</sup> Purdue University (USA)

Received 26 February 2024 | Accepted 27 February 2024 | Published 28 February 2024



## ABSTRACT

Generative Artificial Intelligence (GenAI) has emerged as a promising technology that can create original content, such as text, images, and sound. The use of GenAI in educational settings is becoming increasingly popular and offers a range of opportunities and challenges. This special issue explores the management and integration of GenAI in educational settings, including the ethical considerations, best practices, and opportunities. The potential of GenAI in education is vast. By using algorithms and data, GenAI can create original content that can be used to augment traditional teaching methods, creating a more interactive and personalized learning experience. In addition, GenAI can be utilized as an assessment tool and for providing feedback to students using generated content. For instance, it can be used to create custom quizzes, generate essay prompts, or even grade essays. The use of GenAI as an assessment tool can reduce the workload of teachers and help students receive prompt feedback on their work. Incorporating GenAI in educational settings also poses challenges related to academic integrity. With availability of GenAI models, students can use them to study or complete their homework assignments, which can raise concerns about the authenticity and authorship of the delivered work. Therefore, it is important to ensure that academic standards are maintained, and the originality of the student's work is preserved. This issue highlights the need for implementing ethical practices in the use of GenAI models and ensuring that the technology is used to support and not replace the student's learning experience.

## KEYWORDS

Artificial Intelligence, Ethical Implications, Ethical Principles, Generative Artificial Intelligence, Large Language Model.

DOI: 10.9781/ijimai.2024.02.011

## I. LARGE LANGUAGE MODELS TAKE ARTIFICIAL INTELLIGENCE FROM DECEPTIVE TO DISRUPTIVE

**T**ECHNOLOGY has evolved rapidly in the last few years, affecting many areas, including education. The launch of ChatGPT on November 30, 2022, was a key event in the history of the Artificial Intelligence (AI). For the first time, a technology labeled AI went mainstream, becoming the fastest-growing consumer product of all time, getting 1 million users in just five days (See Fig. 1) and reaching 100 million active users in less than two months (Fig. 2) [1]. The new chatbot has had a deep cultural impact, bringing the rapidly advancing field of AI and its societal impacts to the forefront of public attention.

The model of the 6Ds of digitized technologies introduced by Peter Diamandis and Steven Kotler [2] showcases the significance of the Generative AI (GenAI) [3] moment in 2023. The 6D model states that when something is digitized, it goes through six phases:

1. Digitized. A resource, a technology, a process, or a social or economic activity becomes digital, it will evolve at an exponential

pace following the pace of improvement described by Moore's Law [4] and other exponential behaviors observed in digital technologies (computing, memory, digital storage, bandwidth, etc.).

2. Deceptive. In the first stages, the digitized version will be inferior to the old analog version, and its evolution will be deceptively slower than the linear, steady improvements of analog alternatives. A classic example is digital photography, invented in the 70s by Kodak, which was inferior to chemical film for over 30 years.
3. Disruptive. The exponential curve of growth kicks and the technological improvement mimics a hockey stick curve. The digitized version becomes disruptive, deeming the previous technology obsolete in a very short period of time. The following phases are observed after the disruption.
4. Demonetization. Marginal costs tend to be zero. Taking one more digital picture is close to zero, just like doing a web search, watching an online video, or making a social media post. Kodak went bankrupt in 2010, the same year that Instagram was acquired by Facebook (now Meta) for an unprecedented sum. Instagram's business model relied on the zero marginal cost of taking and posting a picture online.
5. Democratization. Access to the digitized version becomes universal. While the paper volumes of an encyclopedia were expensive and took significant physical space, Wikipedia is open

\* Corresponding author.

E-mail addresses: marc.alier@upc.edu (M. Alier), fgarcia@usal.es (F. J. García-Peñalvo), jdorribo@purdue.edu (J. D. Camba).

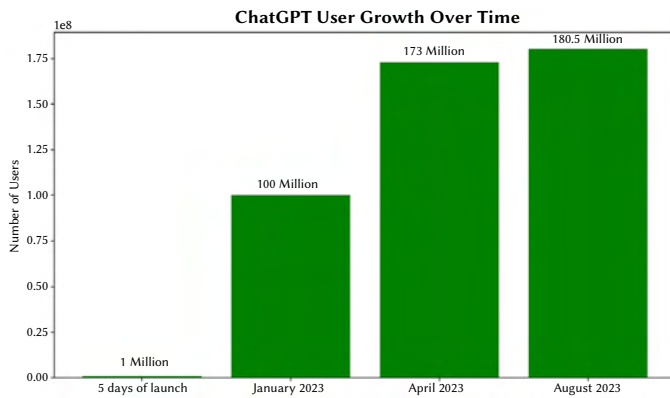


Fig. 1. Number of users of ChatGPT during the first nine months. Source: Own production adapted from <https://nerdynav.com/chatgpt-statistics/>

to anyone with access to an online device, which makes it more valuable to own such devices.

- Dematerialization. The analog artifact is no longer required, just like the traditional encyclopedia and photo albums are prescinding, and we can take back the space they occupied on our shelves.

Leveraging foundational technologies like neural networks [5], deep learning [6], transformers [7], and quantum technologies [8], Large Language Models (LLMs) [9], [10] like ChatGPT are navigating the trajectory outlined in the 6D model. Starting from an understated impact, these models are transitioning to a phase of significant disruption. This ongoing shift is evident in their rapid adoption and increasing prominence across diverse sectors, indicating a growing and extensive impact on the economy and culture.

If the exponential trends in AI continue, we can anticipate significant performance and changes in emergent functionalities. A notable trend is the rise of Large Multimodal Models (LMMs) or Multimodal Large Language Models (MLLMs) [11], such as GPT-4V [12], trained on diverse data types like text, images, sound, video, and infrared images. These LMMs exhibit surprising capabilities such as creating stories from images or performing OCR-free math reasoning [13], suggesting a potential path to Artificial General Intelligence (AGI) [14].

The year 2023 has been remarkable for LLMs, with exponential or sigmoid growth in various dimensions: enhanced capabilities, increased model sizes, new models and projects, heightened investment, and public attention. Nevertheless, it has also led to one of the quickest government reactions to a new technology. The U.S. Federal Elections Commission is investigating misleading political ads, and Congress demands more oversight on how AI firms manage and identify their training data. Likewise, the European Union has updated its AI Act to address GenAI [15]. Additionally, it has emphasized the ethical aspects of Information and Communication Technologies (ICTs) and AI [16]. Significant debates have arisen over the impact of AI technologies on society and the job market, as well as philosophical discussions about the potential catastrophic consequences of AGI and Superintelligent Artificial Intelligence (SIA) [17] for humanity.

#### A. Discovering the Emergent Abilities of Large Language Models

It is important to consider that ChatGPT is just the tip of the iceberg in the innovations emerging from the GenAI sector, a field heavily reliant on transformer models [7] and diffusion techniques [18]. ChatGPT is a chatbot based on an adaptation of the GPT-3 LLM (specifically GPT-3.5-Turbo) [19] (with a 175 billion-parameter architecture capable of handling a context window of 4,096 tokens, about 2,500 words) and, on its enhanced version, the GPT-4 model [20] (with a context window of 32K tokens). Information about GPT 4.0 has not been opened to the community. It is estimated to be a model of

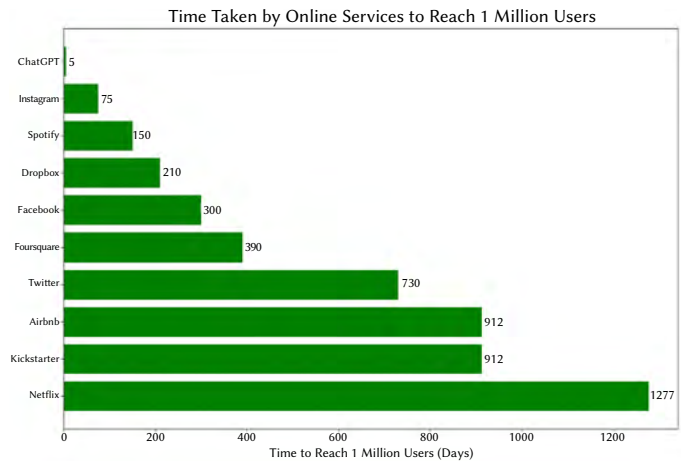


Fig. 2. Time taken by Online Services to reach 1 million users. Source: Own production adapted from <https://nerdynav.com/chatgpt-statistics/>

about 1.8 trillion parameters organized as a MoE (Mixture of Experts), with 16 experts of 111 billion parameters, plus the trunk part of 55 billion parameters, activating only two experts for each inference (280 billion parameters) [21].

LLMs are enormous neural network systems based on the transformer architecture [7], introduced by 2017 DeepMind, a company acquired by Alphabet (Google) in 2017. Since then, transformers have become the go-to architecture in AI research, serving as a kind of *lingua franca* among AI research subfields. Previously, these subfields had diverged so much in their theoretical approaches that innovations in one area rarely permeated others [22].

Creating an LLM involves several key steps:

- Model architecture. This is the code and mathematical framework of the model. Most top-performing LLMs currently use variations of the “decoder-only” transformer architecture.
- Training dataset. This includes all the examples and documents on which the model is trained, shaping its learned patterns. The content typically consists of text in natural or programming languages or structured data like tables or equations.
- Tokenizer. It converts the text from the training dataset into numerical values, as models require numbers for processing. Text is transformed into tokens (words, sub-words, or characters) based on the tokenization method. The size of a dataset is often measured by the number of these tokens, which can range from hundreds of billions to several trillion.
- Training hyperparameters. These define the specifics of the training process, including the rate of parameter adjustments and model updates.
- Computing power and human oversight. Adequate computing resources and skilled personnel are essential for running and monitoring the training process. The training involves setting up the architecture on hardware and running the training algorithm with the chosen hyperparameters, resulting in a set of learned model weights.
- Post-training. LLMs can be specialized or adapted for specific tasks through fine-tuning. This involves additional training on a more specialized dataset, optimizing the model for particular applications. Though costly in terms of computing power, this step is generally less expensive than training a model from scratch. High-quality open-source pre-trained models are valuable in this context, as they allow for community-driven development and application, even with limited computing resources [23].

Recent developments show that as large pre-trained models grow in size with billions of parameters, they reveal unique properties [24]. In particular, it seemed that models going above specific size thresholds jumped in capabilities, two concepts that were dubbed “emergent abilities” and “scaling laws.”

For instance, OpenAI’s GPT models display this evolution. The original GPT could manage basic text labeling but lacked coherence in text generation [25]. GPT-2 improved, offering higher-quality text and some instruction-following capabilities [26]. GPT-3, however, emerged as a versatile and practical LLM for various language tasks. The significant capability leap between these models is mainly due to scaling up computational power and data: GPT-3 required about 20,000 times more computation than the original GPT [27]. Although these models share similar designs, their advancements are largely attributed to breakthroughs in high-performance computing infrastructure rather than specific advancements in language technology model design.

As they scale up, LLMs exhibit new properties that their developers had not anticipated, and we are only now starting to discover them. Among these properties, few-shot learning and chain-of-thought reasoning stand out.

- Few-shot learning enables a sufficiently large LLM to quickly grasp new tasks from just a few examples in a single interaction [26].
- Chain-of-thought reasoning allows the model to articulate its thought process when tackling complex tasks, similar to how the students would explain their reasoning during a math test, thereby enhancing their performance [28].

These GPT-3 capabilities, particularly in few-shot learning and chain-of-thought reasoning, were identified post-training and several months after its widespread public deployment, respectively [29]-[32].

In hindsight, these characteristics are partly the consequence of the LLMs’ ability to “learn” from the information within the context of execution -all the information received from user messages and that the model has generated- like the training and fine-tuning data.

Furthermore, LLMs demonstrate unforeseen skills in programming, arithmetic, correcting misconceptions, and answering exam questions across various domains, and improve as the model size scales up [30], [33].

There is a common belief that LLMs are merely statistical predictors of the next word, limited to text-based learning and reasoning. However, recent evidence suggests that LLMs are developing internal representations of the world, enabling them to reason abstractly beyond the specific linguistic structure of texts [24]. Although this ability is currently limited and inconsistent, it is most evident in larger and newer models, indicating that it could strengthen with further scaling of these systems. Key findings supporting this include:

- LLMs’ internal representations of color words align closely with human color perception [34].
- They can infer authors’ knowledge or beliefs from a document and predict its continuation [35].
- LLMs internally represent properties and locations of objects in stories, evolving as new information is presented. This includes representing spatial layouts in story settings and real-world geography and providing instructions for drawing novel objects.
- LLMs develop internal representations of the game board’s state when trained on board games using descriptions of moves [36].
- LLMs can differentiate between misconceptions and facts, showing calibrated internal representations of truth likelihood [30].
- LLMs pass tests designed for common-sense reasoning, including

those like the Winograd Schema Challenge [37], which lack textual clues for answers.

These findings indicate a growing ability of LLMs to develop complex, abstract internal models that extend beyond simple text processing.

### B. Size Matters: Openness Is the Key

In the previous section, we discussed the matter of size in LLMs. Models above specific size thresholds seemed to jump in “emergent abilities” according to certain “scaling laws.” However, in March 2022, DeepMind released a paper exploring the ideal balance between tokens and model parameters within a set compute budget for LLM training [38]. The study suggests that smaller models with significantly more data are more effective for an average budget. For instance, the Chinchilla model, which is not open source, had 70B parameters (a third the size of larger models) but was trained on 1.4T tokens of data (3 to 4 times more). This approach led to comparable or better performance than larger models, both open and closed source.

According to Clémentine Fourrier [23], “this paradigm shift, while probably already known in a closed lab, took the open science community by storm.” In 2023, we witnessed a wave of open-source releases of pre-trained LLMs released almost daily. Noteworthy releases LLaMA (by Meta) in February, Pythia (by Eleuther AI) in April, MPT (by MosaicML) in May, X-GEN (by Salesforce) and Falcon (by TIUAE) in June, Llama 2 (by Meta) in July. Qwen (by Alibaba) and Mistral (by Mistral AI) in September, Yi (by 01-ai) in November, DeciLM (by Deci), Phi-2, and SOLAR (by Upstage) in December.

These models, with parameters ranging between 3B and 70B, have quickly gained adoption for their performance and varying open-source licenses. Most models incorporate decoder transformer architecture with modifications and varying attention functions. While performance and inference speeds differ, the primary distinctions among these publicly released architectures are their training data and licensing.

These releases of open-source LLMs, along with other notable open-source AI models in image processing like Stability.ai’s Stable Diffusion, and audio processing models, such as OpenAI’s speech-to-text model “Whisper,” have sparked great excitement among the developer community worldwide. Throughout 2023, we have witnessed a surge in the number of software projects related to generative AI (Fig. 3).

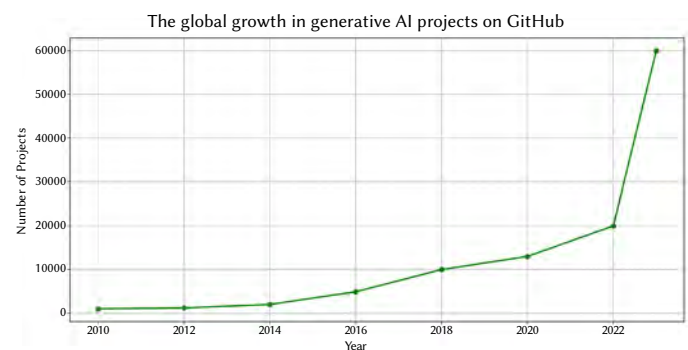


Fig. 3. Global growth in Generative AI software projects. Source: Own production adapted from <https://d66z.short.gy/3f10bE>

Scores, perhaps hundreds of thousands of independent software developers, researchers, and entrepreneurs worldwide, have begun experimenting with these technologies. Whether working with open models or developing against the OpenAI’s Application Programming Interface (API) and other proprietary LLM providers like Google or Anthropic, this vibrant activity leads to experimentation in new use



cases, applications, and technologies based on and complementary to AI models.

In early 2023, a group of Stanford students utilized OpenAI's text\_davinci-003 API to generate a fine-tuning dataset, leading to the development of Alpaca 7B [39]. This model, fine-tuned from the LLaMA 7B model [40], is designed to follow instructions based on 52K demonstrations. Alpaca exhibits similar capabilities to OpenAI's text-davinci-003 but is notably smaller and more cost-effective to reproduce, with an estimated cost of under \$600 [41].

An internal Google document, leaked in spring 2023 (<https://d66z.short.gy/u7blNr>), reveals insights on the competitive landscape of AI. It suggests that open-source AI is outpacing giants like Google and OpenAI, particularly in the realms of LLMs. This shift is attributed to the speed, customization, privacy, and capabilities of open-source models, even with fewer resources. The document highlights open-source models achieving remarkable feats with significantly lower budgets, challenging the traditional approach of building giant, costly models.

At the time of this writing (early 2024), the best-performing published LLM, according to the Aena ELO Rating [42], is OpenAI's GPT-4-Turbo-1106. However, in the top, we find two open-sourced LLMs: Mixtral-8x7b-instruct from the French firm Mistral AI and Tulu-2-DPO-70B from Paul Allen's AllenAI (<https://d66z.short.gy/A5XMno>). While the final draft of this paper is being written, new models such as Gemini Pro 1.5 with 1M tokens of context [43] are being introduced in the leaderboard, still without surpassing the latest GPT-4 on overall performance.

However, just days before the publication of this special issue, the project LoRA [44] (<https://d66z.short.gy/pKHBNG>) has released a specialized set of fine-tuned versions of Mistral 7b, an open-source LLM by the French company Mistral AI, that can be run on a medium spec laptop, where each specialized small LoRA LLM outperforms GPT-4 significantly on a specific benchmark [45].

## II. GENAI AND EDUCATION

### A. Towards the Young Lady's Illustrated Primer

In Neal Stephenson's science fiction novel "The Diamond Age" [46], one of the central pieces of educational technology is the "Young Lady's Illustrated Primer." This device is a highly advanced, interactive book that uses AI to tailor educational content and tutoring to the individual learner. Designed initially for an elite clientele, the Primer adapts to its users' interests, learning pace, and developmental needs, providing personalized education. The story plot places the Illustrated Primer in the hands of a poor girl, who turns her life's path around.

The Primer goes beyond traditional educational tools in several ways. First, it engages with the user through interactive storytelling, making learning an immersive experience. The stories it tells are not static but evolve based on the user's interactions and choices, teaching problem-solving, critical thinking, and moral reasoning. Second, the AI in the Primer is capable of understanding and responding to the emotional and cognitive state of the user, providing support and challenges that are appropriate for the user's current level of understanding. This aspect of the Primer reflects a deep integration of AI into the educational process, offering a vision of how technology might be used to create highly individualized learning experiences.

Just as Stephenson's previous book "Snowcrash" [47] has inspired many modern technologies that are or might become a reality (virtual reality, augmented reality, internet of things, surveillance of workers with data analytics, cryptocurrencies, and smart contracts, networked states [48], and even a virtual librarian character that could easily be

a near future product evolved from ChatGTP), "The Diamond Age's" Illustrated Primer is an inspiration for the next wave of educational technologies.

There is no lack of techno-optimists and capital to push a new wave of technologies that are moving from deceptive to disruptive. Diamandis and Kotler showcase in their book "The future is faster than you think" a student's field trip to a virtual-reality Ancient Rome, accompanied by an AI instructor to illustrate the educational transformative applications of the combination of GenAI, virtual reality, and augmented reality [49].

But, paraphrasing Darth Vader in Star Wars first film (chronologically), before we get too proud of the technological monstrosity we are about to construct, let us take a step back and reconsider what we have learned about educational technologies (EdTech).

### B. Education Is More Than a Marketplace

The worldwide education market was valued at approximately \$6,682.46 billion in 2022 (<https://d66z.short.gy/zYVVYD>). This market encompasses a wide range of segments, including K-12 education, higher education, vocational education, corporate training, and various modes of delivery such as online learning, in-person learning, and blended learning.

In the last 25 years, educational technology has undergone significant transformation. The advent of the internet in the mid-1990s marked the beginning of a new era in education. Early technologies were primarily focused on computer-based learning and multimedia content in classrooms [50]. However, the early 2000s witnessed a surge in online learning platforms [51], revolutionizing access to education. This period saw the introduction of virtual classrooms, e-learning modules, and interactive educational software. The proliferation of mobile technology and tablets in the 2010s further expanded the reach of digital learning, allowing students to access educational resources anytime, anywhere [52]. More recently, advancements in AI, virtual and augmented reality, and adaptive learning systems have further personalized the learning experience, catering to individual learning styles and needs [53], [54]. This rapid evolution of technology has broadened the scope of education and brought about a paradigm shift in teaching methodologies and learning processes.

During all these years, the landscape of educational technology has been marked by a striking duality. On the one hand, there is an undeniable commercialization, with education increasingly influenced by market-driven models and private enterprises [55]. On the other hand, there is a growing movement towards open-source technologies [56] and freely accessible content repositories [57]. This contrast paints a complex picture of the current educational space, where the forces of commodification coexist with a commitment to open access and knowledge sharing.

The current landscape of educational technology, whether open-source or privately owned, demands a critical examination of its approach, implementation, and application. The following are several key issues:

1. Narrow focus on learning. Educational technology often emphasizes "learning" and "learners," a concept termed "learnification." This overlooks vital educational aspects like socialization, subjectification, qualification, and contextual factors [58]. Tools like Learning Management Systems (LMSs) tend to function more as management tools than learning aids, limiting the understanding of digital technology's role in education [59].
2. Technology over pedagogy. The idea that technology should be integrated with teaching methods to enhance education truly is often overlooked. Blending technology with effective



teaching strategies is crucial for real progress in education. This ensures that technology exists in the classroom and supports and improves learning outcomes [60]. This concept has been introduced previously. Back in the 1980s, Seymour Papert [61] observed similar issues within the LOGO community. He criticized the usual ways of evaluating educational technology, such as controlled experiments and product reviews. Papert argued for a more comprehensive approach considering the social and cultural aspects of using computers in education. His viewpoint challenges the common, technology-focused mindset in education. Instead of looking at how technology fits into education, he suggested a more culturally aware evaluation of its role. This approach remains relevant today as we continue to explore the most effective ways to integrate technology in learning environments.

3. Emotional and human impact. Understanding digital tools' emotional and human impact is crucial [62]. These technologies influence students' and staff's emotions, values, and behaviors, and their role in learning environments should be supportive and enriching. Online learning technologies, especially LMS, inherently exhibit an "architecture of control" in their design. The user interface and design choices subtly shape users' behavior and interactions, potentially limiting educational exploration and autonomy. Furthermore, integrating learning analytics introduces continuous monitoring and analysis of student data [63]. While aimed at personalizing and enhancing learning, this constant surveillance raises privacy and psychological concerns [64]. The educational journey can become algorithm-driven, often without transparently acknowledging underlying decision-making processes [65].

### C. ChatGPT Goes to School

The domain of education has historically pioneered the assimilation of technological advancements. In the last decades, many software applications have been developed and evolved to cater to diverse educational requisites, spanning online learning, language acquisition, academic research, pedagogical support, content generation, and professional development.

The infusion of AI into education is not a recent phenomenon [66]. Despite years of dedicated research and substantial financial investments, the field has yet to yield substantial impacts beyond research and development, with only a handful of commercial products achieving limited influence.

The emergence of ChatGPT has metamorphosed AI's role in education from a theoretical construct into an immediate reality. This paradigm shift transpired virtually overnight, organically gaining traction without advertising or marketing campaigns. Stakeholders, including students, educators, and administrators, have instinctively grasped this transformation's significance, urgency, and potential, even though the precise course of action still needs to be discovered [67].

As compelling proof of this rapid and widespread interest, many teachers are enrolling in different courses about integrating ChatGPT and GenAI tools in their classrooms and courses. Most teachers who participated in these courses cited three primary motivations for their interest in ChatGPT's role in education. First, they expressed concerns about the potential for increased plagiarism facilitated by the technology. Second, they were intrigued by the implications of automating academic tasks within their specific fields of expertise. Last but not least, they were interested in how ChatGPT could enhance students' educational experiences and outcomes.

The advent of ChatGPT (as the most known GenAI tool) has further enriched the educational technological landscape, offering, among others, [68]:

- Diverse educational opportunities: ChatGPT and similar LLMs can generate instructional content, facilitate discussions on diversity and inclusion, create quizzes, evaluate assignments, and provide feedback. Their versatility extends to assisting in understanding complex concepts and offering examples of code in programming languages [69].
- Research assistance. ChatGPT can suggest research ideas and methodologies and provide examples from previous studies. It can enhance inclusivity in research, find relationships between subjects, assist in statistical analysis, and suggest further study extensions [70].
- Writing assistance. ChatGPT can offer feedback on writing, provide suggestions on organization, and help make arguments more compelling [71].

However, the integration of LLMs like ChatGPT in education presents also significant risks [68], for example:

- Quality of prompts. ChatGPT and similar models' efficacy heavily relies on the quality of the prompts provided. The users' ability to frame questions effectively is crucial in obtaining accurate and relevant responses [72].
- Response variability. The quality of responses can vary significantly based on the application domain. If the training dataset for a particular domain lacks depth or breadth, the responses in that domain might not meet the desired standards [73].
- Hallucinations. LLMs tend to generate content that, while appearing authoritative, might be entirely fabricated or unrelated to the query [74]. Such "hallucinations" can mislead users, especially in an educational context where accuracy is paramount.
- Over-reliance on technology. There is a risk of decreased creativity and critical thinking [75] due to over-dependence on ChatGPT.
- Inaccurate or biased Information. ChatGPT's responses may unintentionally perpetuate biases and reinforce stereotypes in its training data [76].
- Lack of human interaction: While ChatGPT can assist, it cannot replace the value of human interaction, which is essential for students' social and emotional development [77].
- Ethical concerns. Issues related to data ownership [78], control, consent, and plagiarism [79] may arise.
- Security concerns [80]. Storing sensitive data on ChatGPT could pose a security risk due to OpenAI has openly stated that conversations with ChatGPT are going to be included in datasets for training future models [81].

Moreover, the use of ChatGPT in education brings forth a set of ethical and societal challenges, especially for educational institutions and decision-makers [68], for example:

- Integrating the GenAI into the educational institutions' Information Technology (IT) government policies. Glitches, server downtime, or compatibility issues can disrupt the teaching and research process. Thus, the institutions must redefine their IT government strategies to integrate AI advances into their technological ecosystems [82].
- Development of ethical codes and the establishment of general guidelines regarding generative AI. Ensuring responsible and ethical practices in its implementation [83].
- Compliance with data regulations. Due to the geographical location of OpenAI's servers, compliance with specific data privacy regulations (for example, in the European Union) is compromised [84], [85].
- Limit the educational institution's dependency on third-party enterprises. Universities should not rely solely on third-party

solutions. They should encourage a collaborative approach, promoting development and adopting open-source, ethical, and secure LLMs [82].

To address these challenges, educators must emphasize critical thinking, promote collaboration, establish clear guidelines for using AI technology, and have backup plans [86].

#### *D. AI Plagiarism, the Elephant in the Room*

The evolving landscape of academic integrity is increasingly challenged by the use of writing essays, documentation analysis and research, and even solving math problems, presenting educators with dilemmas over distinguishing genuine student work from AI-generated content. The core of this issue lies in the sophisticated capabilities of AI, which enable the production of text indistinguishable from human-written essays at minimal cost and effort [87]. This accessibility has magnified concerns over academic dishonesty, previously exacerbated by the internet and platforms facilitating the sharing of completed assignments.

However, ways to circumvent system controls have always been used, for example, by inserting Cyrillic characters that look like letters of the Latin alphabet (see the table of confusing characters at <https://d66z.short.gy/qLwNBx>) and easily circumvent anti-plagiarism systems.

Detection tools like Turnitin, designed to identify plagiarism, are now grappling with the nuances of AI-generated texts, often leading to false positives and negatives. The efficacy of these tools diminishes as AI technology advances, a point underscored by research from the University of Maryland [88], which suggests that detecting AI-generated text reliably may be impossible.

An illustrative case discussed by Robert Topinka [89], a Birkbeck, University of London professor, highlights these challenges. He recounts an instance where a top-performing student contested an accusation of submitting an AI-generated essay, underscoring the limitations of current detection methods and the potential for unjust accusations. The lecturers seem to wish for the easy solution, the infallible judgment of the AI tool that will tell whether the student has cheated with AI. They also seem to lose the irony of it and the fact that research indicates that this infallible judgment is not infallible or even capable of outperforming random classifiers [88].

The situation calls for a fundamental reevaluation of academic assessment methods. Alternatives that prioritize critical thinking and creativity, such as presentations and podcasts, are proposed to adapt to the AI era [90]. These methods ensure fairness and encourage genuine student engagement, moving away from traditional essays vulnerable to AI assistance.

This shift also prompts a broader discussion on the ethical responsibilities of educators in deploying AI detection tools. The reliance on imperfect technology risks harming students' academic careers and reflects a deeper issue of educational values. The drive towards easy solutions for maintaining academic integrity may overshadow the essential goal of education: to cultivate understanding, critical thinking, and innovation among students.

#### *E. Safe AI in Education*

In 2023, we have repeatedly heard and read the words "AI Ethics" [91] and "AI Safety" [92]. We have reached a point where we do not have a common definition, and most people using the terms align it with their agenda. We propose a simple definition of "Safe AI in Education," which is an AI system that is used by students that:

1. **Provides a guarantee of privacy of the students' data and interactions with it.** All the information about the students, their identity, roles, academic records, and interactions with the system

are to be secure and used only to provide the service. We also need guarantees that the information is deleted after the academic course is over;

2. **Is aligned with the teaching strategy.** ChatGPT and other GenAI tools are multi-purpose. It can allow a student to learn, create content, and research, but also to cheat and avoid doing the hard work and learning. Students can ask the system for solutions to their assignments or to paraphrase essays to evade proctoring and anti-plagiarism software;
3. **Provides answers and interactions aligned with a didactic purpose.** If an LLM-powered application is used within the context of a learning activity, we need to be able to bind it under certain parameters. For example, Salman Kahn presented at TED 2023 Kahn Amigo [93], an AI system based on GPT-4 that adapted its behavior to a certain study plan, could act as a Socratic teacher, and was able to present relevant questions to the students to help them progress, instead of providing straightforward answers;
4. **Minimizes the risk of hallucinations or incorrect information.** LLMs are trained with vast amounts of information, and the best ones, like GPT-4 in early 2024, are often correct. However, there is no guarantee that the output is correct and relevant. And there is always the possibility of an AI hallucination. It is a tall order, but a safe AI system needs to maximize the relevance of its answers and minimize its mishaps. The intuition is that this task is much simpler when the application context is smaller than when a chatbot like ChatGPT is open to any conceivable task.
5. **Presents a behavior, values, and usefulness that students and teachers understand. The user experience must clarify what the tool is and is not for.**

#### *F. Smart Learning Applications, a Technological Approach to Safe AI in Education*

The idea of a "Smart Learning Application" emerges as a pivotal innovation rooted in the principles of AI safety and educational integrity. This stems from discussions at the 2023 TEEM conference in Bragança, Portugal, particularly during the Managing Generative AI in Educational Settings session. This idea is conceptualized as an advanced AI educational tool that goes beyond traditional learning applications by integrating with an LMS, such as Moodle, where they are appropriately termed "activities" [94]. In contrast to general educational apps like Kahoot, which do not integrate with the LMSs and therefore fall short of our criteria due to their disconnection from the educational framework, Smart Learning Applications are crafted to function within the specific boundaries of a course. These applications stand out for their capacity to:

- Ensure a secure access. Utilizing the LMS for authentication and authorization, they restrict access to legitimate users.
- Adapt to user roles. The LMS customizes the application's features to match the user's role: teacher, student, or administrator.
- Provide course-specific context. Each application instance is directly associated with a course, enabling a customized educational journey. Smart Learning Applications leverage LLMs via APIs to facilitate features such as on-the-fly content creation and personalized learning trajectories. This strategy boosts interactivity and customization, tackling challenges like guaranteeing content accuracy and adhering to data privacy laws. The goal is to present an educational technology that is more closely aligned with educational objectives, capable of upholding academic integrity, and offering a tailored user experience.

### III. A FINAL REFLECTION

There are reasons for excitement and concern with applying GenAI in education. Yet, we must prevent one from overshadowing the other about the leap in AI, and potentially in its educational application, with ChatGPT as the flagship, necessitates relentless study, design, experimentation, and evaluation. This should be done with caution yet boldness, embracing the new possibilities. Let us discard the notion that technology, being material and mercenary, will ruin an education that is spiritual and selfless [95].

Many of the issues and dangers identified in the educational context have yet to arise due to the emergence of ChatGPT or other similar applications. They already existed, have been approached from various perspectives, and have remained unresolved. However, the potential of these technologies and the effect of their rapid penetration in all realms of society are magnifying some of these issues more than ever before [68].

AI, especially with its ability to create content indistinguishable from human production and interact with users through natural language, represents one of our most socially disruptive technological means. We are just beginning to imagine the possibilities, risks, and challenges that this technology opens up. However, it is essential to recognize that the future we may build on this foundation must be in more than just the hands of technologists. There must be spaces for inter- and transdisciplinary co-creation that ensure the ethical, safe, and inclusive development of a technology that, not so long ago, we would have considered science fiction.

### IV. MONOGRAPH CONTENTS

This International Journal of Interactive Multimedia and Artificial Intelligence monograph about Generative Artificial Intelligence in Education comprises seven research papers.

The first paper is entitled “A cybernetic perspective on generative AI in education: From transmission to coordination” by Dai Griffiths, Enrique Frias-Martinez, Ahmed Tlili and, Daniel Burgos. This work examines the impact of LLMs and GenAI on education, highlighting a lack of clarity in human-machine communication within educational models. It introduces two paradigms: the transmission paradigm, which aligns with traditional educational methods and communication models, and the coordination paradigm, which combines constructivist learning models with a coordination communication model. The authors argue that LLMs disrupt the existing balance between these paradigms by creating a simulacrum of intelligence, challenging the transmission paradigm’s validity. They suggest that adopting the coordination paradigm can help educational institutions understand and utilize GenAI more effectively, urging a shift in educational practices to leverage AI’s capabilities fully.

Lin Tang and Yu-Sheng Su, in their work “Ethical implications and principles of using artificial intelligence models in the classroom: A systematic literature review,” conduct a systematic literature review [96], [97] on the ethical implications and principles of using AI models in classrooms, addressing the need for an ethical framework amidst AI’s growing educational application. By analyzing 32 out of 1,445 publications from 2013 to 2023, the authors identified five main ethical concerns: algorithmic bias, data privacy breaches, opacity, diminished autonomy, and academic dishonesty, with algorithmic bias and privacy issues being the most prevalent. They also outline six ethical principles: fairness, privacy, transparency, accountability, autonomy, and beneficence, emphasizing fairness and privacy as critical. The paper highlights the under-researched areas of autonomy and academic misconduct, urging more in-depth discussions and solutions to ethical issues, clarity on implementing ethical principles, and accurate assessment of AI’s ethical implications in education.

The next paper, “A trustworthy automated short-answer scoring system using a new dataset and hybrid transfer learning method,” by Martinus Maslim, Hei-Chia Wang, Cendra Devayana Putra, and Yulius Denny Prabowo, introduces HTL-ASAS, an advanced automated system for scoring short answers, addressing inconsistencies in manual grading by teachers due to various challenges. Utilizing a hybrid transfer learning approach and a new dataset of student answers (QA-CS), the system demonstrates remarkably high accuracy (99.6%) in evaluating responses from introductory IT courses. This high level of precision suggests HTL-ASAS’s potential as a reliable tool in educational settings, promising to reduce teacher workload and improve assessment consistency.

Juan Izquierdo-Domenech, Jordi Linares-Pellicer, and Isabel Ferri-Molla are the authors of the paper “Virtual reality and language models, a new frontier in learning.” They introduce an innovative learning architecture that combines virtual reality and LLMs with Retrieval-Augmented Generation (RAG) to enhance educational experiences across various settings. This approach integrates immersive virtual reality applications with LLMs, allowing students to interactively engage with learning materials through questions and receive answers with textual and visual hints within a virtual reality environment. The paper addresses the challenge of integrating diverse data sources by utilizing RAG to structure information from APIs, PDFs, Structured Query Language (SQL) databases, and more into formats that are easily processed by LLMs. An empirical study involving twenty participants compared the effectiveness of this virtual reality and LLM architecture against traditional learning methods showed significant improvements in learning outcomes for the group using the immersive virtual reality application. This research highlights the potential of combining virtual reality and LLMs to create dynamic, engaging, and effective learning experiences.

The paper “Generative Artificial Intelligence in product design education: Navigating concerns of originality and ethics,” by Kristin A. Bartlett and Jorge D. Camba, explores the integration of image-generative AI in product design education, addressing the technological advancements and their potential future applications. It critically examines the legal and ethical challenges posed by such technology, including issues of bias, exploitation of hidden labor, intellectual property theft, lack of originality, and inadequate copyright protection. The authors offer recommendations for design educators on incorporating AI responsibly into the curriculum. They advocate for AI to be presented as one of many tools available to designers, emphasizing its role in the creative process rather than as a means to produce final designs. The paper also suggests strategies for fostering meaningful discussions about AI among students, aiming to enrich their understanding and ethical use of AI in design.

Verónica Parra, Patricia Sureda, Ana Corica, Silvia Schiaffino, and Daniela Godoy investigate in their work, “Can generative AI solve geometry problems? Strengths and weaknesses of LLMs for geometric reasoning in Spanish” the potential of GenAI, specifically LLMs like ChatGPT, Bard, and others, in solving geometry problems, a key area in high-school curricula. It highlights the growing interest in using LLMs for educational purposes, especially math problem-solving, and notes the usual focus on English language benchmarks. This study differentiates itself by concentrating on Spanish, a comparatively less-resourced language, to explore LLMs’ capabilities in geometric reasoning. By analyzing the performance of chatbots powered by various LLMs, the study assesses their accuracy in solving geometry problems and categorizes errors in their reasoning processes. The findings aim to understand LLMs’ strengths and weaknesses in geometry, paving the way for better classroom integration strategies and developing more advanced generative AI tools for educational support.



The last paper, entitled “Evaluating ChatGPT-generated linear algebra formative assessments, by Nelly Rigaud Téllez, Patricia Rayón Villela, and Roberto Blanco Bautista, delves into the utilization of LLMs, specifically ChatGPT, for creating formative assessments in linear algebra, focusing on the mathematical problem-solving process. It assesses ChatGPT’s performance in generating feedback on linear algebra problems, highlighting deficiencies in reasoning, proofs, and model construction. By comparing feedback from both instructors and ChatGPT against detailed formative feedback criteria, including affective aspects, the study aims to enhance the feedback quality from both sources. A novel framework for formative assessment using LLMs was developed to generate prompts based on common linear algebra errors, facilitating concept development and problem-solving strategies. This approach encourages a dynamic learning cycle where instructors validate tasks. ChatGPT supports query-based learning, revealing insights into improving feedback for advanced math problems and suggesting adaptations in teaching and learning strategies for educators and students.

#### ACKNOWLEDGMENT

The Ministry of Science and Innovation partially funded this monograph through the AvisSA project grant number (PID2020-118345RB-I00). The Departament de Recerca i Universitats de la Generalitat de Catalunya partially funded this monograph through the 2021 SGR 01412 research groups award.

#### REFERENCES

- [1] V. Mahajan, “100+ Incredible ChatGPT Statistics & Facts in 2024,” 2023. [Online]. <https://bit.ly/48M9fdX>
- [2] P. H. Diamandis and S. Kotler, *Abundance: The Future Is Better Than You Think* (Exponential Technology Series). The Free Press, 2012.
- [3] F. J. García-Peñalvo and A. Vázquez-Ingelmo, “What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 7-16, 2023, doi: 10.9781/ijimai.2023.07.006.
- [4] R. R. Schaller, “Moore’s law: past, present, and future,” *IEEE Spectrum*, vol. 34, no. 6, pp. 52-59, 1997, doi: 10.1109/6.591665.
- [5] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” *Psychological Review*, vol. 65, no. 6, pp. 386-408, 1958, doi: 10.1037/h0042519.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436-444, 2015, doi: 10.1038/nature14539.
- [7] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017*, pp. 5998-6008.
- [8] D. Peral-García, J. Cruz-Benito, and F. J. García-Peñalvo, “Systematic literature review: Quantum machine learning and its applications,” *Computer Science Review*, vol. 51, 2024, Art no. 100619, doi: 10.1016/j.cosrev.2024.100619.
- [9] W. X. Zhao *et al.*, “A Survey of Large Language Models,” *arXiv*, 2023, Art no. arXiv:2303.18223v13, doi: 10.48550/arXiv.2303.18223.
- [10] Y. Chang *et al.*, “A Survey on Evaluation of Large Language Models,” *ACM Transactions on Intelligent Systems and Technology*, vol. In Press, 2024, doi: 10.1145/3641289.
- [11] Z. Yang *et al.*, “The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision),” *arXiv*, 2023, Art no. arXiv:2309.17421v2, doi: 10.48550/arXiv.2309.17421.
- [12] OpenAI, “GPT-4V(ision) System Card,” OpenAI, USA, 2023. [Online]. Available: <https://bit.ly/3TOD21h>
- [13] S. Yin, S. Bai, J. Huang, X. Chen, and Y. Zhang, “A Survey on Multimodal Large Language Models,” *arXiv*, 2023, Art no. arXiv:2306.13549v1, doi: 10.48550/arXiv.2306.13549.
- [14] S. Bubeck *et al.*, “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” *arXiv*, 2023, Art no. arXiv:2303.12712v5, doi: 10.48550/arXiv.2303.12712.
- [15] V. Elliott, “Generative AI Learned Nothing From Web 2.0,” *Wired*, December 28, 2023. [Online]. Available: <https://d66z.short.gy/BZ57XL>
- [16] J. M. Flores-Vivar and F. J. García-Peñalvo, “Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4),” *Comunicar*, vol. 31, no. 74, pp. 35-44, 2023, doi: 10.3916/C74-2023-03.
- [17] S. Altman, G. Brockman, and I. Sutskever, “Governance of superintelligence,” In: OpenAI, 2023. Available from: <https://bit.ly/3q6NFjv>.
- [18] F. A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, “Diffusion Models in Vision: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850-10869, 2023, doi: 10.1109/TPAMI.2023.3261988.
- [19] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” *arXiv*, 2020, Art no. arXiv:2005.14165v4 doi: 10.48550/arXiv.2005.14165.
- [20] OpenAI, “GPT-4 Technical Report,” *arXiv*, 2023, Art no. arXiv:2303.08774v4, doi: 10.48550/arXiv.2303.08774.
- [21] D. Patel and G. Wong, “GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. Demystifying GPT-4: The engineering tradeoffs that led OpenAI to their architecture,” 2023. [Online]. <https://bit.ly/3SbiU8r>
- [22] T. Harris, “AI and The Future of Life,” 2023. [Online]. <https://d66z.short.gy/LhQinU>
- [23] C. Fourrier, “2023, year of open LLMs,” In: Hugging Face, 2023. [Online]. Available from: <https://d66z.short.gy/vcHitF>.
- [24] S. R. Bowman, “Eight Things to Know about Large Language Models,” *arXiv*, 2023, Art no. arXiv:2304.00612v1 doi: 10.48550/arXiv.2304.00612.
- [25] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding by Generative Pre-Training,” OpenAI, USA, 2018. [Online]. Available: <https://d66z.short.gy/OHRedH>
- [26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners,” OpenAI, USA, 2019. [Online]. Available: <https://bit.ly/3Mq72Lz>
- [27] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbahn, and P. Villalobos, “Compute Trends Across Three Eras of Machine Learning,” in *2022 International Joint Conference on Neural Networks (IJCNN) (18-23 July 2022, Padua, Italy)*, 2022. doi: 10.1109/IJCNN55064.2022.9891914.
- [28] J. Wei *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” *arXiv*, 2023, Art no. arXiv:2201.11903v6, doi: 10.48550/arXiv.2201.11903.
- [29] M. Nye *et al.*, “Show Your Work: Scratchpads for Intermediate Computation with Language Models,” *arXiv*, 2021, Art no. arXiv:2112.00114v1, doi: 10.48550/arXiv.2112.00114.
- [30] J. Wei *et al.*, “Emergent Abilities of Large Language Models,” *arXiv*, 2022, Art no. arXiv:2206.07682v2 doi: 10.48550/arXiv.2206.07682.
- [31] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large Language Models are Zero-Shot Reasoners,” *arXiv*, 2023, Art no. arXiv:2205.11916v4 doi: 10.48550/arXiv.2205.11916.
- [32] J. Zhou *et al.*, “Instruction-Following Evaluation for Large Language Models,” *arXiv*, 2023, Art no. arXiv:2311.07911v1, doi: 10.48550/arXiv.2311.07911.
- [33] A. Srivastava *et al.*, “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models,” *arXiv*, 2023, Art no. arXiv:2206.04615v3 doi: 10.48550/arXiv.2206.04615.
- [34] M. Abdou, A. Kulmizev, D. Hershovich, S. Frank, E. Pavlick, and A. Søgaard, “Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color,” in *Proceedings of the 25th Conference on Computational Natural Language Learning*, A. Bisazza and O. Abend Eds.: Association for Computational Linguistics, 2021, pp. 109-132. doi: 10.18653/v1/2021.conll-1.9.
- [35] J. Andreas, “Language Models as Agent Models,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang Eds.: Association for Computational Linguistics, 2022, pp. 5769-5779. doi: 10.18653/v1/2022.findings-emnlp.423.
- [36] B. Z. Li, M. Nye, and J. Andreas, “Implicit Representations of Meaning in Neural Language Models,” *arXiv*, 2021, Art no. arXiv:2106.00737v1 doi: 10.48550/arXiv.2106.00737.
- [37] H. Levesque, E. Davis, L. Morgenstern, and T. i. c. o. t. p. o. k. r. a. reasoning.,

- "The winograd schema challenge," in *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*: Association for the Advancement of Artificial Intelligence, 2012, pp. 552-561.
- [38] J. Hoffmann *et al.*, "Training Compute-Optimal Large Language Models," *arXiv*, 2022, Art no. arXiv:2203.15556v1, doi: 10.48550/arXiv.2203.15556.
- [39] R. Taori *et al.*, "Alpaca: A Strong, Replicable Instruction-Following Model," Stanford University, USA, 2023. [Online]. Available: <https://bit.ly/444TrRx>
- [40] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv*, 2023, Art no. arXiv:2302.13971v1, doi: 10.48550/arXiv.2302.13971.
- [41] S. Willinson, "Stanford Alpaca, and the Acceleration of On-Device Large Language Model Development," In: Simon Willinson's Blog, 2023. [Online]. Available from: <https://bit.ly/3r3ahlN>.
- [42] L. Zheng *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *arXiv*, 2023, Art no. arXiv:2306.05685v4 doi: 10.48550/arXiv.2306.05685.
- [43] S. Pichai and D. Hassabis, "Our next-generation model: Gemini 1.5," In: AI. 2024. [Online]. Available from: <https://d66z.short.gy/cT1911>.
- [44] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv*, 2021, Art no. arXiv:2106.09685v2, doi: 10.48550/arXiv.2106.09685.
- [45] T. Wang, J. Zhao, and W. V. Eaton, "LoRA Land: Fine-Tuned Open-Source LLMs that Outperform GPT-4," In: *Predibase*, 2024. [Online]. Available from: <https://d66z.short.gy/4ec8RY>.
- [46] N. Stephenson, *The Diamond Age: Or, A Young Lady's Illustrated Primer*. Bantam Books, 1995.
- [47] N. Stephenson, *Snowcrash*. Bantam Books, 1992.
- [48] B. Srinivasan, *The Network State: How To Start a New Country*. USA: Amazon Kindle, 2022.
- [49] P. H. Diamandis and S. Kotler, *The Future Is Faster Than You Think: How Converging Technologies Are Transforming Business, Industries, and Our Lives*. Simon & Schuster, 2020.
- [50] F. J. García-Peñalvo and J. García Carrasco, "Educational hypermedia resources facilitator," *Computers & Education*, vol. 44, no. 3, pp. 301-325, 2005, doi: 10.1016/j.compedu.2004.02.004.
- [51] F. J. García-Peñalvo and A. M. Seoane-Pardo, "An updated review of the concept of eLearning. Tenth anniversary," *Education in the Knowledge Society*, vol. 16, no. 1, pp. 119-144, 2015, doi: 10.14201/eks201516119144.
- [52] M. Sharples, M. Milrad, I. Arnedillo, and G. Vavoula, "Mobile Learning: Small devices, Big Issues," in *Technology Enhanced Learning: Principles and Products* N. Balacheff, S. Ludvigsen, T. d. Jong, A. Lazonder, and S. Barnes Eds. Heidelberg: Springer, 2009, pp. 233-249.
- [53] C. J. Villagrà-Arnedo, F. J. Gallego-Durán, F. Llorens-Largo, R. Satorre-Cuerda, P. Compañ-Rosique, and R. Molina-Carmona, "Time-Dependent Performance Prediction System for Early Insight in Learning Trends," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 112-124, 2020, doi: 10.9781/ijimai.2020.05.006.
- [54] S. C. Cifuentes, S. G. García, M. P. Andrés-Sebastiá, J. D. Camba, and M. Contero, "Augmented Reality Experiences in Therapeutic Pedagogy: A Study with Special Needs Students," in *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT) (Austin, TX, USA, 25-28 July 2016)*. USA: IEEE, 2016, pp. 431-435. doi: 10.1109/ICALT.2016.23.
- [55] T. A. Urdan and C. C. Weggen, *Corporate e-learning: Exploring a new frontier*. San Francisco, USA: WR Hambrecht, 2000.
- [56] A. Anand and S. A. C. S. S. Eswaran, "A survey of open source learning management systems," *Annals: Computer Science Series*, vol. 16, no. 1, pp. 185-188, 2018.
- [57] C. López, F. García-Peñalvo, and P. Pernías, "Desarrollo de Repositorios de Objetos de Aprendizaje a través de la Reutilización de los Metadatos de una Colección Digital: De Dublin Core a IMS," *RED. Revista de Educación a Distancia*, vol. IV, no. monográfico II, 2005. [Online]. Available: <http://www.um.es/ead/red/M2>.
- [58] L. Castañeda and N. Selwyn, "More than tools? Making sense of the ongoing digitizations of higher education," *International Journal of Educational Technology in Higher Education*, vol. 15, no. 1, p. 22, 2018, doi: 10.1186/s41239-018-0109-y.
- [59] N. Selwyn, "Minding our language: why education and technology is full of bullshit...and what might be done about it?," *Learning, Media and Technology*, vol. 41, no. 3, pp. 437-443, 2016, doi: 10.1080/17439884.2015.1012523.
- [60] A. Bartolomé, L. Castañeda, and J. Adell, "Personalisation in educational technology: the absence of underlying pedagogies," *International Journal of Educational Technology in Higher Education*, vol. 15, no. 1, 2018, Art no. 14, doi: 10.1186/s41239-018-0095-0.
- [61] S. Papert, "Information Technology and Education: Computer Criticism vs. Technocentric Thinking," *Educational Researcher*, vol. 16, no. 1, pp. 22-30, 1987, doi: 10.3102/0013189X016001022.
- [62] A. M. Seoane-Pardo, "Formalización de un modelo de formación online basado en el factor humano y la presencia docente mediante un lenguaje de patrón," PhD, Programa de Doctorado en Formación en la Sociedad del Conocimiento, Universidad de Salamanca, Salamanca, Spain, 2014. [Online]. Available: <https://goo.gl/sNrkHu>
- [63] J. de Souza Zanirato Maia, A. P. Arantes Bueno, and J. R. Sato, "Applications of Artificial Intelligence Models in Educational Analytics and Decision Making: A Systematic Review," *World*, vol. 4, no. 2, pp. 288-313, 2023, doi: 10.3390/world4020019.
- [64] D. Amo-Filva, D. Fonseca, F. J. García-Peñalvo, M. Alier-Forment, and M. J. Casany-Guerrero, "Learning Analytics' Privacy in the Fog and Edge Computing: A Systematic Mapping Review," in *Proceedings TEEM 2022: Tenth International Conference on Technological Ecosystems for Enhancing Multiculturality. Salamanca, Spain, October 19-21, 2022*, F. J. García-Peñalvo and A. García-Holgado Eds. Singapore: Springer Nature, 2023, pp. 1199-1207. doi: 10.1007/978-981-99-0942-1\_126.
- [65] D. Shin, "Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm," *Journal of Information Science*, vol. 49, no. 1, pp. 18-31, 2021, doi: 10.1177/0165551520985495.
- [66] T. Wang and E. C. K. Cheng, "An investigation of barriers to Hong Kong K-12 schools incorporating Artificial Intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 2, 2021, Art no. 100031, doi: 10.1016/j.caeai.2021.100031.
- [67] M. Alier-Forment and F. Llorens-Largo, "Cabalga el Cometa," in EP-31 Las Alucinaciones de ChatGPT con Faraón Llorens, 2023. [Online]. <https://bit.ly/3ZCNBVT>
- [68] F. J. García-Peñalvo, F. Llorens-Largo, and J. Vidal, "The new reality of education in the face of advances in generative artificial intelligence," *RIED: Revista Iberoamericana de Educación a Distancia*, vol. 27, no. 1, pp. 9-39, 2024, doi: 10.5944/ried.27.1.37716.
- [69] F. J. García-Peñalvo, "The perception of Artificial Intelligence in educational contexts after the launch of ChatGPT: Disruption or Panic?," *Education in the Knowledge Society*, vol. 24, 2023, Art no. e31279, doi: 10.14201/eks.31279.
- [70] A. Bahrini *et al.*, "ChatGPT: Applications, Opportunities, and Threats," *arXiv*, 2023, Art no. arXiv:2304.09103v1, doi: 10.48550/arXiv.2304.09103.
- [71] J. Crawford, M. Cowling, and K. A. Allen, "Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI)," *Journal of University Teaching and Learning Practice*, vol. 20, no. 3, 2023, doi: 10.53761/1.20.3.02.
- [72] L. Henrickson and A. Meroño-Peñuela, "Prompting meaning: a hermeneutic approach to optimising prompt engineering with ChatGPT," *AI & SOCIETY*, vol. In Press, 2023, doi: 10.1007/s00146-023-01752-8.
- [73] A. Nazir and Z. Wang, "A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges," *Meta-Radiology*, vol. 1, no. 2, 2023, Art no. 100022, doi: 10.1016/j.metrad.2023.100022.
- [74] H. H. Thorp, "ChatGPT is fun, but not an author," *Science*, vol. 379, no. 6630, p. 313, 2023, doi: 10.1126/science.adg7879.
- [75] Y. K. Dwivedi *et al.*, "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *International Journal of Information Management*, vol. 71, 2023, Art no. 102642, doi: 10.1016/j.ijinfomgt.2023.102642.
- [76] A. Iskender, "Holy or Unholy? Interview with Open AI's ChatGPT," *European Journal of Tourism Research*, vol. 34, 2023, Art no. 3414, doi: 10.54055/ejtr.v34i.3169.
- [77] E. P. H. Choi, J. J. Lee, M. H. Ho, J. Y. Y. Kwok, and K. Y. W. Lok, "Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education," *Nurse Education Today*, vol. 125, 2023, Art no. 105796, doi: 10.1016/j.nedt.2023.105796.
- [78] D. Gašević, G. Siemens, and S. Sadiq, "Empowering learners for the age

of artificial intelligence,” *Computers and Education: Artificial Intelligence*, vol. 4, 2023, Art no. 100130, doi: 10.1016/j.caeai.2023.100130.

- [79] D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, “Chatting and cheating: Ensuring academic integrity in the era of ChatGPT,” *Innovations in Education and Teaching International*, vol. In Press, 2023, doi: 10.1080/14703297.2023.2190148.
- [80] H. Lee, “The rise of ChatGPT: Exploring its potential in medical education,” *Anatomical Sciences Education*, vol. In Press, 2023, doi: 10.1002/ase.2270.
- [81] OpenAI, “Privacy policy” In: OpenAI, 2023. [Online]. Available from: <https://d66z.short.gy/XyWACH>.
- [82] F. Llorens-Largo and F. J. García-Peñalvo, “La inteligencia artificial en el gobierno universitario,” In: Universidad, 2023. [Online]. Available from: <https://bit.ly/46SSxbG>.
- [83] W. M. Lim, A. Gunasekara, J. L. Pallant, J. I. Pallant, and E. Pechenkina, “Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators,” *International Journal of Management Education*, vol. 21, no. 2, 2023, Art no. 100790, doi: 10.1016/j.ijme.2023.100790.
- [84] T. Wang, Y. Zhang, S. Qi, R. Zhao, Z. Xia, and J. Weng, “Security and Privacy on Generative Data in AIGC: A Survey,” *arXiv*, 2023, Art no. arXiv:2309.09435v2, doi: 10.48550/arXiv.2309.09435.
- [85] M. Alier, M. J. Casañ Guerrero, D. Amo, C. Severance, and D. Fonseca, “Privacy and E-Learning: A Pending Task,” *Sustainability*, vol. 13, no. 16, doi: 10.3390/su13169206.
- [86] E. Sabzalieva and A. Valentini, “ChatGPT and artificial intelligence in higher education: Quick start guide,” UNESCO and UNESCO International Institute for Higher Education in Latin America and the Caribbean (IESALC), Paris, France; Caracas, Venezuela, ED/HE/IESALC/IP/2023/12, 2023. [Online]. Available: <https://bit.ly/3oeYm2f>
- [87] F. Llorens-Largo, J. Vidal, and F. J. García-Peñalvo, “Ya llegó, ya está aquí, y nadie puede esconderse: La inteligencia artificial generativa en educación,” In: Aula Magna 2.0, 2023. [Online]. Available from: <https://bit.ly/3tcq5Uh>.
- [88] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can AI-Generated Text be Reliably Detected?,” *arXiv*, 2024, Art no. arXiv:2303.11156v3, doi: 10.48550/arXiv.2303.11156.
- [89] R. Topinka, “The software says my student cheated using AI. They say they’re innocent. Who do I believe?,” *The Guardian*, February 13th. [Online]. Available: <https://d66z.short.gy/T3vBN5>
- [90] F. J. García-Peñalvo, “Cómo afecta la inteligencia artificial generativa a los procesos de evaluación,” *Cuadernos de Pedagogía*, no. 549, 2024.
- [91] UNESCO, “Recommendation on the Ethics of Artificial Intelligence,” UNESCO, Paris, France, 2022. [Online]. Available: <https://bit.ly/40MCNna>
- [92] Y. Hu *et al.*, “Artificial Intelligence Security: Threats and Countermeasures,” *ACM Computing Surveys*, vol. 55, no. 1, 2021, doi: 10.1145/3487890.
- [93] S. Khan, “How AI could save (not destroy) education,” In: TED2023, 2023. [Online]. Available from: <https://d66z.short.gy/gp6isK>.
- [94] M. Alier, M. J. Casañ, and D. Amo, “Smart Learning Applications: Leveraging LLMs for Contextualized and Ethical Educational Technology,” in *Proceedings TEEM 2023: Eleventh International Conference on Technological Ecosystems for Enhancing Multiculturality. Bragança, Portugal, October 25–27, 2023*: Springer, 2024.
- [95] M. Fernández Enguita, “Inteligencia aumentada y avanzada para aprender y enseñar,” *Cuadernos de Pedagogía*, no. 549, 2024.
- [96] A. García-Holgado, S. Marcos-Pablos, and F. J. García-Peñalvo, “Guidelines for performing Systematic Research Projects Reviews,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 136-144, 2020, doi: 10.9781/ijimai.2020.05.005.
- [97] F. J. García-Peñalvo, “Developing robust state-of-the-art reports: Systematic Literature Reviews,” *Education in the Knowledge Society*, vol. 23, 2022, Art no. e28600, doi: 10.14201/eks.28600.



Marc Alier

He is an associate professor at the Polytechnic University of Catalonia (UPC), passionate about e-learning, project management, and computing ethics. Marc holds an engineering degree in computer science and a Ph.D. in Sustainability from UPC. With over 25 years of experience in research and development related to e-learning, he has designed and developed several Learning Management Systems (LMS) and educational content authoring tools. Additionally, Marc has been an online instructor and creator of MOOCs. Marc has been a pioneer in developing and contributing to the Moodle community since 2004, designing and implementing core functionalities such as the Wiki module, the Webservices layer, and the IMS LTI consumer. He has also been the academic director of UPC’s Ph.D. program in Education in Engineering Sciences and Technology. He has authored over 160 academic publications in journals and conferences. He has authored nonacademic books at <http://pansingluten.net> and <http://aprendizdeluthier.com>. Additionally, he is a prolific podcaster at <http://mossegalapoma.cat>, <https://cabalgaelcometa.com>, and <http://zetatesters.com>. Marc has also taught project management, information systems, and computing ethics at universities since 2001 and has been the director of a master’s program in software for organization management and several post-degree courses at UPC School.



Francisco José García-Peñalvo

He received degrees in computing from the University of Salamanca and the University of Valladolid and a Ph.D. from the University of Salamanca (USAL). He is a Full Professor of the Computer Science Department at the University of Salamanca. In addition, he is a Distinguished Professor of the School of Humanities and Education of the Tecnológico de Monterrey, Mexico. Since 2006, he has been the head of the GRIAL Research Group GRIAL. He is head of the Consolidated Research Unit of the Junta de Castilla y León (UIC 81). He was Vice-dean of Innovation and New Technologies of the Faculty of Sciences of the USAL between 2004 and 2007 and Vice-Chancellor of Technological Innovation of this University between 2007 and 2009. He is currently the Coordinator of the Ph.D. Programme in Education in the Knowledge Society at USAL. He is a member of IEEE (Education Society and Computer Society) and ACM.



Jorge D. Camba

He is an Associate Professor in the School of Engineering Technology and the Department of Computer Graphics Technology (by courtesy) at Purdue University in West Lafayette, IN (USA). Before joining Purdue, he was a faculty member at the University of Houston and the Dwight Look College of Engineering at Texas A&M University. Dr. Camba is also a Senior Research Scientist at the Department of Industrial Engineering at the University of Naples Federico II in Naples, Italy, and an I3BH Fellow at the Institute for Research and Innovation in Bioengineering in Valencia, Spain. Dr. Camba is the author of more than 100 peer-reviewed publications and eight books. His research focuses on parametric solid modeling complexity, design automation and reusability, and product lifecycle management.



# A Cybernetic Perspective on Generative AI in Education: From Transmission to Coordination

Dai Griffiths<sup>1\*</sup>, Enrique Frías-Martínez<sup>1</sup>, Ahmed Tlili<sup>2</sup>, Daniel Burgos<sup>1</sup>

<sup>1</sup> Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR) (Spain)

<sup>2</sup> Smart Learning Institute (SLI), Beijing Normal University (BNU) (China)

Received 18 November 2023 | Accepted 12 February 2024 | Published 21 February 2024



## ABSTRACT

The recent sudden increase in the capabilities of Large Language Models (LLMs), and generative AI in general, has astonished education professionals and learners. In formulating a response to these developments, educational institutions are constrained by a lack of clarity concerning human-machine communication and its relationship to models of education. Ideas and models from the cybernetic tradition can help to fill this gap. Two paradigms are distinguished: (1) the transmission paradigm (combining the model of learning implied by the instruments and processes of formal education and the conduit model of communication), and (2) the coordination paradigm (combining the constructivist model of learning and the coordination model of communication). It is proposed that these paradigms have long coexisted in educational practice in a *modus vivendi*, which is disrupted by LLMs. If an LLM can pass an examination, then from within the transmission paradigm this can only be understood as demonstrating that the LLM has indeed learned and understood the material being assessed. At the same time, we know that LLMs do not in fact have the capacity to learn and understand, but rather generate a simulacrum of intelligence. It is argued that this paradox prevents educational institutions from formulating a coherent response to generative AI systems. However, within the coordination paradigm the interactions of LLMs and education institutions can be more easily understood and can be situated in a conversational model of learning. These distinctions can help institutions, educational leaders, and teachers, to frame the complex and nuanced questions raised by GenAI, and to chart a course towards its effective use in education. More specifically, they indicate that to benefit fully from the capabilities of generative AI education institutions need to recognize the validity of the coordination paradigm and adapt their processes and instruments accordingly.

## KEYWORDS

Education, Cybernetics, Generative AI, Human-Machine Communication, Large Language Model (LLM), Machine Learning.

DOI: 10.9781/ijimai.2024.02.008

## I. INTRODUCTION

**T**HE recent sudden increase in the capabilities of Large Language Models (LLMs) and other generative artificial intelligence (GenAI) applications has astonished education professionals and students. A wide-ranging debate has emerged concerning the immediate and future impact of these developments on educational institutions and practice, focusing on topics such as assessment, the role of the teacher, the opportunities for students, and the implications for institutions.

The present paper contributes to the clarification of this discourse in the context of formal education. The core activity of education is communication between humans, often mediated by texts and other media, in conversations between actors that include students, teachers, administrators and policymakers. It is therefore hard to achieve clarity in the understanding of the impact of AI on education without a clear understanding of the nature of human-machine communication. The present lack of consensus on how GenAI could or should be used in

education, and whether its use is constructive or destructive, suggests that this understanding remains problematic. This paper proposes a historical perspective on thinking about models of communication and learning, largely associated with the cybernetic tradition, which has renewed relevance in helping to navigate the complex terrain presented by generative AI and education. We summarize the conclusion of each section in a brief text in italics, to provide an overview of our argument. We commence with a brief review of the technology under discussion.

## II. STATE OF THE ART

Generative modeling, also known as GenAI or generative AI, leverages unsupervised learning techniques such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) to discern patterns in various types of content, ranging from text and images to video. By doing so, it gains the ability to create new content that mirrors these identified patterns. Within text, this technology manifests as Language Models (LMs) and their extensive counterparts, Large Language Models (LLMs). The primary distinctions between these two lie in the scale of data used for training – LMs typically

\* Corresponding author.

E-mail address: david.griffiths@unir.net

utilize smaller, domain-specific datasets, whereas LLMs draw from vastly larger data pools – and their respective use cases, with LMs being more suited for tasks like text prediction and spell checking, and LLMs being designed for text generation.

In this context, transformers, and in particular generative pre-trained transformers (GPT), are the de facto standard to implement LLMs. GPT uses large amounts of text data to create a generative model that captures and replicates the structure of a phrase. As a result, LLMs can process and produce human-like text outputs and open the door to a variety of educational applications.

In this brief review we focus on the most recent applications of LLMs in education. García-Peñalvo and Vázquez-Ingelmo [1] characterize the generative AI landscape, while Zhao *et al.* [2] provide a survey of the underlying technology of large language models. A number of overviews of applications and limitations in educational settings are available [3] [4] [5] [6]. Table I presents examples of generative AI applied to education published in 2023, focusing on the LLM used and the input data. Although LLMs have the potential to be applied to any area of knowledge, applications to date have tended to focus on specific areas like coding and math. We identify five main applications, namely: (1) Automatic Grading; (2) Exam Solution; (3) Educational Content Generation (including tests); (4) Plagiarism Detection; and (5) Tutoring. Other reviews have proposed a higher number of groups in the classification [5] [3].

TABLE I. GENERATIVE AI APPLICATIONS FOR EDUCATION, INCLUDING APPLICATION, LLM MODEL USED AND DATA

Ref.	Application	LLM	Data
[7]	Grading	OpenAI GPT-3	Computer Sci. Exams
[8]	Grading	OpenAI GPT-4	Questionnaire
[9]	Exam Solution	OpenAI GPT-4	Law Exams
[10]	Plagiarism	OpenAI GPT-3	Math Exams
[11]	Tutoring	OpenAI GPT-3	Math Exercises
[12]	Tutoring	OpenAI GPT-3	Word Vocabulary
[13]	Tutoring	GPTeach	Course data
[14]	Test Generation	OpenAI GPT-4	Questionnaire
[15]	Exam Solution	OpenAI GPT-3.5	High School Exams
[16]	Plagiarism	OpenAI GPT-3.5	Human/GPT Texts
[17]	Content Generation	N/A	Python Code
[18]	Content Generation	OpenAI Codex	Python Code
[19]	Exam Solution	OpenAI GPT-4	Medical Exam
[20]	Exam Solution	OpenAI GPT-4	Physics Exam
[21]	Exam Solution	OpenAI GPT-3.5	Medical Exam
[22]	Tutoring	OpenAI GPT 3	Questionnaire

Although there is a variety of LLMs available, both commercial (OpenAI ChatGPT, Bing, etc.), and open source (Llama, Llama2, BLOOM, Alpaca, PaLM2, Bert and its variations, DeepMind Gopher, etc.), ChatGPT has become the standard to implement educational research studies. The use of other LLMs in recent studies seems to be residual. From the examples presented in Table I only one work [13] proposes its own LLM, called GPTeach, but even in this case it uses ChatGPT-3 API for solving questions. This contrasts with the results observed in another study [5] where Bert (and its variants) was used in almost 90% of the studies up to 2022, and ChatGPT in all its versions was used marginally. As a result, the study claims that the most advanced LLMs models have not been the focus of educational tasks. This is not the outcome of other review papers [3] [4] that conclude that ChatGPT-3, which at the time was the most advanced, was being widely used for educational applications. In any event, Table I indicates that current studies are based on the most advanced LLMs implementations such as ChatGPT 3.5 and ChatGPT4.

Current educational applications are largely built using commercial LLMs, although there is a wide range of open source LLMs. This is mainly because of the complexities and cost of training open source LLMs. Commercial implementations (e.g. GPT-4) have been already fine-tuned for conversation (e.g. ChatGPT-4), and the use of this foundation makes it possible to focus directly on the relevant research questions and possible applications. This approach has its drawbacks as there is no control or detailed knowledge over the data used to train the LLM. In contrast a fully open model opens the door to difficult questions about the legality and quality of sources. Fine-tuning of pre-trained LLMs with smaller and more specific datasets that are adapted to a particular domain is less problematic and enables the personalization of learning materials. Following this approach ChatGPT-4 is already being deployed in learning applications such as DuoLingo [23] for learning languages or Khan Academy [24] for personalized learning.

LLMs have the potential to affect the whole educational community, but the papers we have examined show that the focus to date has been on educators/teachers and students. The use of LLMs for other stakeholders such as academic administrators or policy makers seems to be residual, or undocumented, and still needs to be explored. However, the rapid transformation of workplaces through the application of AI [25] raises many open questions about the future of academic management and leadership. Concerns have also been raised about bias in AI applications [26], and their compliance with ethical standards [27], raising a large number of additional open questions.

Table I highlights the recent impact and potential of LLMs (mainly of ChatGPT) for educational applications. Nevertheless, there are many concerns and limitations including: (1) data privacy, (2) bias of generated content (especially regarding the language used [26]); and mainly (3) the potential impact on educational practice. Most of the studies do not evaluate the impact of the application in an educational setting. There are some exemptions, mainly when the application is exam solution as it can be directly compared with previous results [9] [19] [20], and in some cases content generation, for example a study [17] that concludes that the perceived quality of AI-generated resources is largely on par with student-generated resources. In general, it is difficult to ascertain the actual benefits and limitations of the five application areas identified in pedagogical settings.

A key thread of research concerns the degree to which LLMs improve the engagement of students in the learning process, hypothesized improvements comprehension, retention, and overall academic success [28]. Progress on this topic requires not only rigorous experiments, but also increased clarity on the nature of human-machine communication and its implications for education, which is the issue we address in this paper. It can be seen that much valuable research is being carried out into the use of LLMs in education, but this work tends to focus on the results of introducing the technology into an educational activity, without examining the processes involved in human-machine communication. Moreover, we note that the reviewed studies investigate the integration of LLMs in education without using theories of communication and learning as a backbone of their research. This makes it hard to compare like with like, or to cumulate research results.

At the heart of education is the interaction between students, teachers and learning resources. It is therefore unsurprising that a lack of clarity about what is happening during communication between humans and machines generates uncertainty among education professionals and institutions about the position they should adopt when faced by GenAI in general and LLMs in particular. In the discussion below we propose some theoretical tools which can assist in inspection and analysis of interactions between GenAI and

educational actors, and which provide a framework within which educational policy can be formulated.

*There is huge interest in the potential of GenAI in education, and an extensive body of evidence. However, there is a lack of clarity on the nature of the educational interactions which GenAI supports.*

### III. TWO VIEWS OF LEARNING

In most formal education, the design of organizational processes assumes that knowledge can be delivered by a teacher or an institution to a student. This assumption is embedded at all levels of the education system: in national plans, curricula, quality assurance processes, teaching plans, and not least in the fees charged for access to courses. A particularly clear example is the field of knowledge management, which is built on the ideas of capture and delivery of knowledge (see Girard and Girard [29] for an overview).

In contrast, the practice of teaching has been strongly influenced by the constructivist theory of learning. We cannot here provide a detailed account of the many ways in which constructivism has been conceptualized, applied, and critiqued, but the following examples indicate its scope. In his influential ‘Radical Constructivism: A Way of Knowing and Learning’, von Glasersfeld [30] starts his discussion with the sceptics of ancient Greece, but more conventionally the tradition is traced back to Vygotsky and Piaget, with further development being carried out by a host of psychologists, philosophers and educationalists, including Jerome Bruner, Paolo Freire, Seymour Papert and Gert Biesta. The last of these has written that:

The founding intuition of constructivism is that knowing and learning are processes in which knowers and learners actively construct their knowledge and understanding – they make sense – rather than that this should be understood as a process where knowers or learners passively receive such knowledge and understanding. [31]

Constructivism is a theory of learning, and it has been accompanied by theories of pedagogy, notably those known as Learner Centered Pedagogy, which has been widely influential among teachers. Bremner, Sakata and Cameron recently conducted a systematic review of the outcomes of Learner Centered Pedagogy (LCP) [32] which concludes that “there is a real gap in hard data to prove or disprove the value of LCP”, while teachers and students “lean towards positive experiences of LCP”.

Individual teachers and theorists may be convinced constructivists or may vehemently oppose constructivist ideas. At the level of the education system, however, the two contradictory views have cohabited for half a century. On the one hand, the organizational instruments of the education system (such as curricula, learning objectives and lesson plans) assume that it is possible to prescribe what students will learn, how they will learn it, and how long this will take. On the other hand, many teachers are strongly influenced by a belief that the characteristics, prior experiences and activities of students determine what they learn and how fast they learn it, with profound consequences for their classroom teaching practice and informal interactions with students. These two contradictory positions have resolved to a *modus vivendi* which enables educational activities to proceed smoothly. Part of the explanation for this coexistence is that the two theories of learning do not generate mutually exclusive classroom activities. In this context, Richardson points out that “students also make meaning from activities encountered in a transmission model of teaching such as lectures or direct instruction, or even from non-interactive media such as television”. As a result, the coexistence of the two models of education is often not commented upon, or even not perceived. The balance between the two varies from

one place to another and adjusts over time, responding to changing patterns of teaching practice and to the shifting winds of political and social pressures. In the following sections, we discuss two models of communication which are compatible with the two conceptions. These are not the only two available models of communication, and they have nothing to say about the emotional or dialectic aspects of communication. However, we argue that they are of great utility in understanding the communication between humans and machines.

*The transmission and constructivist models of learning coexist in educational practice, in a long-standing modus vivendi.*

### IV. TWO MODELS OF COMMUNICATION IN EDUCATION

#### A. The Transmission Model of Communication

The conception of the communication of knowledge underlying the organizational structures of education has close parallels to Shannon’s mathematical model of the transmission of information (Fig. 1), which was published in 1948, but nevertheless remains a cornerstone of the teaching of telecommunications. Weaver, who collaborated closely with Shannon, explicitly stated that “...information must not be confused with meaning. In fact, two messages, one of which is heavily loaded with meaning and the other of which is pure nonsense, can be exactly equivalent, from the present viewpoint, as regards information.” Nevertheless, there has been confusion about the relationship between information and meaning implied by the theory since its formulation. Indeed, misinterpretation is hard to avoid given the lack of precision in English vocabulary. For example, as Reddy pointed out, the word ‘message’ used in Shannon’s model is ambivalent in English, referring to both the means of communication “I got your message (MESSAGE1) but had no time to read it” and also the understanding of the recipient “Okay, John, I get the message (MESSAGE2); let’s leave him alone” [33].

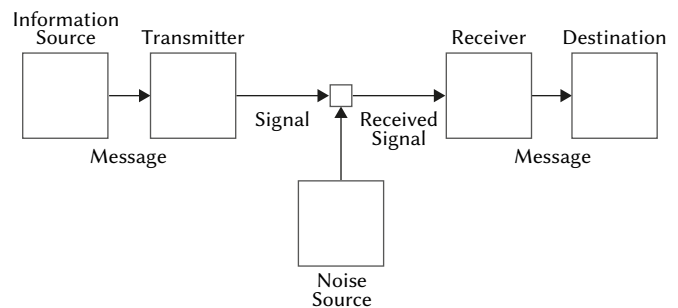


Fig. 1. Shannon’s “Schematic diagram of a general communication system”, adapted from [34].

In 1979, Reddy characterized the merging of these two meanings of ‘message’ as the *conduit metaphor*, which sees language in the following terms: “(1) language functions like a conduit, transferring thoughts bodily from one person to another; (2) in writing and speaking, people insert their thoughts or feelings in the words; (3) words accomplish the transfer by containing the thoughts or feelings and conveying them to others; (4) in listening or reading, people extract the thoughts and feelings once again from the words.” There are, of course, other metaphors for communication. Krippendorff [35] (p.51-70) distinguishes five metaphors for communication in addition to the conduit metaphor: hydraulic, control, transmission, war, and dance-ritual. The first three of these, however, are largely compatible with the conduit metaphor. Moreover, for our present purposes, the importance of the conduit metaphor is that it maps closely onto the aspiration of education to deliver knowledge to the student, and the implied assumptions of its organizational processes.



Pace Weaver, the Shannon / Weaver diagram can be used (or abused) to represent the conduit metaphor, and mapped onto educational processes, as shown in Table II.

TABLE II. THE CONDUIT MODEL MAPPED TO EDUCATION

<b>Information source</b>	A teacher, a video recording, an author, a work of literature, a data set, etc.
<b>Transmitter</b>	A teacher's speech and activities, a copy of a book, a journal paper, etc.
<b>Noise Source</b>	Disturbance in the classroom, students' psychological states, inadequate or faulty equipment, etc.
<b>Receiver</b>	The student (identical with the destination), a computer
<b>Destination</b>	The student

The productive activities of a student can be described in a similar way, as shown in Table III. The model, if accepted as accurate, also functions as a tool for apportioning blame for failure: if a student's achievement is not satisfactory, then the problem can be sought in one of the steps. So, for example, the teacher's transmission of the information may be inadequate, or the classroom environment may be dysfunctional, or the student may not be listening. Certain approaches to educational technology are often (though not necessarily) aligned with the conduit model. Examples of how researchers are integrating machine learning and GenAI methods with them include competence management systems [36], recommender systems [37], and knowledge management [38].

TABLE III. THE CONDUIT MODEL MAPPED TO STUDENTS' ACTIVITIES

<b>Information source</b>	The student
<b>Transmitter</b>	An essay, assignment, multiple choice test, <i>viva voce</i> , etc.
<b>Noise Source</b>	Poor language and writing skills, student's psychological states, noisy environment, inadequate or faulty equipment, etc.
<b>Receiver</b>	The teacher
<b>Destination</b>	The educational institution

### B. The Coordination Model of Communication

The biological theory of communication put forward by Maturana and Varela in the 1980s is based on coordination rather than transmission. We briefly summarize it here, but the theory is complex, and readers are advised to engage with the original exposition, most accessibly presented in the book 'The Tree of Knowledge' [39]. We summarize Maturana and Varela's view of communication as follows:

- Organisms are organizationally closed but structurally open, i.e. organisms have a standard biological plan which is inherited and fixed, but they grow, think and act in different ways in interactions with the environment.
  - Organisms respond to perturbations in their environment with neuronal activity, but in this process, nothing enters the organism from outside.
  - Organisms become structurally coupled to their environment, i.e. a history of recurrent interactions leads the organism to adjust to its environment, and vice versa.
  - Other organisms are part of the environment. Organisms structurally couple to each other, each adjusting its internal structure in response to the actions of the other. These coordinations constitute communication.
- Humans use sounds, letters, images and movements to coordinate their coordination. For example, we have learned from prior interactions to associate the sound and written form of the word 'baby' with a young human. This higher-level coordination constitutes language (or as Maturana and Varela would prefer 'languaging').

Maturana and Varela "...conclude that, biologically, there is no "transmitted information" in communication", and argue that the conduit metaphor "is basically false" [39] p196. From this perspective, the processes of education should be seen as an ongoing structural coupling between teachers and students, mediated by a wide range of language-based activities. Through these recursive coordinations, the structure of the student changes, and they become able to perform the tasks required of a successful student. Approaches to educational technology that are often (though not necessarily) aligned with the coordination model, and where work is underway to integrate AI, include computer supported collaborative learning [40], self-regulated learning [41], and the use of writing productivity tools [42]. Beyond these, however, lies the largely unmapped terrain of students' informal interactions with GenAI, which has consequences for students coordinations with teachers and institutions that have not yet been fully manifested, let alone understood.

We refer to the combinations of the respective models of educational processes and of communication as the 'transmission paradigm' and the 'coordination paradigm'. GenAI

*The transmission and constructivist models of learning are congruent with the conduit and coordination models of communication.*

## V. GENERATIVE AI: A PARADOX FOR EDUCATION

The coexistence of the two conceptions of education that we have described in Section IV.A and IV.B is radically disrupted by generative AI in general and LLMs in particular. Yeadon et al. write that "short-form essays, written by AI software in only a few seconds, can score a First Class for an assignment from an accredited university Physics module. This, we argue, effectively renders the short-form essay obsolete as an assessment tool." [43] The abilities of current AI should not be overstated, as it falls short in some full examinations. While it was successful in radiology [44] it failed in plastic surgery [45] and in the sixth-grade math and science examinations in Singapore [46]. However, the capabilities of AI will only increase, and as Eulerich *et al.* report [47], ChatGPT 4 can pass exams which were too demanding for ChatGPT 3.

The ability of GenAI technology to create acceptable student texts is a practical problem for education, but the challenge is not unprecedented. As Sharples points out "Transformer AI systems belong to an alternative history of educational technology, where students have appropriated emerging devices – pocket calculators, mobile phones, machine translation software, and now AI essay generators – to make their lives easier. The response from teachers and institutions is a predictable sequence of ignore, resist, then belatedly accommodate." [48] Nevertheless, GenAI presents a different and deeper challenge than the technologies, which Sharples mentions. This is because it can disrupt the equilibrium which has developed between the organizational processes of education and the constructivist practices of teachers, corresponding to the transmission and coordination models of communication.

Education institutions use the performance of students in examinations as evidence of whether students have learned and understood the content of a course or not. According to the transmission view of communication, the information which is extracted at the destination is the same as that which was

transmitted by the information source. Consequently, according to the transmission model of education, in combination with the use of the examination as an assessment instrument, the ability of a GenAI to create texts which merit a pass in an examination must mean that the system, which generates them texts, has knowledge and understanding which matches that of the students who pass the same examinations. However, as we discuss in the next section, there is strong evidence that this is not the case. Education institutions are therefore confronted with a paradox: their educational instruments tell them that LLMs are intelligent learners, but the research evidence and close engagement with GenAI systems shows that LLMs are not intelligent learners. Without resolving this paradox, educational institutions cannot formulate a coherent response to GenAI systems.

*Generative AI creates a paradox for the transmission model of education.*

## VI. WHAT DOES GENERATIVE AI GENERATE?

Some have argued that AI systems do indeed have human level knowledge and understanding, including Blake Lemoine, a Google software engineer working on AI, who was fired in 2022 for maintaining that the system he was working on was conscious [49]. To many others with close knowledge of GenAI, these claims seem intuitively absurd. That is not a sufficient refutation, however, and it is important to establish a stronger argument against ascribing human level capabilities to GenAI.

Gregory Bateson, like Maturana and Varela, worked within the cybernetic tradition, and he thought deeply about the nature of mind and machine. In an earlier paper we have discussed in detail the implications of his work for AI [50], and here we summarize two of his arguments which imply that we should not ascribe human-like mental states to present-day computers.

Firstly, Bateson argues that “The question is not “Can machines learn?” but what level or order of learning does a given machine achieve?” [51] (p.284). Bateson’s *Level I learning* involves changes in the responses which a machine or organism gives at different times, possibly as a result of habituation or reinforcement. This level of learning is displayed by LLMs. *Level II learning* involves ‘learning to learn’, for example as one might improve one’s ability to learn musical scales not by continual practice but by a change in learning strategy. The term ‘deep learning’ refers to the depth of layers of neural networks but gives the impression that AI can learn in a more than superficial way. It is true that LLMs have moved AI closer to *Level II*, to the extent that stochastic changes lead to improved algorithms. However, this takes place within a tightly constrained and fixed framework. There is no equivalent in deep learning to the developmental changes that take place when a student acquires an entirely new body of knowledge or skill, transforming the way they go about solving problems and thinking about the world.

Secondly, Bateson argued that information flow takes place within an ‘ecology of mind’. In his view, mental processes include “a number of phenomena which most people do not think of as processes of thought” [52] p.16, including embryology, evolution, and “all those lesser exchanges of information and injunction that occur inside organisms and between organisms, and that, in the aggregate, we call life.” [52] p.17. This ecology of mind “...will usually not have the same limits as the ‘self’” [51] p.317, and includes both animate and inanimate entities. A computer is not equipped with the sensors and effectors, nor the mental processes which are required to create the rich set of interactive loops between itself and the outside world which constitute an ecology of mind. In other terms, it is not embodied, in the sense that Varela, Thompson and Rosch describe: “...

first, cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological, and cultural context.” [53] (p.173).

More recently, Brian Cantwell Smith has argued along similar lines that all AI systems, including GPTs, literally “do not know what they are talking about” [54] p.76. “...there is no reason to suppose, and considerable reason to doubt, that any system built to date, and any system we have any idea how to build, ‘knows’ the difference between: (i) its own (proximal) state, including the states of its representations, inputs and outputs; and (ii) the external (distal) state of the world that we at least take its states, its representations and those inputs and outputs to represent.” AI is able to perform extraordinarily complex manipulations of words and their tokens, and to relate them to each other. But AI does not know that there is a world external to itself, or that its representations are about that world, and it cannot take responsibility for the adequacy of its representations to describe the world [54] p.79. Consequently, when an AI system produces a student essay about, for example, preserving the rain forest, it cannot ‘know’, in any way that is equivalent to human knowing, what a forest is, nor why it might have importance. It can provide only reports on correlations among its internal states. In this sense, AI systems are electronic solipsists, whose processes correspond to Bradley’s characterization of solipsism as the belief that “nothing beyond my self exists; for what is experience is its states” [55] p.248.

These arguments lead us to conclude that a text produced by generative AI is a simulacrum of human communication which, as Baudrillard put it “itself, no longer even knows the distinction between signifier and signified, nor between form and content”. [56] p.63-64.

*Generative AI generates a simulacrum of human intelligence.*

## VII. THE IMPLICATIONS OF AI SIMULACRA FOR EDUCATION

Consider a trap deployed to attract and snare a pest species, for example Zapponi *et al.* [57] describe how pheromones and vibrations are used for to capture stink bugs. In terms of the conduit model, the sense organs of the insect are the *receiver* of information, and the insect itself is the *destination*. The bug perceives the pheromones and vibrations as a signal whose *transmitter* is the organs of a fellow bug, and the *information source* as a potential reproductive partner. But the bug has been tricked, the information source is in fact a device which generates a simulacrum of a reproductive partner, and the bug has no way of detecting the deception.

An educational institution finds itself in a precisely parallel situation when confronted by examination scripts or essays authored by LLMs: the scripts draw humans into inauthentic interactions with a device. When the assessors of exam scripts award a pass mark to a text produced by AI, they are misled into ascribing to the perceived *information source* knowledge and understanding which is not present. It is axiomatic to the transmission model that learning is contained and transmitted within documents. Consequently, from within this model, an LLM’s success in passing an examination can only understood as demonstrating that the LLM which is the source has indeed learned and understood the material being assessed. Like the target of the pheromone trap, the institution has no way to detect or make sense of the deception from within the confines of a transmission model of educational communication, and cannot abandon the model without undermining the credibility of its own instruments. At the same time, it is also clear that LLMs do not have this ability.

A coordination model of educational communication is better equipped to describe educational interactions with LLMs. The

assessment process is seen to be one more example of the coordination around utterances and documents which establish structural coupling between teachers and students, and through which understanding and knowledge are mutually and iteratively probed. A teacher adopting a coordination model when confronted by an AI script is not dissuaded by the logic of the model from concluding that although the text appears to reflect knowledge and understanding, it does not in fact do so. The teacher and the institution are still challenged by LLMs, as the ease with which inauthentic texts can be generated can disrupt the coordination between teachers and students. It is often difficult to distinguish authentic and inauthentic texts, indeed, as Linardaki reports, on the site “Bot or Not” (botpoet.com) a poem by Gertrude Stein was thought by at least 70% of respondents to have been written by a computer [58]. This is a practical challenge for education seen as a process of coordination, of the same order as those presented by the emergence of calculators and the internet. The same cannot be said for education seen as transmission, which finds its foundational axioms to be threatened, undermining the credibility of grades and diplomas.

The challenge of GenAI for education is thus that it disrupts the balance between the instruments of education (transmission model) and the practice of education (often influenced by the coordination model), by undermining the credibility of the transmission paradigm. If a machine which is widely accepted to be incapable of understanding can pass an examination, we are forced to ask if we can take that examination seriously as a measure of learning and understanding, and if the entire edifice of learning objectives and curricula in fact delivers the learning which it claims to do.

An additional consideration is that the predominant manifestation of AI prior to the emergence of LLMs was the expert system. Expert systems are taxonomic in nature, adhering to explicit classifications. The structure of expert systems corresponds to the taxonomic organization of education, which, for example, subdivides knowledge into disciplines, subdisciplines, curricula, learning resources, etc. Expert systems could reflect these structures, making them easy to apply in education (if not easy to create). Generative AI, however, is not taxonomic, but rather (to use McCulloch’s word) “...anastomotic, whereby afferents of any sort could find their way by intersecting paths to any set of efferents, so relating perception to action” [59] p.392. In this sense, the inner workings of a GPT are not inspectable, and it is not possible to say why, precisely, a particular output was generated. This is a poor fit for an education system which is based on the verifiable delivery of taxonomic knowledge and is required to be transparent and answerable for interactions which take place within them.

*The simulacrum of intelligence produced by Generative AI creates a paradox for the transmission model of education. The coordination model of education is better able to describe educational interactions with generative AI.*

### VIII. THE POSSIBLE RESPONSES OF EDUCATION INSTITUTIONS

One possible conclusion from our discussion in section VII would be that formal education is a fundamentally flawed enterprise, and it should be swept away, together with its instruments. We do not take this position. Rather, we propose that the irruption of LLMs, and GenAI in general, means that the *modus vivendi* between the transmission and coordination models of educational communication will have to be revised. The balance has been disturbed and can only be restored by adjusting the relative influence of the two models on the educational process.

We characterize the possible responses of educational institutions to GenAI in terms of three extremes. In practice, it is likely that institutions will not simply adopt one of these models, but rather experiment with aspects of these strategies in parts of the institution.

#### 1. Reject GenAI

- The institution decides that its business model and processes require a transmission model.
- The coordination paradigm, and the practices influenced by the model, are anathematized and suppressed, and replaced with an emphasis on rote learning and reproduction of specified formulations of knowledge.
- GenAI is rigorously excluded as a disruptive force.

This strategy has three drawbacks. Firstly, it prevents teachers from making use of the pedagogical flexibility which the current *modus vivendi* affords, with consequent negative impact on student outcomes. Secondly, it requires increased coercion of teachers and students, with negative consequences for institutional dynamics and recruitment. Thirdly, it prevents institutions from benefiting from the substantial benefits which GenAI can provide.

#### 2. Embrace GenAI to replace teachers.

- The institution observes that GenAI is cheaper than teachers.
- The institution moves all its courses online, run by AI, and fires all its teachers.
- The institution gains competitive advantage by selling its courses more cheaply than institutions that employ teachers.

This strategy has the drawback of failing to recognize the limitations of current GenAI and the consequent fall in the quality of the education offered. It is also vulnerable to a race to the bottom, where all education is provided by large AI companies, and educational institutions as we currently know them disappear.

#### 3. Embrace AI to support teaching and learning.

- The institution recognizes that GenAI has shown that the transmission paradigm is built on unreliable foundations.
- Educational instruments are reconceptualized as supports for education based on the coordination paradigm, and gradually optimizes them for this revised function, with special attention to assessment.
- The crucial role played by teachers in supporting learning and understanding is recognized. Institutional management processes, unique selling points and business models are revised accordingly.
- GenAI is welcomed as a powerful technology which can support the activities of students, teachers, and administrators in many ways.
- The institution prepares itself for a radical transformation of its processes and the roles of teachers and students.

This strategy has one substantial drawback: it requires the institution to expend its time and resources on rethinking what the education it offers consists of, and how it should be managed and marketed. The potential rewards for this effort, however, are more effective teaching and management processes, and enhanced opportunities for learning.

*Generative AI can support education in different ways, but to benefit fully from its capabilities education institutions need to recognize the validity of the coordination paradigm and reform themselves accordingly.*

### IX. THE EDUCATIONAL OPPORTUNITIES OFFERED BY GENERATIVE AI

The simulacra produced by generative AI are of great utility in many domains and can be used as the basis for the creation of many potentially useful educational applications. We indicate the scope of the services being offered in the following examples, without offering any assessment of their value.

For students, GenAI services can offer support for self-regulated learning and enhancement of students’ autonomy [60]. It can provide



tutoring [61] with recommended learning paths and materials, adjusting them for difficulty and focus; support self-evaluation [62]; provide tools to support the writing process [63].

For teachers, as was the case with earlier waves of educational technology, it is proposed that GenAI services can automate some aspects of their work, saving them time for more important teaching activities. Services include automatic generation of exams and class presentations, as well as automated grading. Indeed services are available that create entire courses [64]. GenAI can create sophisticated games and gamified assessments [65], with a GenAI model being fine-tuned to a topic and then generating game mechanics, including points and a leaderboard that can be used to rank students.

Finally, GenAI can help administrators and policy makers in decision making, as discussed in a recent systematic review [66]. GenAI can also provide administrative support for students, while the company Tribal [67] offers AI driven improvements in admission and enrolment, diversity, timetabling, and predicting and responding to inspections.

The argument made in this paper, however, suggests that success of GenAI in supporting students, teachers and managers will not be determined solely by its technical capabilities and the attractiveness of services such as those discussed above. Its effectiveness will also depend on the ability of institutions to create an environment where people can participate in human-machine interactions in ways which are coherent with the organizational structures and teaching activities of the institution. Gordon Pask, working in the cybernetic tradition, developed a framework called ‘conversation theory’ [68], which provides a starting point for imagining how such interactions might be applied in learning activities.

Pask saw learning as taking place through interpreted formal relationships, with a student’s understanding developing through agreements between the participants in a conversation, typically involving a teacher and a student. To support this conversation, Pask argued that it is “necessary to develop a network of topics and concepts which represent the chosen subject matter area. It is also necessary to ensure that the formal relationships between the concepts are made explicit within the network. The final network within which the student work is called an **entailment structure**” [68] (emphasis in the original). There were two practical barriers to adoption of Pask’s framework. Firstly, Pask specified a complex set of structures and organizations for the implementation of conversations [69]. These requirements were not adopted by Laurillard, who adapted some of Pask’s ideas in her own conversational framework [70]. Secondly, the development of entailment structures for any individual topic was hugely time consuming. It is reasonable to propose that LLMs could provide an entailment structure, as there is no doubt that they provide “a network of topics and concepts which represent” any topic that a student might choose. LLMs can also be interrogated regarding formal relationships between concepts, though these not always explicit. Whatever the detail of correspondence with Pask’s theories, there is certainly an opportunity for students to use LLMs as an opportunity to explore concepts and the relationships between them, and as an emulated interlocutor with which to test their understanding, in combination with conversations with humans (including written and other media exchanges).

In addition to the benefits proposed for GenAI in education, several problems have been identified. Daniel Dennett has recently expressed concerns about GenAI creating ‘counterfeit humans’ and proposed that this should be outlawed [71]. This argument is consistent with our discussion in this paper and would serve to clarify human-machine communication. In a similar vein, the European Writers Council [72] has condemned many aspects of GenAI, including that “Uncontrolled AI output is being pushed into the bestseller lists with click farms”,

often with “identity theft and name deception”. There is clearly a danger that such materials will mislead and confuse students. Other studies have reported that the underlying AI models may be biased leading to inaccurate decisions or results [73] and reinforce stereotypes [74]. Guleria and Sood [75] identify a lack of transparency and the explainability of the output of GenAI, contrasting ‘black box’ machine learning systems with ‘white box’ systems based on “inductive logic programming, rule learners, etc.”

These concerns all revolve around the reliability and transparency of GenAI. Greater transparency of training data and Dennett’s proposed prohibition of counterfeit humans would help in this, but it remains impossible to know exactly how and why deep neural networks produce a particular output. It seems more feasible to use these systems to manage uncertainty rather than in an attempt eliminate it, and to treat their output as explorations or predictions with varying degrees of accuracy and relevance. These can feed into human discussion and analysis, a role for which teachers are well suited.

*Given an appropriate understanding of human-machine communication, generative AI has much to offer to education institutions. Pask’s conversation theory provides a starting point for an exploration of the educational potential of GenAI which is compatible with the coordination paradigm of communication and has a clear role for teachers.*

## X. CONCLUSIONS

This paper has discussed four related domains and argues that each of them can be seen as being informed by a transmission paradigm or by a coordination paradigm. This is summarized in Table IV.

TABLE IV. DOMAINS AND PARADIGMS

Domain	Transmission paradigm	Coordination paradigm
Nature of communication	Conduit of information (misapplication of Shannon and McCulloch)	Coordination (Maturana and Varela’s autopoietic theory)
Models of learning	Delivery model of learning, knowledge and understanding	Constructivist view of learning, knowledge and understanding
Implication for understanding of GenAI in teaching and learning	AI passes exams, so it must have human-like intelligence. But we know that it does not. Result: paradox and rejection	AI disrupts teacher-student interactions but creates many opportunities for learning. Result: challenge and adaptation
Expected institutional response to GenAI	Applications of Gen AI focused on selected existing functions, and retrenchment of traditional educational organization.	Broad application of Gen AI, and rethinking of educational organization and instruments

We have argued that the two paradigms are strongly interconnected vertically in Table IV: i.e. the model of the nature of communication that is adopted determines the model of learning, which in turn molds the response of teachers and institutions to GenAI. Because of the vertical interconnection of the paradigms, contradictions will be generated if an institution seeks to make use of the benefits of GenAI in its teaching and learning, while maintaining its existing use of organization and instruments based on the conduit paradigm. It may be expected that this will then disturb the *modus vivendi* between the organizational structures and instruments of the institution and the practice of teachers and create tensions within the institution. This

implies that institutions should recognize that the educational use of GenAI has greater systemic implications for pedagogy than earlier generations of learning technology, and implications for educational organization and instruments which are greater than any seen since the emergence of the internet.

Educational institutions will have to decide to what extent they will persist with the present model of education in the face of a far greater degree of tension between the transmission paradigm and the realities of teaching and learning, or if they will undertake a serious re-examination of educational processes in the light of developments in AI. Similarly, educational researchers will have methodological challenges in understanding and measuring educational processes based on coordination rather than transmission. Researchers, teachers, and educational administrators will need to take a position on these questions, if they are to avoid confusion in their practice, research, and findings.

As authors of the present study, we are fully aware that we have not provided a complete survey of the fields of information, communication, and pedagogy. Nor would this be possible within the confines of a journal paper. Rather, our purpose has been to distinguish and characterize two paradigms which we believe clarify the questions raised by GenAI for institutions, and to explore their implications. We believe that the distinctions which we have made can help institutions, educational leaders, and teachers to frame the complex and nuanced questions raised by GenAI, and to chart a course towards its effective use in education.

#### ACKNOWLEDGMENT

The authors would like to thank the Research Institute for Innovation and Technology in Education (UNIR iTED), the Universidad Internacional de La Rioja, Logroño, Spain; the Smart Learning Institute (SLI) at Beijing Normal University, China; and the Horizon Europe research project GREAT (grant agreement 101094766), which partially co-funded this research.

#### REFERENCES

- [1] F. García-Peñalvo and A. Vázquez-Ingelmo, "What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, p. 7, 2023, doi: 10.9781/ijimai.2023.07.006.
- [2] W. X. Zhao *et al.*, "A Survey of Large Language Models". arXiv, Sep. 11, 2023. Accessed: Sep. 28, 2023. [Online]. Available: <http://arxiv.org/abs/2303.18223>
- [3] E. Kasneci *et al.*, "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education", *Learning and individual differences*, vol. 103, no. 102274, 2023.
- [4] S. Grassini, "Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings", *Education Sciences*, vol. 13, no. 7, p. 692, Jul. 2023, doi: 10.3390/educsci13070692.
- [5] L. Yan *et al.*, "Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review", *British Journal of Educational Technology*, vol. 55, no. 1, pp. 90–112, 2024, doi: 10.1111/bjet.13370.
- [6] A. Caines *et al.*, "On the application of Large Language Models for language teaching and assessment technology". arXiv, Jul. 17, 2023. Accessed: Oct. 23, 2023. [Online]. Available: <http://arxiv.org/abs/2307.08393>
- [7] I. Drori *et al.*, "From Human Days to Machine Seconds: Automatically Answering and Generating Machine Learning Final Exams", in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Long Beach CA USA: ACM, Aug. 2023, pp. 3947–3955. doi: 10.1145/3580305.3599827.
- [8] S. Moore, H. A. Nguyen, T. Chen, and J. Stamper, "Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods", presented at the European Conference on Technology Enhanced Learning, O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, and T. Papathoma, Eds., in *Lecture Notes in Computer Science*, vol. 14200. Cham: Springer Nature Switzerland, 2023, pp. 229–245. doi: 10.1007/978-3-031-42682-7\_16.
- [9] J. H. Choi, K. E. Hickman, A. Monahan, and D. B. Schwarcz, "ChatGPT Goes to Law School", *SSRN Journal*, 2023, doi: 10.2139/ssrn.4335905.
- [10] D. Yan, M. Fauss, J. Hao, and W. Cui, "Detection of AI-generated Essays in Writing Assessments", *Psychological Test and Assessment Modeling*, vol. 65, pp. 125–144, 2023.
- [11] Z. Liang, W. Yu, T. Rajpurohit, P. Clark, X. Zhang, and A. Kaylan, "Let GPT be a Math Tutor: Teaching Math Word Problem Solvers with Customized Exercise Generation". arXiv, May 22, 2023. Accessed: Oct. 23, 2023. [Online]. Available: <http://arxiv.org/abs/2305.14386>
- [12] G. Zografos and L. Moussiades, "A GPT-Based Vocabulary Tutor.", *Springer Nature*, 2023, pp. 270–280.
- [13] J. M. Markel, S. G. Opferman, J. A. Landay, and C. Piech, "GPTeach: Interactive TA Training with GPT-based Students", in *Proceedings of the Tenth ACM Conference on Learning @ Scale*, Copenhagen Denmark: ACM, Jul. 2023, pp. 226–236. doi: 10.1145/3573051.3593393.
- [14] S. L. Fleming *et al.*, "Assessing the Potential of USMLE-Like Exam Questions Generated by GPT-4", *Medical Education*, preprint, Apr. 2023. doi: 10.1101/2023.04.25.23288588.
- [15] J. C. F. De Winter, "Can ChatGPT Pass High School Exams on English Language Comprehension?", *International Journal of Artificial Intelligence in Education*, Sep. 2023, doi: 10.1007/s40593-023-00372-z.
- [16] M. S. Orenstrakh, O. Karnalim, C. A. Suárez, and M. Liut, "Detecting LLM-Generated Text in Computing Education: A Comparative Study for ChatGPT Cases", *arXiv preprint arXiv:2307.07411*, 2023.
- [17] P. Denny, H. Khosravi, A. Hellas, J. Leinonen, and S. Sarsa, "Can We Trust AI-Generated Educational Content? Comparative Analysis of Human and AI-Generated Learning Resources". arXiv, Jul. 03, 2023. Accessed: Oct. 23, 2023. [Online]. Available: <http://arxiv.org/abs/2306.10509>
- [18] S. Sarsa, P. Denny, A. Hellas, and J. Leinonen, "Automatic Generation of Programming Exercises and Code Explanations Using Large Language Models", in *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 1*, Lugano and Virtual Event Switzerland: ACM, Aug. 2022, pp. 27–43. doi: 10.1145/3501385.3543957.
- [19] A. Gilson *et al.*, "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment", *JMIR Medical Education*, vol. 9, no. 1, p. e45312, Feb. 2023, doi: 10.2196/45312.
- [20] W. Yeadon and D. P. Halliday, "Exploring Durham University Physics exams with Large Language Models". arXiv, Jun. 27, 2023. Accessed: Oct. 23, 2023. [Online]. Available: <http://arxiv.org/abs/2306.15609>
- [21] Y. Kaneda *et al.*, "Assessing the Performance of GPT-3.5 and GPT-4 on the 2023 Japanese Nursing Examination", *Cureus*, Aug. 2023, doi: 10.7759/cureus.42924.
- [22] J. M. Markel, S. G. Opferman, J. A. Landay, and C. Piech, "GPTeach: Interactive TA Training with GPT-based Students", in *Proceedings of the Tenth ACM Conference on Learning @ Scale*, Copenhagen Denmark: ACM, Jul. 2023, pp. 226–236. doi: 10.1145/3573051.3593393.
- [23] Duolingo team, "Introducing Duolingo Max, a learning experience powered by GPT-4", *Duolingo Blog*. Accessed: Oct. 23, 2023. [Online]. Available: <https://blog.duolingo.com/duolingo-max/>
- [24] S. Kahn, "Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access", *Kahn Academy*. [Online]. Available: <https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/>
- [25] A. Varma, C. Dawkins, and K. Chaudhuri, "Artificial intelligence and people management: A critical assessment through the ethical lens", *Human Resource Management Review*, vol. 33, no. 1, p. 100923, Mar. 2023, doi: 10.1016/j.hrmmr.2022.100923.
- [26] R. González-Sendino, E. Serrano, J. Bajo, and P. Novais, "A Review of Bias and Fairness in Artificial Intelligence", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. In press, no. In press, p. 1, 2023, doi: 10.9781/ijimai.2023.11.001.
- [27] A. J. López Rivero, M. E. Beato, C. Muñoz Martínez, and P. G. Cortiñas Vázquez, "Empirical Analysis of Ethical Principles Applied to Different AI Uses Cases", *International Journal of Interactive Multimedia and Artificial*

- Intelligence*, vol. 7, no. 7, p. 105, 2022, doi: 10.9781/ijimai.2022.11.006.
- [28] E. R. Kahu, "Framing student engagement in higher education", *Studies in Higher Education*, vol. 38, no. 5, pp. 758–773, Jun. 2013, doi: 10.1080/03075079.2011.598505.
- [29] J. Girard and J. Girard, "Defining knowledge management: Toward an applied compendium", vol. 3, no. 1, 2015.
- [30] E. V. Glasersfeld, *Radical Constructivism: A Way of Knowing and Learning*. Falmer Press, 1996.
- [31] G. Biesta, "Freeing teaching from learning: Opening up existential possibilities in educational relationships", *Studies in Philosophy and Education*, vol. 34, no. 3, pp. 229–243, 2015.
- [32] N. Bremner, N. Sakata, and L. Cameron, "The outcomes of learner-centred pedagogy: A systematic review", *International Journal of Educational Development*, vol. 94, no. article 102649, Oct. 2022, doi: 10.1016/j.ijedudev.2022.102649.
- [33] M. J. Reddy, "The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language", *Metaphor and thought*, vol. 2, pp. 285–324, 1979.
- [34] C. E. Shannon, "A mathematical theory of communication", *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [35] K. Krippendorff, *On Communication*. New York: Routledge, 2009.
- [36] G. J. Booth *et al.*, "Competency-Based Assessments: Leveraging Artificial Intelligence to Predict Subcompetency Content", *Academic Medicine*, vol. 98, no. 4, pp. 497–504, Apr. 2023, doi: 10.1097/ACM.00000000000005115.
- [37] W. Wang, X. Lin, F. Feng, X. He, and T.-S. Chua, "Generative Recommendation: Towards Next-generation Recommender Paradigm". arXiv, Apr. 07, 2023. Accessed: Feb. 01, 2024. [Online]. Available: <http://arxiv.org/abs/2304.03516>
- [38] M. H. Jarrahi, D. Askay, A. Eshraghi, and P. Smith, "Artificial intelligence and knowledge management: A partnership between human and AI", *Business Horizons*, vol. 66, no. 1, pp. 87–99, Jan. 2023, doi: 10.1016/j.bushor.2022.03.002.
- [39] U. Maturana and F. Varela, *The Tree of Knowledge*. Boston MA: Shambala, 1987.
- [40] U. Cress and J. Kimmerle, "Co-constructing knowledge with generative AI tools: Reflections from a CSCL perspective", *International Journal of Computer-Supported Collaborative Learning*, vol. 18, no. 4, pp. 607–614, Dec. 2023, doi: 10.1007/s11412-023-09409-w.
- [41] I. Molenaar, "The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning", *Computers and Education: Artificial Intelligence*, vol. 3, p. 100070, Jan. 2022, doi: 10.1016/j.caeai.2022.100070.
- [42] D. E. Salinas-Navarro, E. Vilalta-Perdomo, R. Michel-Villarreal, and L. Montesinos, "Using Generative Artificial Intelligence Tools to Explain and Enhance Experiential Learning for Authentic Assessment", *Education Sciences*, vol. 14, no. 1, p. 83, Jan. 2024, doi: 10.3390/educsci14010083.
- [43] W. Yeadon, O.-O. Inyang, A. Mizouri, A. Peach, and C. P. Testrow, "The death of the short-form physics essay in the coming AI revolution", *Physics Education*, vol. 58, no. 3, p. 035027, May 2023, doi: 10.1088/1361-6552/acc5cf.
- [44] S. C. Shelmerdine, H. Martin, K. Shirodkar, S. Shamsuddin, and J. R. Weir-McCall, "Can artificial intelligence pass the Fellowship of the Royal College of Radiologists examination? Multi-reader diagnostic accuracy study", *BMJ*, p. e072826, Dec. 2022, doi: 10.1136/bmj-2022-072826.
- [45] M. Kelloniemi and V. Koljonen, "AI did not pass Finnish plastic surgery written board examination", *Journal of Plastic, Reconstructive & Aesthetic Surgery*, p. S1748681523005983, Oct. 2023, doi: 10.1016/j.bjps.2023.10.059.
- [46] M. R. Das, "Not smarter than a 6th grader: ChatGPT fails Singapore's 6th-grade maths and science exams", Firstpost. Accessed: Oct. 30, 2023. [Online]. Available: <https://www.firstpost.com/world/chatgpt-fails-singapore-6th-grade-maths-and-science-exams-12189482.html>
- [47] M. Eulerich, A. Sanatizadeh, H. Vakilzadeh, and D. A. Wood, "Can Artificial Intelligence Pass Accounting Certification Exams? ChatGPT: CPA, CMA, CIA, and EA?", *SSRN Journal*, 2023, doi: 10.2139/ssrn.4452175.
- [48] M. Sharples, "Automated Essay Writing: An AIED Opinion", *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, pp. 1119–1126, Dec. 2022, doi: 10.1007/s40593-022-00300-7.
- [49] N. Tiku, "The Google engineer who thinks the company's AI has come to life," Washington Post. Accessed: Feb. 16, 2024. [Online]. Available: <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lambda-blake-lemoine/>
- [50] D. Griffiths, "Artificial Intelligence Seen Through the Lens of Bateson's Ecology of Mind", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 1, pp. 62–70, 2021, doi: 10.9781/ijimai.2021.08.004.
- [51] G. Bateson, *Steps to an Ecology of Mind (first published 1972)*. Chicago and London: The University Of Chicago Press, 1987.
- [52] G. Bateson and M. C. Bateson, *Angels fear: Towards an epistemology of the sacred (first published 1987)*, no. 1. Hampton Press, 2005. doi: 10.1016/0732-118x(90)90037-3.
- [53] F. J. Varela, E. Thompson, and E. Rosch, *The Embodied Mind*. Cambridge, MA: MIT Press, 1993.
- [54] B. Cantwell-Smith, *The promise of artificial intelligence: Reckoning and judgment*. MIT Press, 2019.
- [55] F. H. Bradley, *Appearance and Reality, A Metaphysical Essay*. Allen and Unwin, 1893.
- [56] J. Baudrillard, *Simulacra and Simulation translated by Sheila Faria Glaser*. Ann Arbor: University of Michigan Press, 1994.
- [57] L. Zapponi, R. Nieri, V. Zaffaroni-Caorsi, N. M. Pugno, and V. Mazzoni, "Vibrational calling signals improve the efficacy of pheromone traps to capture the brown marmorated stink bug", *Journal of Pest Science*, vol. 96, no. 2, pp. 587–597, Mar. 2023, doi: 10.1007/s10340-022-01533-0.
- [58] X. A. (Christina Linardaki), "Artificial Intelligence vs. Human Intelligence: The case of poetry", *Academia Letters*, Jan. 2021, Accessed: Sep. 08, 2023. [Online]. Available: [https://www.academia.edu/49359147/Artificial\\_Intelligence\\_vs\\_Human\\_Intelligence\\_The\\_case\\_of\\_poetry](https://www.academia.edu/49359147/Artificial_Intelligence_vs_Human_Intelligence_The_case_of_poetry)
- [59] W. S. McCulloch, *Embodiments of Mind*. Cambridge, MA: MIT Press, 1988.
- [60] "War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education", *Journal of Applied Learning and Teaching*, vol. 6, no. 1, Apr. 2023, doi: 10.37074/jalt.2023.6.1.23.
- [61] TutorAI, "Learn anything". Accessed: Oct. 31, 2023. [Online]. Available: <https://www.tutorai.me/>
- [62] Formative, "Formative AI". Accessed: Oct. 31, 2023. [Online]. Available: <https://en-gb.formative.com/ai-powered>
- [63] jenni, "Supercharge Your Next Research Paper". Accessed: Oct. 31, 2023. [Online]. Available: <https://jenni.ai/>
- [64] CourseAI, "Our AI Course Creator tool streamlines the process of creating online courses and empowers financial independence using AI technology", CourseAI. Accessed: Oct. 31, 2023. [Online]. Available: <https://courseai.com/>
- [65] Taskade, "AI Educational Game Generator". Accessed: Oct. 31, 2023. [Online]. Available: <https://www.taskade.com/generate/education/educational-game>
- [66] J. de Souza Zanirato Maia, A. P. A. Bueno, and J. R. Sato, "Applications of Artificial Intelligence Models in Educational Analytics and Decision Making: A Systematic Review", *World*, vol. 4, no. 2, Art. no. 2, Jun. 2023, doi: 10.3390/world4020019.
- [67] Tribal, "How AI is assisting students teachers and administrators today". Accessed: Oct. 31, 2023. [Online]. Available: <https://www.tribalgrou.com/blog/how-ai-is-assisting-students-teachers-and-administrators-today>
- [68] G. Pask, "Conversational techniques in the study and practice of education", *British Journal of Educational Technology*, no. 46, pp. 12–25, 1976.
- [69] G. Pask, B. C. E. Scott, and D. Kallikourdis, "A theory of conversations and individuals (Exemplified by the Learning Process on CASTE)", *International Journal of Man-Machine Studies*, vol. 5, no. 4, pp. 443–566, Oct. 1973, doi: 10.1016/S0020-7373(73)80002-1.
- [70] D. Laurillard, *Rethinking University Teaching: a framework for the effective use of educational technology*. London: Routledge, 1993.
- [71] D. C. Dennett, "The Problem With Counterfeit People", The Atlantic. Accessed: Oct. 24, 2023. [Online]. Available: <https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/>
- [72] European Writers Council, "Analysis: The success of Generative AI in the book sector is based on theft". Accessed: Oct. 30, 2023. [Online]. Available: <https://europeanwriterscouncil.eu/gai-is-based-on-theft/>
- [73] L. Lucy and D. Bamman, "Gender and Representation Bias in GPT-3 Generated Stories", in *Proceedings of the Third Workshop on Narrative Understanding*, Virtual: Association for Computational Linguistics, Jun. 2021, pp. 48–55. doi: 10.18653/v1/2021.nuse-1.5.



- [74] V. Turk, "How AI reduces the world to stereotypes", Rest of World. Accessed: Oct. 24, 2023. [Online]. Available: <https://restofworld.org/2023/ai-image-stereotypes/>
- [75] P. Guleria and M. Sood, "Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling", *Education and Information Technology*, vol. 28, no. 1, pp. 1081–1116, Jan. 2023, doi: 10.1007/s10639-022-11221-2.



Dai Griffiths

Dai Griffiths, also known as David, has a background in the arts, and holds a PhD from Universitat Pompeu Fabra. He spent the first part of his career working as a teacher in primary, secondary and higher education, as well as in interpersonal skills training in industry, before becoming fascinated by the potential of computers in education. For the past twenty-five years he has worked in the development of educational applications, and as an educational technology researcher, and has published extensively. In this work he became deeply engaged with the tradition of cybernetics. He was appointed Professor at the Institute for Educational Cybernetics at the University of Bolton, where he worked with the Centre for Education Technology Interoperability and Standards (Cetis). He then took on a role in the Department of Education of Bolton University, leading the Department's PhD and Doctor of Education programs. Dai Griffiths is currently a Senior Researcher at the Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR). ORCID: 0000-0002-6863-2456.



Enrique Frias-Martinez

Enrique Frias-Martinez is an applied researcher in Artificial Intelligence with over 15 years of experience working mainly on EdTech, urban computing and AI for Social Good. Currently he works as a Senior Researcher at UNIR-iTED (Research Institute for Innovation and Technology in Education). He has held positions at Telefonica Research, at the Department of Biomedical Engineering of the University of California, Los Angeles (UCLA), at the Department of Information Systems & Computing of Brunel University London, and at the Courant Institute of Mathematical Sciences, New York University. Enrique received a Ph.D. in Computer Science from the Universidad Politécnic de Madrid, Spain, in 2001, and a Ph.D. degree in Information Systems from Brunel University London in 2007. He was the recipient of the Best Ph.D. Thesis Award of the School of Computer Science 2001, Universidad Politécnic de Madrid. ORCID: 0000-0001-5348-3120



Ahmed Tlili

Prof. Ahmed Tlili is an Associate Professor at Beijing Normal University, China, Adjunct Associate Professor at An-Najah University, Palestine and a Visiting Professor at UNIR, Spain. He is the Co-Director of the OER Lab at the Smart Learning Institute of Beijing Normal University (SLIBNU), China. He serves as the Editor of Springer Series Future Education and Learning Spaces, and the Executive Editor of Smart Learning Environments. He is also the Associate Editor of IEEE Bulletin of the Technical Committee on Learning Technology, and the Journal of e-Learning and Knowledge Society. Prof. Tlili serves an expert at the Arab League Educational, Cultural and Scientific Organization (ALECSO). He has edited several special issues in several journals. He has also published several books, as well as academic papers in international referred journals and conferences. He has been awarded the Martin Wolpers 2021 Prize by the Research Institute for Innovation and Technology in Education (UNIR iTED) in recognition of excellence in research, education and significant impact on society. He also has been awarded the IEEE TCLT Early Career Researcher Award in Learning Technologies for 2020. ORCID: 0000-0003-1449-7751



Daniel Burgos

Daniel Burgos works as a full professor of Technology for education and communication and vice-rector for International Research at the Universidad Internacional de La Rioja (UNIR). He holds a UNESCO Chair on eLearning. He is the Director of the Research Institute for Innovation and Technology in Education (UNIR iTED, <http://ited.unir.net>). He has implemented more than 80 European and worldwide R&D projects and published more than 250 scientific papers, 58 books and special issues, and 11 patents. He is an adjunct professor at Universidad Nacional de Colombia (UNAL, Colombia), a full professor at An-Najah National University (Palestine), an extraordinary professor at North-West University (South Africa), a visiting professor at Coventry University (United Kingdom), and at the China National Engineering Research Center for Cyberlearning Intelligent Technology (CIT Research Center, China); and a Research Fellow at INTI International University (Malaysia). He works as a consultant for the United Nations (UNECE), ICESCO, the European Commission and Parliament, and the Russian Academy of Science. He holds 12 PhD degrees, including Computer Science and Education. ORCID: 0000-0003-0498-1101.

# Ethical Implications and Principles of Using Artificial Intelligence Models in the Classroom: A Systematic Literature Review

Lin Tang<sup>1</sup>, Yu-Sheng Su<sup>2</sup> \*

<sup>1</sup> Nanjing Normal University, (China)

<sup>2</sup> Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi county (Taiwan)

Received 2 October 2023 | Accepted 13 February 2024 | Published 27 February 2024



## ABSTRACT

The increasing use of artificial intelligence (AI) models in the classroom not only brings a large number of benefits, but also has a variety of ethical implications. To provide effective education, it is now necessary to understand the ethical implications of using AI models in the classroom, and the principles for avoiding and addressing these ethical implications. However, existing research on the ethical implications of using AI models in the classroom is rather sparse, and a holistic overview is lacking. Therefore, this study seeks to offer an overview of research on the ethical implications, ethical principles and the future research directions and practices of using AI models in the classroom through a systematic literature review. Out of 1,445 initially identified publications between 2013 and 2023, 32 articles were included for final coding analysis, identified using explicit inclusion and exclusion criteria. The findings revealed five main ethical implications, namely algorithmic bias and discrimination, data privacy leakage, lack of transparency, decreased autonomy, and academic misconduct, with algorithmic bias being the most prominent (i.e., the number of existing studies is the most), followed by privacy leakage, whereas decreased autonomy and academic misconduct were relatively understudied; and six main ethical principles, namely fairness, privacy, transparency, accountability, autonomy and beneficence, with fairness being the most prominent ethical principle (i.e., the number of existing studies is the most), followed by privacy, while autonomy and beneficence were relatively understudied. Future directions of research are given, and guidelines for future practice are provided: (1) further substantive discussion, understanding and solution of ethical implications are required; (2) the precise mechanism of ethical principles of using AI models in the classroom remains to be elucidated and extended to the implementation phase; and (3) the ethical implications of the use of AI models in the classroom require accurate assessment.

## KEYWORDS

Artificial Intelligence, Classroom, Education, Ethical Implications, Ethical Principles, Systematic Literature Review.

DOI: 10.9781/ijimai.2024.02.010

## I. INTRODUCTION

**A**RTIFICIAL intelligence (AI) is defined as a branch of computer science that simulates intelligent behavior in computers, in an attempt to develop human-like intelligence machines [1]. In recent years, AI has become an indispensable part of people's lives with its powerful functions, and it deeply affects all areas of human activities, including education [2]. In order to achieve Sustainable Development Goal 4 (SDG4) of UNESCO's Agenda 2023 on quality and inclusive education [3], many AI models have been applied in real classrooms to promote instruction and learning, such as Google Classroom (which was applied to online teaching management) [4], Google Dialogflow (which was used as a virtual education assistant) [5], and GPT (which was applied to automatic question generation and essay scoring) [6].

These AI models are based on powerful algorithms and generating capabilities to support personalized learning systems and automated assessment systems that facilitate students' learning and teachers' teaching [7]. They contribute to students' learning and can also free teachers from heavy work [8]. Compared with traditional computer-based models, these AI models can provide more dynamic and realistic learning experiences [9].

However, despite the use of AI models in the classroom having many undeniable benefits, it also raises potentially extensive ethical implications, for instance, leakage of personal private data caused by the collection of large amounts of data, discrimination and unfairness caused by algorithmic bias, and lack of integrity caused by the abuse of technology [10], [11]. Thus, using AI models in the classroom is seen as a complex and highly controversial issue [12]. Actually, however, we just need to assume what our response speed will be, rather than ignoring or banning AI, thus avoiding extremism [13]. SDG4 emphasizes that AI technologies must be applied to ensure equitable and inclusive access to education [3]. Hence, AI should be used to

\* Corresponding author.

E-mail addresses: 200602093@njnu.edu.cn (L. Tang), ccucssu@gmail.com (Y. S. Su).

enhance and amplify the ability of teachers to teach and students to learn, instead of being replaced by them. In fact, both discriminative AI models and generative AI models are the result of data-driven model training, not a mystical magic [14]. Therefore, AI models are not a panacea, and this understanding of AI will help address the ethical implications explored in this study.

In order to avoid and address the ethical implications resulting from the use of AI models in the classroom, more ethical principles need to be considered, such as privacy, fairness, transparency and accountability [15]. Recently, some researchers and international organizations have specifically studied the ethical principles when applying AI in the field of education [16]. It is worth noting that some ethical principles overlap in these reports, but few studies have systematically examined the global consensus on the ethical principles of using AI in the classroom [17]. At a more formal and legal level, some countries and organizations have developed or are developing general laws about AI, such as the United States' AI Bill of Rights [18] and Canada's Artificial Intelligence and Data Act [19], which also cover the ethical aspects of using AI. Most notably, the European Union approved the Artificial Intelligence Act in December 2023 [20], which is the first global comprehensive regulation of the field of AI. Some rules for AI development and use have also been developed, such as human oversight, security, privacy, transparency, non-discrimination, and social and environmental well-being [21]. These efforts aimed to achieve a consensus on the rational and regulated use of AI through ethical and legal constraints.

In general, using AI models has brought about some ethical implications while improving the quality of instruction and learning. However, the ethical implications, ethical principles, and related research directions of using AI models in the classroom still need to be clarified. Previous systematic review work has provided some substantial insights into AI in education, including theoretical paradigms, applications, benefits, challenges and trends [8], [22]-[24]. However, literature reviews on the ethical implications about the use of AI models in the classroom are limited, and there is no research involving a systematic literature review, resulting in the lack of a holistic view. Additionally, the reviewing which principles are required to avoid and address the ethical implications of AI model use in the classroom remains inadequate and has only been macroscopically articulated in a few studies [11],[17]. Meanwhile, the future research and practice directions of related research could be clearer. Further research is urgently needed to clarify the ethical implications, ethical principles, and future research directions of using AI models in the classroom. Compared with the general literature review method, the systematic literature review method based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) emphasizes following strict steps to extract valid information from existing literature, to draw comprehensive conclusions [25], which are conducive to providing evidence for solving the research questions of this study. Therefore, to make up for this lack of research, this systematic literature review collected, reviewed, and summarized the research on the ethical implications, principles and future research directions of using AI models in the classroom.

The rest of this paper is structured as follows. Section II critically reviews the ethical implications and principles of using AI models in the classroom and raises the research questions. Section III describes the PRISMA method used in this study. Next, the results and findings of the literature analysis regarding the research questions are presented in Section IV. Based on the literature analysis, Section V provides further discussion of the results and findings. In Section VI, the implications, limitations and future work of the current study are illustrated.

## II. LITERATURE REVIEW

### A. The Ethical Implications of Using AI Models in the Classroom

In recent years, AI models have been widely used in the classroom. For example, Zhang et al. [26] introduced AI models into flipped classrooms to digitize and visualize course material preparation, supporting AI-assisted interactive classroom learning. Suresh et al. [27] discussed the application of several deep learning models for promoting fair classroom discussion. Recently, some studies have reported on generative language models such as chatbots being applied in the classroom to support student learning [28], [29]. However, despite the use of AI technology in teaching and learning bringing huge benefits to revolutionize education, the integrating AI models into the classroom could have significant ethical implications [7]. Some typical ethical implications have been identified. For instance, powerful algorithm-based AI predictive models can indeed provide personalized learning for students [30], assist teachers in instructional design [31], and provide references for administrators in making educational decisions [32]. However, one potential ethical issue is that AI algorithms can be biased. Verma [33] argued that if the data are biased, the AI models perpetuate those biases, thereby exacerbating existing discrimination in educational systems. In addition, data-driven AI models require collecting and storing large amounts of sensitive student data. This could raise another potential ethical issue, namely that these data could be used for unintended purposes or be accessed by unauthorized individuals, leading to students' privacy disclosure [16], [34]. Moreover, lack of accountability and transparency are also major ethical implications of using AI models in the classroom, which leads to the question of who is responsible for the accuracy of educational decisions based on AI models and how they are made [35]. Recently, Naidu and Sevnarayan [36] reported on the potential crisis of academic integrity arising from using ChatGPT, an emerging large AI language model, for online assessment in distance education.

In addition, the systematic literature review approach has provided comprehensive views of other aspects of using AI in education, such as paradigms, applications, benefits, challenges, and trends [8], [22]-[24]. For instance, Tahiru systematically reviewed the challenges of implementing AI in education, including ethics, privacy, and trust [24]. These are echoed in Murphy's research. He systematically summarized the major applications of AI models in education, such as rule-based expert systems, intelligent tutor systems (including a student model and a teacher model), and machine learning (including automated scoring systems and early warning systems). Meanwhile, he pointed out that these models are error-prone when used in different scenarios, which can have ethical implications including bias, transparency and trust [8]. Based on the e-learning background, Tang et al. [23] systematically summarized the future research trends of the use of AI in education, especially emphasizing the assessment of the environment and its implications.

However, in these studies, the ethical implications were not a major part of the review and were only briefly summarized. Moreover, previous research has not specifically elaborated the ethical implications about the use of AI models from a teaching and learning perspective, and there has been a lack of attention to the future direction of related research. Hence, it is not clear what the main aspects and concrete content of the ethical implications are, and what future research directions will be. Additionally, in contrast to other review methods, the approach of PRISMA emphasizes the search and selection of literature guided by the research question [25], which has the potential to provide a complete picture of ethical implications of AI models use in the classroom, as demonstrated by research on the ethical implications about AI in other fields [37]. Therefore, there is an urgent need for a systematic review to clarify the main ethical implications using AI models in the classroom, to guide instruction practice.



### B. The Ethical Principles of Using AI Models in the Classroom

The ethical principles are the guidelines that should be followed for the ethical use of AI models in the classroom to avoid and address ethical implications [15]. Hagendorff [38] emphasized that ethical principles for using AI are necessary and must be aligned with societal values. Some international organizations (e.g., UNESCO Education & AI and the European Commission) have reported the general ethical principles that should be followed from AI design and development to its use, such as security, privacy, transparency, accountability, inclusiveness, sustainability, and human-centeredness [39], [40]. Likewise, these ethical principles have been further discussed in education. For example, in terms of the principle of privacy, Miao et al. [39] considered that to protect the privacy of teachers and students, it is necessary to collect and analyze the points of teachers and students before using AI models to decide how to deploy AI in the classroom. Additionally, the large collection of student and teacher data highlighted the need for transparency in using AI models [41]. The principle of transparency refers to the detailed explanation of using AI models, including what the data are, how they are collected, how they are interpreted, and how they are used [15]. Slimi and Carballido [42] emphasized that the principle of transparency is critical for teachers and students because data visualization can be used to analyze student learning behaviors and trajectories and to provide additional support for teachers' instruction. Moreover, the principle of accountability ethics has also been called for in some studies. For example, Klimova et al. [11] highlighted the primary responsibility for clarifying the use of AI-driven mobile apps in education. Hong et al. [43] pointed out that when AI is applied in education, it should be determined who is responsible for the consequences of the data use. These studies required clear subject responsibility for educational decision making based on AI models.

However, review work on ethical principles of using AI models in the classroom is still insufficient, and only a few reviews have been conducted [11], [17]. Specifically, Klimova et al. [11] synthesized eight articles on the ethical principles of using AI in education, and concluded four major principles, namely beneficence, accountability, justice and human values. Regrettably, this study reviewed only a few articles, and needed to provide further analysis of these ethical principles. Additionally, in Memarian and Doleck's research [17], they examined the fairness, accountability, transparency, and ethics in AI in the context of higher education, but they did not define ethical principles as the primary research focus. Importantly, previous studies did not explore the principles that should be followed to avoid and address the ethical implications of using AI models in the context of the classroom. More perspectives have focused on the macro context of education. However, some of the ethical implications arising from the current use of AI models in teaching urgently require a research perspective focused on the classroom. Further, future directions for related research have not been specifically discussed in previous studies. Therefore, it is still unclear what the main aspects of the ethical principles of using AI models in the classroom are, and what the future research directions are. Since the PRISMA method can extract and interpret data more accurately than the general review method [25], it is conducive to providing more accurate answers to the questions in this study. All in all, based on clarifying the ethical implications of using AI models in the classroom, this study further systematically reviewed the ethical principles that should be considered.

### C. Research Questions

To further understand the ethical implications when applying AI models in the classroom, this systematic review examined the ethical implications and ethical principles from the teaching and learning perspective. Additionally, the future research directions of related

research were also investigated. Specifically, the following research questions were proposed in this study:

RQ1: What are the ethical implications of using AI models in the classroom?

RQ2: What are the required principles of using AI models in the classroom to avoid and address the ethical implications?

RQ3: What are future directions of research and practice regarding the ethical implications and principles of using AI models in the classroom?

## III. METHOD

The systematic review method was adopted in this study based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) principles, comprising a total of four phases: identification, screening, eligibility, and included [25]. This methodology was designed to answer specific research questions through clear, systematic, and repeatable search strategies [44]. Next, the procedure of systematic literature review in this study will be described.

### A. Literature Search Process

Two international publication databases were selected to search the full-text archives, including the Web of Science and Scopus, which are the most comprehensive databases of academic literature [45]. In both databases, most journals are predominantly in the English language [46]. Burnham [47] insisted that the Web of Science and Scopus can complement each other to improve the coverage rate of related articles. Importantly, the works contained in these databases are seen to be of high quality and to have significant impact in social science, and the databases cover a wide range of educational journals [48]. Compared with other databases, these two provide a variety of search methods and browsing options, including standard, basic and advanced methods, which is more conducive to the accuracy of literature searches [46]. Additionally, the high accessibility of their journals in the academic community is more conducive to the conduct of this research. Furthermore, the literature searched for in this study covers articles published between 2013 and July 2023, because AI began to make significant progress in education from around 2013 [45]. To ensure the quality of the review, the selected articles were only from peer-reviewed papers, because they have a high degree of credibility and have undergone a rigorous review process [49]. In addition, conference proceedings (if available) were included in this study to obtain up-to-date information on the ethical aspects of the use of AI models in the field of education. Based on the Cochrane Handbook of Systematic Reviews of Interventions, searching for conference proceedings is considered a highly reliable practice, because it is beneficial to capture as many studies as possible, and can greatly reduce the risk of publication bias [50]. After the full text was filtered, according to the guidelines [51], the snowball method was applied to find further papers which were not retrieved through the search strings.

The structured search strategy was adopted in this study to search the databases. To find the most relevant literature in this field, the PICOC (Population, Intervention, Comparison, Outcome and Context) framework, proposed by Petticrew and Roberts [52], was adopted to define the search string and the scope of this study (see the details below):

- a) Population: this study deals with terms, keywords, or some variation of the same meaning related to AI models, classrooms, and ethics. Therefore, the search string was defined according to these criteria.

- b) Intervention: to implement this theme, some exclusion criteria were designed, as shown in Table I. Articles that did not meet these specific requirements were excluded.
- c) Comparison: emphasis was on the specific ethical implications and principles of using AI models, rather than on a broader picture of their use.
- d) Outcome: this step determined which outcomes were the most relevant to answering the research questions [52]. Hence, in addition to the ethical implications and principles of using AI models in the classroom, future research directions related to the topic were also included as outcomes.
- e) Context: the last step is the “context” that defines the boundaries of the questions, which was defined as classroom teaching and education.

Ultimately, the following search string and Boolean operators AND/OR were utilized: (“Artificial intelligence” OR “AI” OR “AI model”) AND (“classroom” OR “educat\*”) AND (“ethic\*” OR “moral\*”). This literature search was conducted in August 2023 and initially identified 1,445 records (1,063 from WOS, 382 from SCOPUS).

### B. Inclusion and Exclusion Process

To improve the pertinence of the literature in the analysis of the research questions, a set of inclusion and exclusion criteria was designed to identify better the papers that focused on the ethical implications of the use of AI models in the classroom (shown in Table I). Specifically, these criteria were mainly based on the following considerations: (a) published between 2013 and 2023, as AI has made significant progress in the field of education since 2013 [45]; (b) written in English, not only because English is the internationally recognized language in the field of science, but also the same language is more conducive to text-mining analysis; (c) research from articles or conference proceedings were chosen, because they are highly scholarly; (d) sourced from peer-reviewed scientific papers, as these papers are typically evaluated by experts in their subject area, thus ensuring some form of quality check; (e) conducted in the field of education, because this was in line with the background of this study, for example, research in the field of medicine was excluded, but research in the field of medical education was included; (f) focus on the use of AI models in education, rather than the design and development of AI models; and (g) focus on the ethical implications of using AI models, rather than simply mentioning them, and discussion of the ethical implications as an important part of the research.

TABLE I. INCLUSION AND EXCLUSION CRITERIA

Inclusion criteria	Exclusion criteria
Research must be published from 2013 to 2023.	Research published before 2013.
Research must be written in English.	Research written in any other languages.
Research from articles or conference proceedings.	Research from book chapters, magazines, news, and posters.
Research must be sourced from peer-reviewed scientific papers.	Research not sourced from peer-reviewed scientific papers.
Research must be carried out in the field of education.	Research conducted in fields other than education.
Research must focus on using AI models in education.	Research not focused on using AI models in education.
Research must focus on the ethical implications of using AI models.	Research not focused on the ethical implications of using AI models.

After deleting 151 duplicates, the remaining 1,294 articles were screened according to the inclusion and exclusion criteria. The

number of articles that did not meet the criteria by reviewing the titles and abstracts was 121. Subsequently, 79 articles that were inconsistent with the research purpose were further excluded by full text reading, and 28 relevant articles were identified. According to Webster and Watson’s suggestions [53], a forward and backward reference search was carried out for these articles to identify further relevant records. In the backward search, references for 28 articles were analyzed, and in the forward reference, Google Scholar was used to analyze and identify articles that cited reservations; as a result, four articles were added after review according to the inclusion and exclusion criteria. Eventually, 32 eligible articles were identified for systematic review. The PRISMA flow diagram of the study is summarized in Fig. 1.

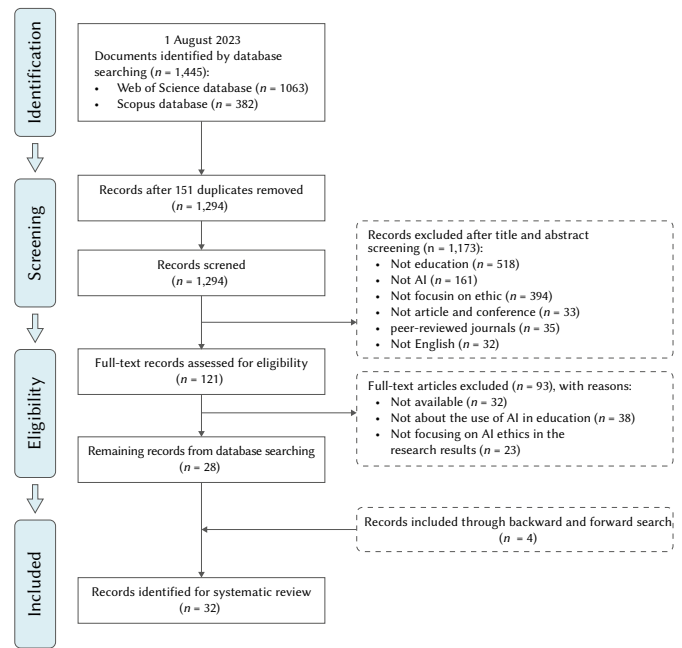


Fig. 1. PRISMA flow diagram of the study.

To ensure rigorous evaluation of articles included in the review, the following quality criteria were developed: (a) the article clearly defined the purpose of the research; (b) the article disclosed the research methodology used; (c) the article clearly presented one or more ethical implications or principles for the use of AI models in the classroom; and (d) a comprehensive description of the results was provided in the study.

As shown in Table II, among the 32 included articles, 23 were journal papers and nine were conference papers. In terms of regional distribution of the literature, the most prolific region for relevant literature was Europe (N = 17, 53%), followed by North America (N = 11, 35%), Asia with less (N = 3, 9%), and Oceania with the least (N = 1, 3%). In terms of the distribution of educational stages, except for discussion in which the scope was not specified (N = 21, 66%), the research was mainly concentrated on higher education (N = 8, 25%). In addition, the major research methods for the 32 studies were also identified, including literature study, quantitative survey, interview and observations, exploratory research, perspective, case study and mixed methods. Among them, literature study was the most commonly used method (N = 12, 38%), followed by perspective (N = 9, 28%).

### C. Data Analysis

The inductive grounded method was applied to analyze and classify the information in the 32 eligible articles relevant to the research question [54]. This classification method identifies and refines topics through data rather than pre-determined categories or theories,

TABLE II. THE GENERAL INFORMATION OF THE 32 INCLUDED ARTICLES

General information	Category	N	%
Article type	Journal	23	72
	Conference paper	9	28
Country/Region	Europe	17	53
	North America	11	35
	Asia	3	9
	Oceania	1	3
	Higher education	8	25
Educational sector	K-12	3	9
	Unspecified field	21	66
	Literature study	12	38
Method	Quantitative survey	3	9
	Interview and observations	4	13
	Exploratory research	2	6
	Perspective	9	28
	Case study	1	3
	Mixed methods	1	3

which facilitates the extraction of new findings and protects the richness of the data [55]. To help answer the research questions, the following extraction framework was identified to extract data from the 32 included articles: study objectives, study design, study type, educational topic, AI application, main findings, ethical implications, ethical principles and the considerations of future work.

To answer RQ1—What are the ethical implications of using AI models in the classroom?—inductive analysis was performed to extract information about the ethical implications by manually mapping each article. These descriptive data were reconstructed through the process of coding, conceptualization, and classification. Specifically, coding was used to identify sentences or paragraphs from the data that were relevant to the ethical implications, and to describe them with short phrases. Then, the constant comparison method was applied to combine the similar codes to form categories about ethical implications [55]. To answer RQ2—What are the ethical principles of using AI models in the classroom?—the researchers read through the full text and adopted the above analytical steps to form categories about ethical principles. Meanwhile, the frequencies of each category coded were calculated in this study. To answer RQ3—What are future directions of research and practice regarding the ethical implications and principles of using AI models in the classroom?—the recommendations made by each researcher in the discussion and conclusion sections were read manually.

Additionally, three strategies were adopted in this study to ensure the reliability of the literature analysis. Firstly, two trained researchers carried out constant discussions until agreement was reached to verify the categories [56]. Secondly, according to Hsieh and Shannon [57], the Results section of this study explains in detail the categories of the literature findings for each research question. Finally, some examples are presented within each category to prove how well categories represent the data in response to the research questions [58].

#### IV. RESULTS

##### A. What Are the Ethical Implications of Using AI Models in the Classroom?

Content analysis of the current literature revealed five categories of ethical implications about the use of AI in the classroom: data privacy leakage, algorithmic bias and discrimination, lack of transparency, decreased autonomy, and academic misconduct (shown in Table III).

TABLE III. THE ETHICAL IMPLICATIONS OF USING AI MODELS IN THE CLASSROOM

Category	Example	Na	Sample studies
Algorithmic bias and discrimination	Gender discrimination.	7	Ghotbi and Ho [68]
	Racial or ethnic discrimination.	6	Ghotbi et al. [69]
Loss of privacy	Class discrimination.	2	Matias and Zipitria [70]
	Cultural bias.	2	Masters [71]
	Personal information is leaked.	6	Köbis and Mehner [72]
Lack of transparency	Increasing culture of algorithmic surveillance.	5	Reiss [35]
	The absence of the right to be forgotten and to give informed consent.	3	Adams et al. [73]
Academic misconduct	Teachers and students may have difficulty understanding predictions related to learning performance.	5	Hong et al. [43]
	The explanation of potential or actual disadvantages or risks of using AI models in the classroom is not apparent.	3	Slimi and Carballido [42]
Decreased autonomy	The ability of students and teachers to manage their own lives is reduced.	6	Han et al. [74]
Academic misconduct	Cause cheating and plagiarism issues.	5	Adams et al. [75]

<sup>a</sup> The number of studies added up to more than 32 because multiple ethical implications of using AI models in the classroom are described in several studies.

##### 1. Algorithmic Bias and Discrimination

Algorithmic bias and discrimination is the most studied ethical implication in the reviewed studies ( $N = 17$ ). Due to the training data, AI models have been shown to exhibit bias and discrimination, which reinforce inherent stereotypes. The existing literature focuses on bias around databases and algorithms of AI models against certain groups of students (often underrepresented minorities), involving aspects such as gender, race or ethnicity, class and cultural background. Specifically, gender discrimination is one of the most apparent forms of this issue. Akgun and Greenhow revealed the gender stereotypes of using AI models in language learning classrooms. When students learn the translation of sentences using generative language models, such as those about doctors and soldiers, they are often translated as male, which exacerbates some social prejudices and gender stereotypes [10]. Additionally, Holmes et al. noted that the application of AI models could be influenced by “cultural imperialism,” leading to ethical issues of cultural discrimination in the classroom [16]. Furthermore, biased decision algorithms have been shown in AI models, such as personalized learning, automated assessment, facial recognition systems and predictive systems in education [58]. Kooli considered that AI models, such as chatbots, can produce inaccurate results or misleading information that can lead to decisions being made against specific groups of students [59].

##### 2. Data Privacy Leakage

Data privacy leakage is a critical ethical issue in debates of using AI models in the classroom ( $N = 14$ ). AI models are used to analyze, assess and predict students’ learning performance by accumulating large amounts of diverse data, such as personal background information, academic performance, facial expressions, and verbal records [35]. While these models can optimize the learning experience in the



classroom, they also raise some ethical problems about data privacy, including personal information leakage, surveillance and student tracking, the absence of informed consent and the right to be forgotten. Leakage of students' personal information is a frequently reported problem [60]. Kowch posed that the long-term tracking of students by AI has led to privacy disclosure, and AI surveillance is a difficult ethical issue that has long been considered [61]. Holmes et al. considered that students still lack a real opportunity to choose whether to opt in or out of educational AI systems, and that the right to informed consent and to be forgotten are important [62]. Therefore, enhancing privacy protection is a must for using AI models in the classroom.

### 3. Lack of Transparency

Lack of transparency is the third most ethical concern in the studies reviewed ( $N = 8$ ). It is worth noting that transparency is also directly called explainability in some studies, such as Jang et al.'s [63] and Farrow's research [64]; it refers to the detailed explanation of algorithmic decisions or the collection and processing of data. There is a general decline in transparency around the use of AI models [17]. Kooli considered that the current application of AI models in the classroom still has a lack of explanation, that is, most teachers and students do not understand the process of AI decision making, and how and under what conditions to use these data [59]. Importantly, the former risks and practical downsides of using AI models in the classroom are not spelled out in detail [41]. Chen et al. [29] demonstrated that when a chatbot was designed for use in the classroom to support students' learning, it was not always able to identify spelling mistakes or understand colloquial speech. Further, the chatbot lacked a deeper understanding of the emotions expressed by students, such as sarcasm. As noted in Hong et al., a lack of transparency has led some teachers and students to question the results of AI algorithm-based learning predictions and decision models [35].

### 4. Decreased Autonomy

Decreased autonomy is also a serious ethical issue discussed in the reviewed studies ( $N = 6$ ). It is worth noting that autonomy is also directly called agency in some studies, such as Tuomi [65] and Holmes et al. [16]; it refers to individuals being free to pursue goals and values that they deem important. Schiff suggested that in the case of using AI models in the classroom, inappropriate decision-making empowerment could potentially decrease and even undermine the autonomy of teachers, students, and parents, and he further emphasized that such problems have already arisen [66]. For example, Chen et al. [29] found that when chatbots were used in the classroom, some students only skimmed the learning content superficially rather than constructing their own thoughtful answers, and others even tended to engage in "smart loafing" in the classroom, handing the responsibilities for collaborative learning to the AI virtual assistant. Similarly, Akgun and Greenhow [10] and Klimova et al. [11] considered that algorithm-driven prediction systems and AI-driven mobile apps for education decrease the ability of students and teachers to manage their own lives, which may even lead to their conforming to norms in specific "data points."

### 5. Academic Misconduct

Academic misconduct is the least ethical concern in the reviewed studies ( $N = 5$ ). However, the misuse of AI technology has led to some academic misconduct issues. The automatic generation function of AI models, such as ChatGPT, may be used by students to cheat and plagiarize while completing assignments and participating in assessments, which devalues the efforts of others and thus produces unfairness [59]. For example, Adams et al. mentioned that with regard to student writing, the use of AI models has caused the boundaries of who is writing to begin to blur: the student or AI [67]. Therefore, the issue

of academic integrity caused by the misuse of technology must be paid special attention to, because it not only involves the ethical use of AI during teaching and learning, but may also lead to educational inequity.

### B. What Are the Ethical Principles of Using AI Models in the Classroom?

A count summary of ethical principle terms from the reviewed studies is presented in Table IV. Through content analysis, six ethical principles of using AI models in the classroom were summarized: fairness, privacy, transparency, accountability, autonomy, and beneficence.

TABLE IV. THE ETHICAL PRINCIPLES OF USING AI MODELS IN THE CLASSROOM

Category	Example	Na	Sample studies
Principle of fairness	Ensure that educational opportunities are equal among the students recommended by AI algorithms.	8	Matias and Zipitria [70]
	Ensure the accessibility of (digital) educational resources.	6	Köbis and Mehner [72]
	Inclusive of students from diverse backgrounds.	5	Schiff [66]
Principle of privacy	Keep the data provided by the students confidential.	9	Nguyen et al. [15]
	Acquire the students' active and full consent to access and use their personal data.	8	Masters [71]
Principle of transparency	Ensure that the educational decision-making process of AI models is explainable and understandable.	7	Chaudhry et al. [80]
	Specify the benefits, actual drawbacks, and possible risks of using AI models in the classroom.	4	Memarian and Doleck [17]
	Protect students' data ownership.	2	Nguyen et al. [15]
Principle of accountability	Open communication regarding the expectations of using AI models in the classroom.	2	Köbis and Mehner [72]
	Be responsible for the actions and decisions of using AI models.	9	Mouta et al. [81]
Principle of autonomy	Ensure teachers and students the right to access data.	4	Jang et al. [63]
	Teachers and students always maintain self-determination in deciding whether and how to adopt AI models.	8	Schiff [66]
Principle of beneficence	Provide comprehensive training before using AI models to enhance AI literacy.	4	Busch et al. [78]
	Support students' development and teacher well-being.	2	Adams et al. [75]

<sup>a</sup> The number of studies added up to more than 32 because multiple ethical principles of using AI models in the classroom are described in several studies.

### 1. Principle of Fairness

Fairness (or justice, or equity) is the most mentioned ethical principle in the reviewed studies ( $N = 19$ ). According to the results of the systematic literature review, fairness generally subsumes representation, accessibility, and inclusiveness. First, AI models must be designed, developed and deployed with non-discriminatory and unbiased data and algorithms to ensure representation and equality between different educational groups. For example, when AI models are applied to student services, such as admissions and financial aid, they should ensure that they do not exacerbate existing biases and discrimination based on race, class, gender, or socioeconomic status [41]. In addition, Nguyen et al. posited that infrastructure, skills and social acceptance should be taken into account when using AI models, allowing equitable access and use by all teachers and students [15]. Schiff considered that the AI tutoring system needs to fit the background of students, such as their local customs, cultural background and learning styles [66].

### 2. Principle of Privacy

Privacy is the second ethical principle of using AI models in the reviewed studies ( $N = 17$ ). First, the use of the AI-assisted tutor system in the classroom should protect students' personal information, such as gender, age, family address and mobile phone number, to avoid information leakage and personal harassment [76]. Moreover, Jang et al. further pointed out that students' data and privacy should be protected throughout the life cycle of using AI models, both in terms of raw data provided by students and new data generated about students during the interactions with AI systems (such as learning outcome analysis and recommendations) [63]. On the other hand, when collecting data about students, for whatever reason, it should be ensured that the student is giving active and not passive consent to the collection of personal data [71]. Meanwhile, Hong et al. considered that the use of AI models should also obtain full informed consent on how personal information and data are collected, shared and used [43].

### 3. Principle of Transparency

Transparency is the third ethical principle in the reviewed studies ( $N = 15$ ). According to the results of the systematic literature review, the principle of transparency mainly includes interpretability, traceability, data ownership and communication. First, Holmes et al. emphasized that teachers and students should be provided with detailed explanations of the rationale, operational processes and outcomes of using AI models, so that they can better understand and apply the results [16]. For instance, when using AI models to make teacher ratings, student evaluations, and other educational decisions, the process, results and application condition of AI algorithm decisions must be explained in detail. Moreover, in addition to displaying the benefits of AI models, teachers and students must be informed of the actual drawbacks and potential risks of using AI models in the classroom, and even remedial suggestions [41]. Additionally, Nguyen et al. argued that data ownership, which relates to who owns and has access to students' personal data, is an important aspect of the principle of transparency [15]. The open communication regarding the expectations of using AI models in the classroom is also considered essential to promote trust [72].

### 4. Principle of Accountability

Accountability is the fourth ethical principle in the studies reviewed ( $N = 13$ ). This principle requires responsibility for the actions and decisions of using AI models in the classroom and ensures that teachers and students have the right to access their data. Celik considered that teachers need to understand who the developers responsible for the design and decision-making of AI models are [77]. In addition, Hong et al. suggested that the principle of accountability can also be

considered as the capacity to verify actions and decisions, so teachers and students must be provided with the right to own and control how AI models are used to facilitate their own teaching and learning [43]. Therefore, it is necessary to clearly state the acknowledgment and responsibility of the actions of every relevant person involved in using AI models.

### 5. Principle of Autonomy

Autonomy is the fifth ethical principle in the studies reviewed ( $N = 8$ ). According to the literature reviewed, the principle of autonomy is generally associated with these key words, such as freedom, self-determination, independence, and empowerment. For example, Busch et al. emphasized that AI models should be considered as an addition to teaching and learning, rather than completely replacing traditional teaching materials and approaches, so that teachers and students can decide at any time whether or not to apply AI models [78]. Köbis and Mehner believed that it is essential to ensure the decisions made when using AI models in the classroom are aligned with human values and prevent compromising human independence [72]. Therefore, learner-centered use of AI models must be cultivated to strengthen students' authority and autonomy over their own learning.

### 6. Principle of Beneficence

Beneficence is the sixth ethical principle in the reviewed studies ( $N = 6$ ). In the context of using AI models in the classroom, the principle of beneficence is always described in terms of providing appropriate training about AI applications, benefiting the development of students, and promoting the well-being of teachers. First, training courses on using AI models should cover knowledge, skills, and ethical considerations to improve the AI literacy of teachers and students [79]. Busch et al. considered that proper education and training on using AI models can not only effectively integrate AI into the classroom, but can also foster the AI literacy of students and teachers, and enhance autonomy and justice [78]. In addition, the use of AI models must meet the developmental needs of students and stay consistent with the educational goals [66]. Similarly, teacher well-being was also considered an important principle in Adams et al. [75] and Adams et al. [73]; it refers to the needs and the physical and mental health of teachers faced with the challenge of using AI models in the classroom.

### C. What Are Future Directions of Research and Practice of Related Research?

Continuous discussions are required to comprehensively understand, prevent and overcome the ethical implications of AI model's use in the classroom. Table V displays the proposed future research and practice directions regarding the ethical considerations about the use of AI in the classroom.

First, while AI models have the capability to revolutionize education, it also raises a number of ethical implications. The results of this literature analysis show that the ethical implications of the use of AI in the classroom are not limited to bias and data privacy disclosure, but are also related to the ethical implications of reduced autonomy and academic integrity (see Table 2). Although these implications are mentioned or discussed in the existing literature, many of them have not been studied in detail. Hence, further substantive discussion, understanding and solution of these implications are required (see Table 5). On the one hand, educational decisions made by algorithm-based AI predictive models lead to bias and discrimination, but it is not clear what algorithmic features and attributes are needed to reduce such data bias. Future work should continue to optimize educational AI predictive models by training them using unbiased data. On the other hand, when AI is applied in the classroom, the algorithm-driven education prediction systems and AI-driven mobile learning apps decrease students' autonomy. In the future, learner-centered use of

TABLE V. THE PROPOSED FUTURE DIRECTIONS OF RESEARCH AND PRACTICE OF RELATED RESEARCH

Implications	Direction for research	Guideline for practice
The ethical implications of using AI models in the classroom need to be addressed further.	What algorithmic features and attributes are needed to reduce data bias in AI prediction models?	Continuously optimize educational AI prediction models by training them with unbiased data.
	How can student autonomy in the use of AI models be maintained?	Cultivate learner-centered use of AI models. Teaching AI and ethics lessons in educational contexts.
The ethical principles of using AI models in the classroom lack elucidation of the precise mechanism.	How do the ethical principles lead to ethical functioning of using AI models?	Strong policy guidance for educators is needed.
	How can the ethical principles be integrated into teachers' teaching practice?	Adopt more pedagogical responsive and context-sensitive ethical approaches in the use of AI models.
The ethical implications of using AI models in the classroom lack accurate assessment.	How are the ethical implications of AI models use investigated in the classroom?	Various approaches, such as case studies or interviews, and more clear-cut empirical research is required.
		Clarify the definition of ethics of using AI models in the classroom.
	What should be considered to evaluate the ethical implications of AI models use in the classroom?	Positive and negative impacts should both be considered. The precise needs of the stakeholders should be taken into account.

AI models should be cultivated, and AI ethics courses are required in educational settings to learn about the ethical use of AI [82].

Second, in fact, most of the ethical principles discussed in the literature are more applicable to general AI systems or computing and design environments, and there is a lack of research on ethical principles for specific use cases in the classroom [15]. Thus, the precise mechanism of ethical principles of using AI models remains to be elucidated and extended to the implementation phase in the classroom. For example, how ethical principles lead to the ethical function of using AI models in practice remains ambiguous, and more robust policy guidance for educators is needed. Additionally, how to integrate the ethical principles of using AI models into the teaching practice of teachers remains to be further explored. More pedagogical responsive and context-sensitive ethical approaches should be designed and adopted in the use of AI models, to avoid and address these ethical implications [73].

Third, the ethical implications need to be more accurate assessment. On the one hand, how to investigate the ethical implications when applying AI model's during teaching and learning is a direction that needs further research. The existing discussion on the ethical implications about the AI models use is mostly descriptive research, and more clear empirical research is required [63],[11]. Moreover, future research needs to clarify the ethical definition of using AI models in the classroom to help identify the ethical implications. On the other hand, future work will be necessary to develop or customize ethical implication assessments for specific AI models use cases in classroom contexts. Thus, not only both positive and negative effects, but also the precise needs of the relevant stakeholders should be considered when assessing the ethical implications of AI models use in the classroom.

## V. DISCUSSION

### A. Five Ethical Implications of Using AI Models in the Classroom

The first research question identifies the five major ethical implications of the use of AI models in the classroom, namely algorithmic bias and discrimination, data privacy leakage, lack of transparency, decreased autonomy, and academic misconduct. First, in terms of algorithmic bias and discrimination, although the main promise of AI models is to improve the objectivity and accuracy

of instruction, the fact is that when AI is applied in the classroom, these inherent social biases, discrimination, and power structures are naturally embedded in them, and are even further perpetuated and exacerbated [10]. Masters emphasized that there is no such thing as ethically neutral AI, as all AI models react and make decisions that favor specific groups, leading to bias and discrimination in the classroom [71]. Second, another ethical implication surrounding the use of AI models in the classroom is data privacy leakage. The disclosure of personal information, surveillance and student tracking, lack of informed consent, and the right to be forgotten were often considered in the use of AI models. Previous review work has also identified the ethical implications of data privacy leakage [10], but in this study, the absence of students' right to be forgotten was further reviewed. Through the review, this study found that students lack the chance to choose whether to enter the educational AI system or not, but also lack the chance to opt out of the system. Third, in terms of the lack of transparency, when using AI models in the classroom, there is not only no clear explanation of the process and results, but also no detailed explanation of the actual shortcomings and potential risks. The latter, in particular, has not been discussed in great detail, but it does in fact exist [41]. In particular, while AI models provide personalized learning for students, they also have the problem of not always being able to understand the open-ended needs of students. Fourth, in terms of reduced autonomy for teachers and students, algorithm-based forecasting and decision-making systems and inappropriate delegation of authority have led to this ethical implication. Therefore, it is essential to consider the long-term consequences of using AI models for students' learning and cognitive abilities. Fifth, academic misconduct also emerges when AI technologies are misused by students. However, a previous review study has paid less attention to this ethical implication [10]. Therefore, this study extended the previous review work.

### B. Six Ethical Principles of Using AI Models in the Classroom

The second research question revealed six ethical principles of using AI models in the classroom, namely fairness, privacy, transparency, accountability, autonomy and beneficence.

First, the principle of fairness, as the most mentioned ethical principle in the review study, requires representation, accessibility, and inclusiveness of using AI models, in order to achieve algorithmic processes and results without discrimination or bias for students and teachers [41]. As noted in the previous section, when AI models fail to



understand the needs of underrepresented students, such as minority students, this group of students may already feel marginalized. Hence, unbiased data training for AI models is recommended. Surprisingly, however, the AI Act recently approved by the EU did not even mention the principle of “fairness,” but explicitly mentioned the term “non-discrimination.” Actually, the “fairness” in the Act is relatively hidden, and the expression “non-discrimination” is intended to reflect specific regulatory objectives, because it has a more specific measure than the concept of fairness [83]. Therefore, the principle of fairness in this study is essentially consistent with the term “non-discrimination” in the Act.

Second, the principle of privacy calls for the protection of personal data and information of teachers and students in the use of AI models. It is worth noting that, based on the perspective of the AI life cycle, Jang et al. further pointed out the need to protect new data generated in the use of AI models [63]. This was not examined in the previous literature review work [10]. In addition to highlighting the full life cycle of AI, at the legal level, the EU’s AI Act protects personal privacy by assessing the categories of AI risks. For example, the Act classifies the use of “real-time” remote biometrics in public places for law enforcement purposes as high risk. The practice is prohibited because it poses a great risk to an individual’s private life [84]. Therefore, it is suggested that in the classroom, both aspects of the full life cycle of AI models and risk categories need to be considered to protect personal privacy.

Third, the principle of transparency mainly includes interpretability, traceability, data ownership and communication. Similar to the definition of transparency in the AI Act proposed by the EU [85], transparency here is meant not just as an algorithmic attribute, but as a means of supporting broader and different values. This act further distinguishes among technical, enabling and protective transparency. In particular, in addition to presenting the conditions, process, and results of using AI models in detail, the actual shortcomings and potential risks of using AI models in the classroom should be clearly stated, and even relevant remedial suggestions should be made [16]. It is worth noting that although some of the complex AI models, such as deep learning neural networks, have techniques for interpreting and proving results, there is still a need to customize different interpretations for different audiences [29]. In the field of education, it is necessary to provide detailed explanations for the use of AI models to teachers and students.

Fourth, the principle of accountability is closely related to the previous principle of transparency, and both principles are mentioned simultaneously in multiple studies [11] [17]. Canada’s AI and Data Act also created a strong link between accountability, transparency and privacy provisions [86]. This means that individuals who use AI models responsibly also have an obligation to be transparent and provide data subjects with an explanation of the information intended or actually used by the AI model. In this Act, the principle of accountability specifically emphasizes responsible anonymization of data. In fact, accountability focuses more on requiring the establishment of mechanisms to ensure responsibility and accountability for AI models before and after their use [63]. The EU’s AI Act adopted different regulatory measures and accountability based on classifying different risk levels of using AI [87]. Therefore, this study recommends that systems of responsibility for the possible consequences of using AI models should be developed and implemented, to clarify the obligations of teachers and students in the classroom, and especially to protect their privacy rights.

Fifth, the principle of autonomy emphasizes that teachers and students have the ability and right to act in accordance with their own interests and values, despite being under the monitoring system of AI. In previous studies, it is included in the human-centered principle, but

these studies all emphasize human values in the use of AI [11] [15]. From a legal point of view, the US AI Bill of Rights also mentioned that the use of AI must follow this principle: where appropriate, individuals can voluntarily opt out of the automated system and choose a human alternative [18]. However, the expression of the term “where appropriate” is vague and subject to different interpretations. Therefore, the boundary and degree of autonomy should be clearly defined. When teachers apply AI models to assist classroom teaching, the key is to reasonably design learning materials and tasks, and to consider in what dimensions and to what extent students’ autonomy can be guaranteed, so as to avoid reducing students’ learning efforts and their learning autonomy.

Finally, the principle of beneficence calls for attention to the sustainable development of teachers and students when using AI models. Importantly, this principle emphasizes appropriate education and training on AI for teachers and students, which would help students critically understand AI and promote the development of teachers’ intellectual competence [65]. Hence, specialized AI ethics courses and lectures on improving AI literacy for teachers and students are suggested. This is similar to the “social and environmental well-being” mentioned in the EU’s AI Act, which refers to the idea that AI should be developed and used in a sustainable and environmentally friendly way, while monitoring and assessing the long-term impacts on individuals and society [88]. Slightly different, training to enhance individual AI literacy is not mentioned in the Act, which focuses more on sustainable considerations in the development and use of AI.

### *C. Future Directions of Research and Practice of Related Research*

The third research question concerns the main future directions of research and practice regarding the ethical implications and principles of AI models use in the classroom. Firstly, due to the lack of research on the specific solutions to the ethical implications about the use of AI in the existing literature, most of the discussions remain at the macro level, and so further exploration is needed in the future. From a technical point of view, how to train unbiased algorithms and what characteristics they should have needs to be explored. From a teaching point of view, how teachers can maintain students’ autonomy when using AI models, and how teaching materials should be properly designed should be examined. On the one hand, it is recommended that AI ethics courses and lectures support students’ autonomous development when using AI models. This has also been considered in previous studies [82]. However, it is further suggested that when teachers design teaching materials, they should allow ample opportunity for students’ autonomous development, and some traditional classroom teaching is still valuable, such as class discussion, rather than relying entirely on AI models. Secondly, future research needs to further elucidate the precise mechanisms of ethical principles of using AI models in the classroom, and extend them to the implementation phase because the ethical principles discussed in the existing literature lack research on specific classroom use cases. This includes how ethical principles are translated into ethical functions and how they are integrated into teachers’ instruction practices, all of which are unclear. In fact, more ethical principles in the context of teaching situations should be explored, which have not been mentioned in previous studies. Thus, future research should be based on different classroom types, such as online classes and flipped classes, to conduct different specific discussions. Finally, the accurate assessment of the ethical implications of AI use in the classroom is required, because what and how to assess it remains unclear, and more empirical research is called for. This finding echoes Memarian and Doleck’s research [17], which reviewed the existing investigation methods of ethical implications and revealed the deficiency of quantitative research methods. However, this study further reported that ethical implication assessments for the future focus on teaching

and learning in the classroom, and developing or customizing ethical implication assessments for specific AI models use cases in classroom contexts. Due to the complexity of the situations presented by the real classroom, the evaluation of the ethical implications of using AI models should consider various factors, especially when it comes to sensitive topics such as student privacy aspects.

## VI. CONCLUSIONS

### A. Implications

The main theoretical contribution of this study is to outline the five ethical implications (including algorithmic bias and discrimination, data privacy leakage, lack of transparency, decreased autonomy, and academic misconduct), six ethical principles (including fairness, privacy, transparency, accountability, autonomy and beneficence), and the main future research directions and practices of the related research. This structure stems from a systematic review that helps to understand and conceptualize practice and research of using AI models ethically in the classroom. Additionally, this review is conducive to validating some less explored areas to help researchers determine the direction of future research efforts on the ethical implications of the AI models use in the classroom, for example, the strategic and evaluation study of the ethical implications, which still remain less researched. Meanwhile, some specific guidance schemes are provided in this study. On a practical level, this research helps educators and learners to understand which behaviors are ethical when using AI models for education-related purposes, which could lead to the implementation of appropriate regulation. Importantly, the study provides a detailed elaboration of ethical principles and practical recommendations to better promote the ethical use of AI models in the classroom.

### B. Limitations and Future Work

However, several limitations of this systematic review must be acknowledged. First, the literature reviewed in this study mainly comes from two databases. Future research can consider other databases, such as Science Direct and Google Scholar, to retrieve suitable papers. In addition, during the eligibility phase of this systematic review, 32 articles were excluded because the full text was not available. Finally, since the articles reviewed in this study are mainly from Europe and North America, most represent Western perspectives. Therefore, there should be further reviews of the research from other continents or in other languages to gain a broader understanding of the ethical implications of AI model use in the classroom.

## ACKNOWLEDGMENT

This study was supported by the Taiwan Comprehensive University System (TCUS) and the National Science and Technology Council, Taiwan, under grant NSTC 111-2410-H-019-006-MY3 and NSTC 111-2423-H-153-001-MY3.

## REFERENCES

- [1] S. K. Das, A. Dey, A. Pal, N. B. Roy, "Applications of Artificial Intelligence in Machine Learning: Review and Prospect," *International Journal of Computer Applications*, vol. 115, pp. 31-41, 2015, doi:10.5120/20182-2402.
- [2] W. Holmes, M. Bialik, C. Fadel, "Artificial Intelligence in Education: Promises and Implications for Teaching and Learning," Boston, MA: Center for Curriculum Redesign, 2019.
- [3] J. M. Flores-Vivar, F. J. García-Peñalvo, "Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4)," *Comunicar*, vol. 31, no. 74, pp. 37-47, 2023, doi:10.3916/C74-2023-03.
- [4] J. A. Kumar, B. Bervell, S. Osman, "Google classroom: insights from Malaysian higher education students' and instructors' experiences." *Education and Information Technologies*, vol. 25, no. 5, pp. 4175-4195, 2020, doi:10.1007/s10639-020-10163-x.
- [5] R. Reyes, D. Garza, L. Garrido, V. De la Cueva, J. Ramirez, "Methodology for the Implementation of Virtual Assistants for Education Using Google Dialogflow," In: Martínez-Villaseñor, L., Batyrshin, I., Marín-Hernández, A. (eds) *Advances in Soft Computing. MICAI 2019*. Lecture Notes in Computer Science, Springer, Cham, vol. 11835, pp. 440-451, 2019, doi:10.1007/978-3-030-33749-0\_35.
- [6] K. Naidu, K. Sevnarayan, "ChatGPT: An ever-increasing encroachment of artificial intelligence in online assessment in distance education," *Online Journal of Communication and Media Technologies*, vol. 13, no. 3, 2023, doi:10.30935/ojcm/13291.
- [7] M. Zafari, J. S. Zafari, A. Sadeghi-Niaraki, C. M. Choi, "Artificial Intelligence Applications in K-12 Education: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 61905-61921, 2022, doi: 10.1109/ACCESS.2022.3179356.
- [8] R. F. Murphy, "Artificial intelligence applications to support k-12 teachers and teaching: a review of promising applications, challenges, and risks," *Perspective*, pp. 1-20, 2019, doi:10.7249/PE315.
- [9] M. Karabacak, B. B. Ozkara, K. Margetis, M. Wintermark, S. Bisdas, "The Advent of Generative Language Models in Medical Education," *JMIR Medical Education*, vol. 9, 2023, doi: 10.2196/48163.
- [10] S. Akgun, C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in K-12 settings," *AI and Ethics*, vol. 2, no. 3, pp. 431-440, 2022, doi:10.1007/s43681-021-00096-7.
- [11] B. Klimova, M. Pikhart, J. Kacetl, "Ethical issues of the use of AI-driven mobile apps for education," *Frontiers in Public Health*, vol. 10, no. 1118116, 2023, doi:10.3389/fpubh.2022.1118116.
- [12] N. Selwyn, "Debería los robots sustituir al profesorado? La IA y el futuro de la educación," *Ediciones Morata*, 2019. Available: <https://bit.ly/3zxyPmO>.
- [13] F. J. García Peñalvo, F. Llorens-Largo, J. Vidal, J. "The new reality of education in the face of advances in generative artificial intelligence," *ITEN - Ibero-American Journal of Distance Education*, vol. 27, no. 1, pp. 9-39, doi:10.5944/ried.27.1.37716.
- [14] F. J. García-Peñalvo, A. Vázquez-Ingelmo, "What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 7-16, doi:10.9781/ijimai.2023.07.006.
- [15] A. Nguyen, H. N. Ngo, Y. Hong, Dang, Belle, B. P. T. Nguyen, "Ethical principles for artificial intelligence in education," *Education and Information Technologies*, vol. 28, pp. 4221-4241, 2023, doi:10.1007/s10639-022-11316-w.
- [16] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. B. Shum, O. C. Santos, M. T. Rodrigo, M. Cukurova, I. I. Bittencourt, K. R. Koedinger, "Ethics of AI in Education: Towards a Community-Wide Framework," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 3, pp. 504-526, 2021, doi:10.1007/s40593-021-00239-1.
- [17] B. Memarian, T. Doleck, "Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 5, no. 100152, 2023, doi:10.1016/j.caeai.2023.100152.
- [18] E. Hine, L. Floridi, "The Blueprint for an AI Bill of Rights: In Search of Enaction, at Risk of Inaction," *Minds & Machines*, vol. 33, no. 2, pp. 285-292, 2023, doi:10.1007/s11023-023-09625-1.
- [19] A. E. Muhammad, K. -C. Yow, "Demystifying Canada's Artificial Intelligence and Data Act (AIDA): The good, the bad and the unclear elements," *2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Regina, SK, Canada, pp. 510-515, 2023, doi: 10.1109/CCECE58730.2023.10288878.
- [20] "EU AI Act: First regulation on artificial intelligence," European Parliament: News, 8 June 2023; [www.europarl.europa.eu/news/en/headlines/society/20230610STO93804/eu-ai-act-first-regulation-on-artificial-intelligence](http://www.europarl.europa.eu/news/en/headlines/society/20230610STO93804/eu-ai-act-first-regulation-on-artificial-intelligence).
- [21] European Parliament (2023) MEPs ready to negotiate first-ever rules for safe and transparent AI. European Parliament. 14.07.2023. Available at: <https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai>

- [22] F. Ouyang, P. Jiao, "Artificial Intelligence in Education: The Three Paradigms," *Computers & Education: Artificial Intelligence*, vol. 100020, 2021, doi:10.1016/j.caeai.2021.100020.
- [23] K. Y. Tang, C. Y. Chang, G. J. Hwang, "Trends in artificial intelligence supported e-learning: A systematic review and co-citation network analysis (1998-2019)," *Interactive Learning Environments*, vol. 31, no. 4, pp. 2134-2152, 2021, doi.org/10.1080/10494820.2021.1875001.
- [24] F. Tahiru, "AI in Education: A Systematic Literature Review," *Journal of Cases on Information Technology*, vol. 23, no. 1, pp. 1-20, 2021, doi:10.4018/JCIT.2021010101.
- [25] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, Prisma Group, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *PLoS medicine*, vol. 6, no. 7, pp. e1000097, 2009, doi:10.1371/journal.pmed.1000097.t001.
- [26] L. Zhang, P. P. Wei, Y. W. Zhang, N. Wang, "Artificial Intelligence and Edge Computing Technology Promote the Design and Optimization of Flipped Classroom Teaching Models for Higher Vocational, Ideological, and Political Courses," *Mobile Information Systems*, vol. 2022, doi:10.1155/2022/5385386.
- [27] A. Suresh, J. Jacobs, C. Clevenger, V. Lai, C. H. Tan, J. H. Martin, T. Sumner, "Using AI to Promote Equitable Classroom Discussions: The TalkMoves Application," *22nd International Conference on Artificial Intelligence in Education (AIED)*, vol. 12749, pp. 344-348, 2021, doi:10.1007/978-3-030-78270-2\_61.
- [28] M. Karabacak, B. B. Ozkara, K. Margetis, M. Wintermark, S. Bisdas, "The Advent of Generative Language Models in Medical Education," *JMIR Medical Education*, vol. 9, 2023, doi:10.2196/48163.
- [29] Y. Chen, S. Jensen, L. J. Jensen, et al. "Artificial Intelligence (AI) Student Assistants in the Classroom: Designing Chatbots to Support Student Success," *Information Systems Frontiers*, vol. 25, no. 1, pp. 161-182, 2023, doi:10.1007/s10796-022-10291-4.
- [30] X. Chen, H. Xie, G. J. Hwang, "A multi-perspective study on artificial intelligence in education: grants, conferences, journals, software tools, institutions, and researchers," *Computers and Education: Artificial Intelligence*, vol. 1, no. 100005, 2020, doi:10.1016/j.caeai.2020.100005.
- [31] B. C. L. Christudas, E. Kirubakaran, P. R. J. Thangaiyah, "An evolutionary approach for personalization of content delivery in e-learning systems based on learner behavior forcing compatibility of learning materials," *Telematics and Informatics*, vol. 35, no. 3, pp. 520-533, 2018, doi:10.1016/j.tele.2017.02.004.
- [32] G. George, A. M. Lal, "Review of ontology-based recommender systems in e-learning," *Computers & Education*, vol. 142, no. 7, pp. 103642, 2019, doi:10.1016/j.compedu.2019.103642.
- [33] S. Verma, "Weapons of math destruction: how big data increases inequality and threatens democracy," *Vikalpa*, vol. 44, no. 2, pp. 97-8, 2019, doi:10.1177/0256090919853933.
- [34] P. M. Regan, J. Jesse, "Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking," *Ethics and Information Technology*, vol. 21, pp. 167-179, 2019, doi:10.1007/s10676-018-9492-2.
- [35] M. J. Reiss, "The use of AI in education: Practicalities and ethical considerations," *London Review of Education*, vol. 19, no. 15, pp. 1-14, 2021, doi:10.14324/LRE.19.1.05.
- [36] K., Naidu, K. Sevnanarayan, "ChatGPT: An ever-increasing encroachment of artificial intelligence in online assessment in distance education," *Online Journal of Communication and Media Technologies*, vol. 13, no. 3, 2023, doi: 10.30935/ojcm/13291.
- [37] N. R. Möllmann, M., Mirbabaie, S. Stieglitz, "Is it alright to use artificial intelligence in digital health? A systematic literature review on ethical considerations," *Health Informatics Journal*, vol. 27, no. 4, pp. 1-17, 2021, doi:10.1177/14604582211052391.
- [38] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99-120, 2020, doi:10.1007/s11023-020-09517-8.
- [39] F. Miao, W. Holmes, R. Huang, H. Zhang, "AI and education: Guidance for policy-makers," *United Nations Educational, Scientific and Cultural Organization*, 2021. Available: <https://unesdoc.unesco.org/ark:/48223/pf0000376709>.
- [40] European Commission, "The European Commission's high-level expert group on artificial intelligence: Ethics guidelines for trustworthy AI," *European Union Publications Office*, 2019 Available: <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>
- [41] S. Larsson, F. Heintz, "Transparency in artificial intelligence," *Internet Policy Review*, vol. 9, no. 2, pp. 1-16, 2020, doi:10.14763/2020.2.1469.
- [42] Z. Slimi, B. V. Carballido, "Navigating the Ethical Challenges of Artificial Intelligence in Higher Education: An Analysis of Seven Global AI Ethics Policies," *TEM Journal-Technology Education Management Informatics*, vol. 12, no. 2, pp. 590-602, 2023, doi:10.18421/TEM122-02.
- [43] Y. Hong, A. Nguyen, B. Dang, B. P. T. Nguyen, "Data Ethics Framework for Artificial Intelligence in Education (AIED)," *International Conference on Advanced Learning Technologies (ICALT)*, pp. 297-301, 2022, doi:10.1109/ICALT55010.2022.00095.
- [44] D. Gough, S. Oliver, J. Thomas, "An introduction to systematic reviews, (2nd ed.,)," Los Angeles: SAGE, 2017.
- [45] C. Guan, J. Mou, Z. Jiang, "Artificial intelligence innovation in education: A twenty-year data-driven historical analysis," *International Journal of Innovation Studies*, vol. 4, no. 4, pp. 134-147, 2020, doi:10.1016/j.ijis.2020.09.001.
- [46] A. A. Chadegani, H. Salehi, M. M. Yunus, H. Farhadi, M. Fooladi, M. Farhadi, N. A. Ebrahim, "A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases," *Asian Social Science*, vol. 9, no. 5, pp. 18-26, 2013, doi:10.5539/ass.v9n5p18.
- [47] J. F. Burnham, "Scopus database: a review," *Biomedical digital libraries*, vol. 3, pp. 1-8, 2006, doi:10.1186/1742-5581-3-1.
- [48] D. T. K. Ng, M. Lee, R. J. Y. Tan, et al. "A review of AI teaching and learning from 2000 to 2020," *Education and Information Technologies*, vol. 28, pp. 8445-8501, 2023, doi:10.1007/s10639-022-11491-w.
- [49] D. Nicholas, A. Watkinson, H. R. Jamali, E. Herman, C. Tenopir, R. Volentine, et al. "Peer review: still king in the digital age," *Learned Publishing*, vol. 28, no. 1, pp. 15-21, 2015, doi:10.1087/20150104.
- [50] C. Lefebvre, J. Glanville, S. Briscoe, A. Littlewood, C. Marshall, M.-I. Metzendorf, A. Noel-Storr, T. Rader, F. Shokrane, J. Thomas, L. S. Wieland, "Chapter 4: Searching for and selecting studies," In H. JPT, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions (Version 6.2)*, 2021. Available:<https://www.training.cochrane.org/handbook>.
- [51] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," *In Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pp. 1-10, 2014, doi:10.1145/2601248.2601268.
- [52] M. Petticrew, H. Roberts, "Systematic reviews in the social sciences," 11, Blackwell Publishing Ltd, Oxford, UK, 2006, doi:10.1002/9780470754887.
- [53] J. Webster, R. T. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly*, vol. 26, no. 2, pp. xiii-xxiii, 2022, doi:10.2307/4132319.
- [54] V. Braun, V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77-101, 2006, doi:10.1191/1478088706qp0630a.
- [55] A. Strauss, J. Corbin, "Grounded theory methodology: An overview," In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 273-285). 1995, SAGE
- [56] U. H. Graneheim, B. Lundman, "Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness," *Nurse Education Today*, vol. 24, no. 2, pp. 105e112, 2004, doi:10.1016/j.nedt.2003.10.001.
- [57] H. F. Hsieh, S. E. Shannon, "Three approaches to qualitative content analysis," *Qualitative Health Research*, vol. 15, no. 9, pp. 1277-1288, 2005, doi:10.1177/1049732305276687.
- [58] E. Dieterle, C. Dede, M. Walker, "The cyclical ethical effects of using artificial intelligence in education," *AI & Society*, 2022, doi:10.1007/s00146-022-01497-w.
- [59] C. Kooli, "Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions," *Sustainability*, vol. 15, no. 7, pp. 5614, 2023, doi:10.3390/su15075614.
- [60] A. Latham, S. Goltz, "A Survey of the General Public's Views on the Ethics of Using AI in Education," *International Conference on Artificial Intelligence in Education (AIED)*, vol. 11625, pp. 194-296, 2019, doi:10.1007/978-3-030-23204-7\_17.
- [61] E. Kowch, "Ethics to Prepare teachers for Professional Service Robots in Classrooms," *International Joint Conference on Information, Media and*



- Engineering (IJCIME), 2021, doi:10.1109/IJCIME49369.2019.0010.
- [62] W. Holmes, F. Iniesto, S. Anastopoulou, J. Boticario, "Stakeholder Perspectives on the Ethics of AI in Distance-Based Higher Education," *The International Review of Research in Open and Distributed Learning*, vol. 24, pp. 96-117, 2023, doi:10.19173/irrodl.v24i2.6089.
- [63] Y. Jang, S. Choi, H. Kim, "Development and validation of an instrument to measure undergraduate students' attitudes toward the ethics of artificial intelligence (AT-EAI) and analysis of its difference by gender and experience of AI education," *Education and Information Technologies*, vol. 27, pp. 11635-11667, 2022, doi:10.1007/s10639-022-11086-5.
- [64] R. Farrow, "The possibilities and limits of XAI in education: a socio-technical perspective," *Learning, Media and Technology*, vol. 48, no. 2, pp. 266-279, 2023, doi: 10.1080/17439884.2023.2185630.
- [65] I. Tuomi, "A Framework for Socio-Developmental Ethics in Educational AI," *In the Proceedings of the 56th Hawaii International Conference on System Sciences*, 2022, doi:10.13140/RG.2.2.36133.58089.
- [66] D. Schiff, "Education for AI, not AI for Education: The Role of Education and Ethics in National AI Policy Strategies," *International Journal of Artificial Intelligence in Education*, vol. 32, pp. 527-563, 2022, doi:10.1007/s40593-021-00270-2.
- [67] C. Adams, P. Pente, G. Lernermeier, J. Turville, G. Rockwell, "Artificial Intelligence and Teachers' New Ethical Obligations," *The International Review of Information Ethics*, vol. 31, no. 1, 2022, doi:10.29173/irie483.
- [68] N. Ghotbi, M. T. Ho, "Moral Awareness of College Students Regarding Artificial Intelligence," *Asian Bioethics Review*, vol. 13, pp. 421-433, 2021, doi:10.1007/s41649-021-00182-2.
- [69] N. Ghotbi, M. T. Ho, P. Mantello, "Attitude of college students towards ethical issues of artificial intelligence in an international university in Japan," *AI & Society*, vol. 37, pp. 283-290, 2022, doi:10.1007/s00146-021-01168-2.
- [70] A. Matias, I. Zipitria, "Promoting Ethical Uses in Artificial Intelligence Applied to Education. In: Frasson, C., Mylonas, P., Troussas, C. (eds) Augmented Intelligence and Intelligent Tutoring Systems," *International Conference on Intelligent Tutoring Systems*, vol. 13891, pp. 604-615, 2023, doi:10.1007/978-3-031-32883-1\_53.
- [71] K. Masters, "Ethical use of Artificial Intelligence in Health Professions Education: AMEE Guide No. 158," *Medical Teacher*, vol. 45, no. 6, pp. 574-584, 2023, doi: 10.1080/0142159X.2023.2186203.
- [72] L. Köbis, C. Mehner, "Ethical Questions Raised by AI-Supported Mentoring in Higher Education," *Frontiers in Artificial Intelligence*, vol. 4, no. 624050, 2021, doi: 10.3389/frai.2021.624050.
- [73] C. Adams, P. Pente, G. Lernermeier, G. Rockwell, "Ethical principles for artificial intelligence in K-12 education," *Computers and Education: Artificial Intelligence*, vol. 4, no. 100131, 2023, doi:10.1016/j.caeai.2023.100131.
- [74] B. Han, S. Nawaz, G. Buchanan, and D. McKay, "Ethical and Pedagogical Impacts of AI in Education," *International Conference on Artificial Intelligence in Education (AIED)*, vol. 13916, pp. 667-673, 2023, doi:10.1007/978-3-031-36272-9\_54.
- [75] C. Adams, P. Pente, G. Lernermeier, G. Rockwell, "Artificial Intelligence Ethics Guidelines for K-12 Education: A Review of the Global Landscape," *International Conference on Artificial Intelligence in Education (AIED)*, vol. 12749, pp. 24-28, 2021, doi.org/10.1007/978-3-030-78270-2\_4.
- [76] L. H. Yu, Z. G. Yu, "Qualitative and quantitative analyses of artificial intelligence ethics in education using VOSviewer and CitNetExplorer," *Frontiers in Psychology*, vol. 14, no. 1061778, 2023, doi: 10.3389/fpsyg.2023.1061778.
- [77] I. Celik, "Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education," *Computers in Human Behavior*, vol. 138, no. 107468, 2023, doi:10.1016/j.chb.2022.107468.
- [78] F. Busch, L. C. Adams, K. K. Bressemer, "Biomedical Ethical Aspects Towards the Implementation of Artificial Intelligence in Medical Education," *Medical Science Educator*, vol. 33, no. 1007-1012, 2023, doi.org/10.1007/s40670-023-01815-x.
- [79] J. M. Flores-Vivar, F. J. García-Peñalvo, "Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4)," *Comunicar*, vol. 31, no.74, pp. 37-47, 2023, doi:10.3916/C74-2023-03.
- [80] M. Chaudhry, M. Cukurova, R. Luckin, "A Transparency Index Framework for AI in Education," *23rd International Conference on Artificial Intelligence in Education (AIED)*, Durham Univ, Durham, ENGLAND, vol. 13356, pp. 195-198, 2022, doi:10.1007/978-3-031-11647-6\_33.
- [81] A. Mouta, A. M. Pinto-Llorente, E. M. Torrecilla-Sánchez, "Blending machines, learning, sense of agency, and ethics: Designing an in-depth framework with Experts using the Delphi Method approach," *In Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)*, Association for Computing Machinery, New York, NY, USA, pp. 665-670, 2021, doi:10.1145/3486011.3486545.
- [82] T. Foltynnek, S. Bjelobaba, I. Glendinning, et al. "ENAI Recommendations on the ethical use of Artificial Intelligence in Education," *International Journal for Educational Integrity*, vol. 19, no. 1, 12, 2023, doi:10.1007/s40979-023-00133-4.
- [83] J. Pfeiffer, J. Gutschow, C. Haas, et al, "Algorithmic Fairness in AI," *Business & Information Systems Engineering*, vol. 65, no. 2, pp. 209-222, 2023, doi:10.1007/s12599-023-00787-x.
- [84] D. Svantesson, "The European Union Artificial Intelligence Act: Potential implications for Australia," *Alternative Law Journal*, vol. 47, no. 1, pp. 4-9, doi:10.1177/1037969X211052339.
- [85] B. Gyevnara, N. Fergusona, B. Schaferb, "Bridging the Transparency Gap:What Can Explainable AI Learn From the AI Act?," *European Conference on Artificial Intelligence*, vol. 372, pp. 964-971, doi:10.3233/FAIA230367.
- [86] A. E. Muhammad, K. -C. Yow, "Demystifying Canada's Artificial Intelligence and Data Act (AIDA): The good, the bad and the unclear elements," *2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Regina, SK, Canada, 2023, pp. 510-515, doi: 10.1109/CCECE58730.2023.10288878.
- [87] R. Paul, "European artificial intelligence 'trusted throughout the world': Risk-based regulation and the fashioning of a competitive common AI market," *Regulation & Governance*, 2023, doi:10.1111/rego.12563.
- [88] [K. Kalodanis, P. Rizomiliotis, D. Anagnostopoulos, "European Artificial Intelligence Act: an AI security approach," *Information and Computer Security*, vol. ahead-of-print, no. ahead-of-print, 2023, doi:10.1108/ICS-10-2022-0165.



Lin Tang

Lin Tang is a graduate student at Nanjing Normal University in China. She has experience in data mining, data analysis and text analysis. Her research interests are now focused on deep learning, modern educational technology, technical education, and other artificial intelligence applications in education.



Yu-Sheng Su

Dr. Yu-Sheng Su is currently an associate professor of the Department of Computer Science and Information Engineering at National Chung Cheng University, Taiwan. His research interests include Cloud Computing, Big Data Analytics, Intelligent Systems, and the Metaverse.

# A Trustworthy Automated Short-Answer Scoring System Using a New Dataset and Hybrid Transfer Learning Method

Martinus Maslim<sup>1,2</sup>, Hei-Chia Wang<sup>1,3\*</sup>, Cendra Devayana Putra<sup>1</sup>, Yulius Denny Prabowo<sup>4</sup>

<sup>1</sup> National Cheng Kung University (Taiwan)

<sup>2</sup> Universitas Atma Jaya Yogyakarta (Indonesia)

<sup>3</sup> Center for Innovative FinTech Business Models, National Cheng Kung University (Taiwan)

<sup>4</sup> Bina Nusantara University (Indonesia)

Received 1 September 2023 | Accepted 22 January 2024 | Published 5 February 2024



## ABSTRACT

To measure the quality of student learning, teachers must conduct evaluations. One of the most efficient modes of evaluation is the short answer question. However, there can be inconsistencies in teacher-performed manual evaluations due to an excessive number of students, time demands, fatigue, etc. Consequently, teachers require a trustworthy system capable of autonomously and accurately evaluating student answers. Using hybrid transfer learning and student answer dataset, we aim to create a reliable automated short answer scoring system called Hybrid Transfer Learning for Automated Short Answer Scoring (HTL-ASAS). HTL-ASAS combines multiple tokenizers from a pretrained model with the bidirectional encoder representations from transformers. Based on our evaluation of the training model, we determined that HTL-ASAS has a higher evaluation accuracy than models used in previous studies. The accuracy of HTL-ASAS for datasets containing responses to questions pertaining to introductory information technology courses reaches 99.6%. With an accuracy close to one hundred percent, the developed model can undoubtedly serve as the foundation for a trustworthy ASAS system.

## KEYWORDS

Automated Short Answer Scoring, Hybrid Transfer Learning, Student Answer Dataset, Trustworthy System.

DOI: 10.9781/ijimai.2024.02.003

## I. INTRODUCTION

THE objective of schools is to educate students through the teaching of academic subjects. To determine the quality of schools and students, it is crucial to measure student competencies [1]. Student competencies can be evaluated by analyzing the outcomes of student learning. The quality of learning is established through the assessment and test of outcomes [2]-[4]. Assessments and evaluations measure students' knowledge and proficiency in each subject [5], [6]. A reliable assessment tool reveals not only the students performing inadequately but also the areas where they will succeed in the future [7]. The assessment process helps teachers analyze patterns in student errors. Teachers can use information from assessments to correct students and advise them about their errors in future classes, and students can subsequently learn from their mistakes [8]. Assessments are supported by various inquiry-based grading approaches and diverse question forms [2].

Some question formats, such as essay, multiple-choice, and short-answer, can be employed to assess the level of student comprehension [7], [9], [10]. Essay writing assessments are critical in gauging the logical reasoning, critical thinking, and foundational writing

proficiencies of students [11]. While multiple-choice questions do prove to be an effective approach for assessing a considerable quantity of students, they are most suitable for evaluating knowledge and skills that are specific, well-defined, and often discrete [12], [13]. On the other hand, short-answer questions are a highly effective evaluative tool; they enable teachers to gauge students' comprehension of a subject matter through the provision of concise textual responses [14], [15]. Short-answer questions require students to provide responses ranging in length from three words to two paragraphs [7].

Although short answers are an effective evaluation method, teachers still struggle to use them, particularly in manual grading. Manual answer grading can be inconsistent since human graders must infer meaning from the student's answer [16]. Human graders may become fatigued after reviewing many responses, and the way they correct remaining responses may also vary [17]. This situation may be caused by fatigue, prejudice, or ordering effects [2], [8], [18]. Another reason for the discrepancy is that manual grading is subjective [19, 20] and highly dependent on the moods of the graders [21]. Moreover, the number of students [1], [5], [7] and the time-consuming [22]-[24] aspect of manually scoring short-answer questions pose difficulties. Approximately thirty percent of a teacher's time is spent evaluating students [25]. This problem is genuinely concerning since it means teachers cannot concentrate on their primary task, teaching. This condition will negatively affect teachers' and students' teaching and learning processes.

\* Corresponding author.

E-mail address: hcwang@mail.ncku.edu.tw

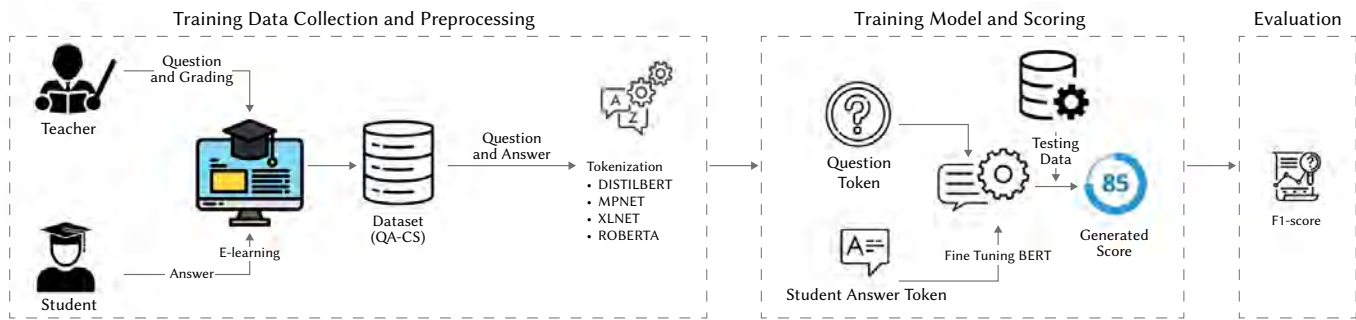


Fig. 1. Phases of the Proposed Research Framework.

Implementing artificial intelligence (AI) is the answer to this issue. AI's capability to produce innovative and real outcomes has elevated it to the forefront of attention in numerous industries, especially education [67]. Natural language processing (NLP) is an AI technology that has the potential to solve the issue of manual grading by enabling the development of a system that can grade responses to short-answer queries automatically; this is referred to as automated short-answer scoring system. The automatic scoring of short answers is one of the most important applications of NLP [26], [27]. In education, automated scoring of short answers has become increasingly popular, allowing for efficient and objective evaluation of student responses. Automatic short answer scoring (ASAS) assigns an output score to a given input answer [28]. The objective of ASAS is to develop a predictive model that takes as input a text response to a specific prompt (e.g., a question about a reading passage) and generates a score expressing the correctness of that response [10], [29], [30]. ASAS systems have garnered much interest because of their capacity to deliver fair and inexpensive grading of large-scale examinations and enhance learning in educational environments [31]. Many studies have focused on the creation of automatic short-answer grading systems, such as C-Rater [32], AutoSAS [25], and AutoMark [33]. However, the accuracy and reliability of these systems can be problematic, particularly as grading becomes more subjective and complex.

Many strategies are utilized to attain high accuracy in the automated scoring of short answer questions. Deep learning approaches have shown promise in enhancing the accuracy of automated scoring systems by enabling them to learn from large datasets and recognize patterns that conventional algorithms may overlook. This study explores constructing a reliable, efficient, and accurate ASAS system via deep learning. Our ultimate objective is to prove that deep learning can be utilized to improve automated scoring systems. While prior research has demonstrated that deep learning techniques can increase the accuracy of ASAS, our study uses a novel approach that focuses on constructing a reliable and more accurate system. Our model incorporates hybrid transfer learning for automated short answer scoring (HTL-ASAS). HTL-ASAS uses various pre-trained tokenizers in combination with the bidirectional encoder representations from transformers (BERT) to increase the accuracy of predictions. We also created a novel student-collected answer dataset for this study. This dataset was acquired without regard to gender or name to eliminate subjectivity and improve system reliability. By emphasizing accuracy and reliability, we seek to contribute to developing more dependable and trustworthy automated short-response scoring systems and enhance the educational experience for all students.

The following is the structure of this document: In section II, prior research concerning ASAS is discussed. The proposed development of the framework is illustrated in Section III. The findings and experimental context are detailed in Section IV. A discussion of the results of the findings is provided in Section V. In section VI, the concluding remarks are provided.

## II. LITERATURE REVIEW

ASAS is a challenging task that requires the capacity to evaluate the semantic content of a student's response accurately. In recent years, this topic has been the subject of numerous studies, and many techniques have emerged as potentially effective methods for enhancing the accuracy of ASAS systems. The fundamental concept of ASAS is to compare student responses to teacher responses, sometimes known as the "gold standard." Studies have utilized various approaches to calculate text similarity. One type of approach involves calculating text similarity based on semantic [34] or grammatical characteristics [13] or with the word overlapping approach [35]. Many advancements have been made to this fundamental idea, including using a semantic similarity measuring approach based on word embedding techniques and syntactic analysis to evaluate the learner's accuracy [5]. Combining semantic analysis with orthography and syntax analysis [36] or with graph-based lexico-semantic text matching is a further advancement that can be implemented [37].

Machine learning is another topic that can be applied to automatically scoring short answers. Term frequency inverse-document frequency (TF-IDF) [26], [38], long short-term memory (LSTM) [39, 40], support vector machines (SVMs) [7], [9], [41], latent semantic analysis [42], Gini [7], k-Nearest Neighbors (KNN) [7], finite state machine [18], and bagging and boosting [7] have all been employed. In addition to applying machine learning, various studies have employed deep learning to increase the accuracy of ASAS. Earlier research employed the concept of deep learning by utilizing transformers for data training. Transformers can be converted into graph transformers, which generate relation-specific token embeddings within each subgraph, which are subsequently aggregated to produce a subgraph representation [43]. Other studies have utilized pre-trained models such as BERT [22], [26], [31], [44]-[48], XLNET [49], [50], MPNET [51], [52], RoBERTa [50], [53], and Distil BERT [53].

In conclusion, many deep learning techniques can be applied to the ASAS system. These techniques have demonstrated potential for enhancing the accuracy of ASAS systems. Current research shows that deep learning models are excellent at enhancing the reliability of these systems.

## III. PROPOSED FRAMEWORK

In the proposed framework, there are three different procedures. The first step is data collection and preprocessing. The second procedure consists of training and testing the model that has been trained using the created dataset. Evaluation of the trained model is the final phase. The results of this evaluation are then compared to those of other studies to show the strengths of the framework proposed in this study. Fig. 1 displays the phases of the proposed research framework.



### A. Data Collection Module

First, we consider the data collection process. Teachers and students of an introductory information technology course participated in this phase. The teacher administered questions via an e-learning platform. The students then responded to the teacher’s questions. Students responded to ten questions related to the course. Within two weeks, the answers of 229 students who had responded to the ten questions posed by the teacher were gathered. After the collection of student responses was complete, the teacher manually evaluated the results of the students’ work. The teacher conducted the evaluation based on a previously prepared answer key. The teacher scored 50 for incorrect responses and 100 for correct responses. To recognize the students’ effort in responding to the question, teachers award 50 points to those who provide incorrect answers. For student responses like the teacher’s, values between 50 and 100 were awarded in multiples of 10, namely, 60, 70, 80, and 90. The teacher also assigned a score of zero to students who did not respond to the given questions.

### B. Data Preprocessing Module

After data collection, the next stage is data preprocessing. Tokenization is performed during this phase. In natural language processing, tokenization is often used to extract needed abstract information of paragraphs or sentences into smaller units that can be assigned meaning more readily by machine. Tokenizers are typically either carefully constructed systems of language-specific rules, which are expensive and require both manual feature engineering and linguistic expertise, or data-driven algorithms that split strings based on frequencies within a corpus, which are more flexible and easier to scale but are ultimately too simplistic to handle the wide range of linguistic phenomena that are not captured by their string-splitting [54]. In deep learning, the fundamental model for extracting contextualized word embeddings is called a Transformer [64]. The central concept of the Transformer architecture is to employ multi-head attention for concurrent data processing while preserving the temporal sequence characteristic of time-series data by the inclusion of positional embedding within the embedding layer [66].

Table I illustrates the tokenization of the phrase “An artificial intelligence robot.” This table displays the transformation of the sentence into word embedding and attention mask embedding, following the Transformer architecture. This tokenization process distinguishes our study from previous research. We employ four distinct tokenizers within the proposed hybrid transfer learning framework.

TABLE I. EXAMPLES OF TOKENIZATION

Before tokenization	An artificial intelligence robot					
<b>Token</b>	[cls]	‘An’	‘artificial’	‘intelligence’	‘robot’	[pad]
<b>Word Embedding</b>	$WE_1$	$WE_2$	$WE_3$	$WE_4$	$WE_5$	$WE_n$ 0
<b>Attention Mask Embedding</b>	$Att_1$	$Att_2$	$Att_3$	$Att_4$	$Att_5$	$Att_n$ 0

Several other types of tokenizers were subjected to experimentation before the identification of the four types that would be utilized in this study. The findings of this experiment indicate that the input format of the BERT model, which was utilized during the training phase, is compatible with the four tokenizers selected for this study: DistilBERT, MPNet, XLNet, and RoBERTa. In this study, various experiments were conducted in which the outcomes of the chosen tokenizer were combined with the BERT model’s training data. Hybrid MPNet refers to the output of the MPNet tokenizer when combined with the BERT

model. Hybrid DistilBERT is the name given to the combination of the DistilBERT tokenizer and the BERT model. The combination of the XLNet tokenizer and the BERT model is called Hybrid XLNet. The hybrid name for the RoBERTa tokenizer and the BERT model is Hybrid RoBERTa. A challenge appeared during the procedure of identifying the optimal combination: the lengthy duration of one experiment. To circumvent this, we conducted experiments on two separate servers. This is greatly beneficial in establishing correspondence between the tokenizer and the BERT model that was employed during the data training phase.

DistilBERT [55] is derived from BERT [56] by employing knowledge distillation. To create a more compact version of BERT, the architects of DistilBERT eliminated token-type embeddings and the pooler from the architecture and reduced the number of layers by a factor of 2. DistilBERT is a lightweight variant of BERT that is pre-trained using only the masked language model (MLM) task but with the same corpus: BookCorpus, which contains 800 million words; English Wikipedia, which contains 2,500 million words, a 30,000 token vocabulary, and WordPiece tokenization. Given an evolving word definition, the WordPiece model is combined with a data-driven approach to maximize the language-model likelihood of the training data. Given a training corpus and D desired tokens, the optimization problem is to select D word pieces such that when they are segmented according to the selected WordPiece model, the resulting corpus contains the fewest number of word pieces [57].

The masked and permuted pretraining model (MPNet) tokenizer was developed in collaboration with researchers from Microsoft and the Nanjing University of Science and Technology in 2020 [58]. MPNet incorporates the benefits of MLMs, such as BERT, and Pre-trained Language Models (PLMs), such as XLNet, by incorporating additional positional information into the permutation-based loss function. The MPNet tokenizer employs a byte-level byte pair encoding (BPE) algorithm to generate a vocabulary of subwords with a fixed size. The BPE algorithm iteratively replaces the most frequent pairs of consecutive bytes in the input text with a single new byte. This procedure is repeated until the desired vocabulary size has been attained. It can, therefore, comprehend a text based on its positional and nonpositional information. The tokenizer utilized by MPNet is inherited from BERT. MPNet was trained on many corpora of text totaling over 160 GB in size and optimized for multiple downstream NLP tasks [59].

The XLNet tokenizer is comparable to the tokenizers used in other transformer-based models but has some distinctive characteristics. Like other tokenizers, it transforms unprocessed text into a sequence of tokens the model can process. The tokenizer employs a subword-based approach, which divides words into smaller subwords and assigns a unique token to each subword. The total size of XLNet using subword fragments for Wikipedia, BooksCorpus, Giga5, ClueWeb, and Common Crawl is 32.89B [60]. XLNet uses the SentencePiece tokenization algorithm. SentencePiece consists of a natural language processing tokenizer and detokenizer. It performs subword segmentation, supports the BPE algorithm and unigram language model, and converts this text into an id sequence while ensuring perfect reproducibility of the normalization and subword segmentation [61]. BPE is an algorithm for subword segmentation that encodes uncommon and unknown terms as sequences of subword units. The assumption is that various word classes can be translated using units smaller than words, such as names (via character reproduction or transliteration), compounds (via compositional translation), and cognates and loanwords (via phonological and morphological transformations) [62].

The RoBERTa tokenizer generates subword tokens using a variant of the BPE algorithm. BPE functions by iteratively merging the most frequently occurring character or character sequence pairs in the

training corpus until the maximum vocabulary size is attained. This method can produce a vocabulary of variable-length subword units that more accurately represent the morphology and syntax of the language than traditional word-based tokenization. Additionally, the RoBERTa tokenizer employs a variety of optimizations to enhance the quality of the tokenization procedure. It employs, for instance, dynamic masking to prevent overfitting during pretraining and removes whitespace from the input text to increase efficiency. RoBERTa was trained on a combined dataset for the same number of steps as before (100K). RoBERTa preprocessed over 160 GB of text in total [63].

Table II displays the tokenizers utilized in this study. Each tokenizer uses a distinct corpus for recognizing terms. DistilBERT and MPNet utilize scholarly sources such as BookCorpus and Wikipedia, whereas XLNet adds Giga5 and ClueWeb. RoBERTa expands its corpus to include CC-News, OpenWebText, and Stories, among others. This distinction results in distinct text representations. This study investigated the appropriate tokenizer for short-answer question tasks.

TABLE II. SUMMARY OF EACH TOKENIZER

Tokenizer	Corpus	Embedding Technique	#Tokens	#Positions
DistilBERT	BookCorpus, English Wikipedia	WordPiece Embedding	85%	100%
MPNet	BookCorpus, English Wikipedia	WordPiece Embedding	95%	100%
XLNet	BookCorpus, Wikipedia, Giga5, ClueWeb,	WordPiece Embedding	92.5%	92.5%
RoBERTa	BookCorpus, English Wikipedia, CC-News, OpenWebText, Stories + pretrain for even longer	Byte Pair Encoding	-	-

### C. Question and Answer Embedding Modules

The preprocessing dataset is trained after the tokenization of teacher questions and student responses is performed. BERT accomplished natural language understanding by considering input consistency. BERT extracts a single token sequence from a single text sentence, and for the NSP objective, it extracts a single token sequence from two text sentences (adding a [SEP] token as a separator) [56]. Each sequence has the specific classification embedding [CLS] added before it, and it serves as the input of the classification-task layer. Combining the corresponding token, segment, and position embeddings puts the corresponding representation of the input together. Each provided input token receives this kind of approach [56]. Fig. 2 illustrates the architecture of BERT fine-tuning for this study. The tokenization

process is initiated once the learners have provided answers to the teacher's questions and the encoder layer has become working, as represented in Fig. 2. The BERT encoder has two primary sublayers: the multihead self-attention layer and the positionwise fully connected feedforward network layer [56]. The output of the Question and Answer embedding of Hybrid MPNet is the latent embedding of the answer and question. The question-and-answer embedding sizes are  $1 \times 768$ . These two embeddings are concatenated to predict the potential score for students and produce an object of size  $2 \times 768$ . Finally, we connect the regression layer to predict the probability of the score and use the highest probability as the predicted score.

### D. Evaluation Technique

In this research, k-fold cross-validation is used as the evaluation methodology. The dataset is first divided into k folds, with k-1 folds used for training and the remaining fold used for evaluation. The folds are then switched until all folds have been trained and evaluated against the remaining k-1 folds, and an average is then calculated. This study utilizes ten-fold cross-validation. The F1-score is utilized for the evaluation matrices in the study. Formula 1 defines the F1-score as the weighted average of precision and recall based on the weight function  $\beta$ . Formula 2 defines the F1-score as the harmonic mean of precision and recall. The F1 score is also referred to as the F-measure. Different F1-score indices can assign distinct weights to precision and recall. Precision is calculated by dividing the number of correct instances retrieved by the total number of instances retrieved, as in Formula 3. Recall is calculated by dividing the number of correct instances retrieved by the total number of correct instances, as in Formula 4.

$$F\text{-score} : F_{\beta} = (1 + \beta^2) * \frac{P * R}{\beta^2 * P + R} \quad (1)$$

When  $\beta = 1$ , the standard F1-score is obtained

$$F\text{-score} : F_1 = F = 2 * \frac{P * R}{P + R} \quad (2)$$

$$\text{Precision} : P = \frac{tp}{tp + fp} \quad (3)$$

$$\text{Recall} : R = \frac{tp}{tp + fn} \quad (4)$$

## IV. RESULTS

In this section, the experimental results of the proposed Hybrid MPNet are presented. We collect and sign the score of each answer. Then, we propose a new deep-learning technique to predict the score. Finally, we evaluate the accuracy of our proposed method using the F1-score, precision, and recall.

### A. Dataset

This study employs a dataset compiled by the authors. The collected dataset comprises four columns: teacher-initiated questions, teacher-prepared responses, student responses, and students' grading in numerical form. The example of the collected dataset can be seen in Table III. The questions given to students were related to an introductory course in information technology. There were five categories and two questions per category for ten questions. The five categories were: 1) data and information, 2) the most recent technology, 3) software, 4) hardware, and 5) the development of computer networks. The scores assigned by the teacher have a value of 0, 50, 60, 70, 80, 90, or 100. 229 students responded to the questions, so the total data collected contained 2290 data. After the data were collected, they were cleaned. First, responses with a zero value were removed, indicating that the student did not answer the question. This

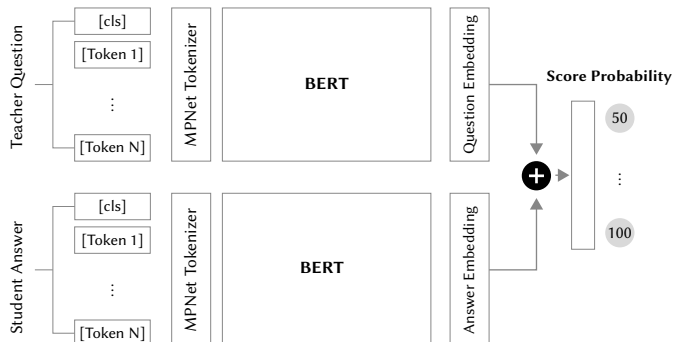


Fig. 2. BERT-Based Question-Answer Embedding Architecture.

was done to reduce the homogeneity of the training data and ensure that the resulting model is highly accurate and creates a trustworthy system. Following the data cleaning procedure, 2023 data points remained. The distribution of word length in student responses is illustrated in Fig. 3. The majority of responses are typically between zero and two hundred words in length. The longest response exceeds 400 words in length.

TABLE III. EXAMPLES OF COLLECTED DATASET

Question	Teacher Answer	Student Answer	Grade
Please define what a “computer network” is.	A computer network can be defined as a communication system that connects two or more computers and peripheral devices and allows data transfer between components.	A computer network is a link between one computer/device and another computer/device that uses network media as an intermediary.	70
		A unit that causes a computer to run.	50
		A communication network that allows computers to communicate with each other by exchanging data.	100
Please explain the definition of Artificial Intelligence (AI).	A field of computer science devoted to solving cognitive problems commonly associated with human intelligence, such as learning, problem solving and pattern recognition.	-	0
		A program with a neural network that can think like humans in carrying out tasks.	70
		A field of computer science that tries to solve cognitive problems like learning, solving problems, and recognizing patterns, that are often associated with human intelligence.	100
		A smart technology embedded in a device.	60
		An artificial intelligence robot.	50

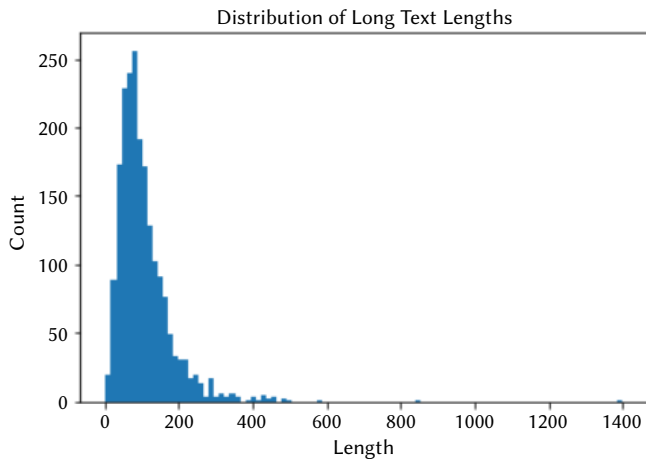


Fig. 3. Distribution of Word Length in Student Responses.

As a consequence of the data cleansing process, the number of responses varied for each question. After the data cleansing procedure, the number of student responses categorized by question type is presented in Table IV. There are three questions to which fewer than 200 students respond. One question pertains to the data

and information category, one concerns software, and one investigates the development of computer networks. Because numerous students failed to respond to that question, the instructor therefore assigns a zero grade.

Table V displays the quantity of student responses used in this study based on assigned scores after the data cleaning process. As shown in Table V, the quantity of responses for each value is not distributed exactly equally. The answers provided by the majority of students yield scores ranging from 60 to 80. From the complete data for 2023, this is evident from the 1,419 answers that obtained a score within that range. At 50, 90, and 100, the remainder was balanced. It is possible to conclude from this distribution that a small number of students submitted responses attaining a perfect score of 100. Also, a small number of students submitted responses that received a minimum score of 50.

TABLE IV. NUMBER OF STUDENT ANSWERS BASED ON QUESTION TYPE

Category	Question	Number of Answers
Data and information	What are data and information? Please compare the differences.	215
	What is the information processing cycle?	198
Recent technology	What is Augmented Reality (AR)?	203
	Please explain the definition of Artificial Intelligence (AI).	203
Software	Please define what “freeware” is.	164
	Please define what an “operating system” is and explain its function.	219
Hardware	Explain the function of a router.	213
	What is a Central Processing Unit (CPU)?	213
Development of computer networks	Please define what a “computer network” is.	183
	Please give the definition of the Internet of Things (IoT).	212

TABLE V. NUMBER OF STUDENT ANSWERS BASED ON TEACHER GRADING

Grading	Number of Answers
50	247
60	403
70	586
80	430
90	174
100	183

### B. Parameter Setting

We propose hybrid transfer learning as a model for ASAS. Before conducting model training experiment, we set our parameters. Table VI summarizes some of the study’s parameters.

TABLE VI. PARAMETER SETTINGS OF THE AUTOMATED SHORT ANSWER GRADING MODEL

Parameter	Value
Batch size	10
Optimizer	Adam
Learning rate	0.00001
Embedding size	300
Activation function	ReLU
The final layer activation function	Sigmoid



### C. Result

**Experiment 1:** In the first experiment, the authors trained the proposed model (HTL-ASAS) for various epochs to determine which provided the most accurate model. The epochs tested were 10, 20, 30, and 40. Training used ten-fold cross-validation to validate the model. Comparison between the evaluation generated by the computer and the evaluation conducted by the teacher yields the F1 score accuracy. When both the machine and the teacher arrive at the same evaluation, this represents a true positive. False positives happen when the assessments of the machine and the instructor differ. Fig. 4 displays each model's F1 score after 10, 20, 30, and 40 epochs for each tokenizer. The average F1 score, as shown in Fig. 4, is the result of the evaluation that was performed.

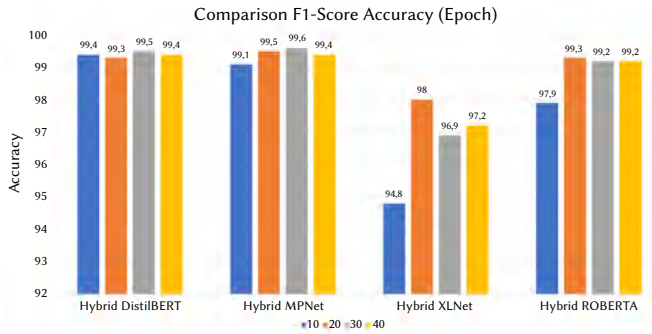


Fig. 4. Results of F1-Score by Epoch.

**Experiment 2:** In the second experiment, the highest F1-score obtained by the various hybrid transfer learning algorithms in the proposed framework, namely, Hybrid DistilBERT, Hybrid MPNet, Hybrid XLNet, and Hybrid RoBERTa, were compared. Fig. 5 displays the comparison of F1 scores.

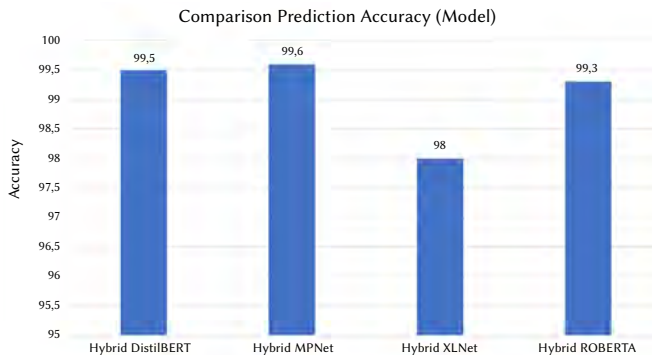


Fig. 5. Results of F1-Score by Proposed Framework Model.

**Experiment 3:** The objective of the third experiment was to compare the F1-scores of the proposed framework with the F1-scores models based on prior research. The prior research models were trained using the dataset from this study. The F1 scores were then obtained from the results of the training model. The following models were used for comparison:

1. BERT architecture to grade short answers [47], [53], [65]. A pre-trained version of the BERT base model was utilized in these experiments.
2. MPNet, specifically mpnet-base-v2 model, which has been used to determine the similarity of short texts [52].
3. DistilBERT, which has been used to grade short answer responses [53]. This study utilized a pre-trained DistilBERT model, namely, the DistilBERT base model.

4. XLNet, which is a pre-trained model used to assess short-answer responses [50].
5. Pre-trained RoBERTa and RoBERTa base architectures used in previous studies [50], [53], [65].

Each model from previous research was trained on this study's dataset and then compared to the framework proposed in this study. For previous BERT research, ten epochs were used to train the model. For MPNet, 30 epochs were used, the same number used for the MPNet hybrid proposed in this study. For DistilBERT, the number of epochs was set to 30, the optimal number for the DistilBERT hybrid. The final two models, XLNet and RoBERTa, were trained in 20 epochs, the optimal number of epochs for hybrid XLNet and RoBERTa. Table VII compares the F1 scores of the proposed framework and models from previous studies.

TABLE VII. COMPARISON OF F1-SCORE ACCURACY

Research	Model	F1-Score Accuracy
[47], [53], and [65]	BERT	0.992
[52]	MPNet	0.952
[53]	DistilBERT	0.961
[50]	XLNet	0.899
[50], [53], and [65]	RoBERTa	0.963
	Hybrid DistilBERT	0.995
Our Proposed Framework (*)	Hybrid MPNet	<b>0.996</b>
	Hybrid XLNet	0.98
	Hybrid RoBERTa	0.993

### V. DISCUSSION

The first experiment's results, depicted in Fig. 4, indicate that increasing the number of epochs has no effect on the accuracy of predictions made for the framework proposed in this study. Each tokenizer employed by the proposed framework requires a distinct number of epochs to attain the highest level of accuracy. Except for the DistilBERT tokenizer, the accuracy of each tokenizer is the lowest for epoch 10. In this investigation, the sample size at epoch 10 was insufficient for each model to achieve maximum accuracy. Upon entering epoch 20, the F1-score of several models increased. Only the DistilBERT tokenizer experienced a reduction in score. The XLNet and RoBERTa tokenizers reached their maximal F1-scores of 98% and 99.3%, respectively, in the 20th epoch; thus, the F1-scores of the corresponding hybrid models at epochs 30 and 40 were lower than at epoch 20. At epoch 30, the MPNet tokenizer attained its highest F1-score of 99.6%, and the DistilBERT tokenizer reached its maximum accuracy rate of 99.5%. The first experiment's results were used for comparison in the second experiment. The MPNet tokenizer paired with the BERT layer (Hybrid MPNet) achieved the greatest accuracy of 99.6%, as shown in Fig. 5. The DistilBERT tokenizer paired with the BERT layer (Hybrid DistilBERT) achieved the next-best accuracy of 99.5%. The accuracy of Hybrid RoBERTa was 99.3%, while the accuracy of Hybrid XLNet was only 98%. This comparison demonstrates that by employing hybrid transfer learning, accuracy increases, and the resulting data can enable the development of a more reliable ASAS system.

The results of Experiment 3, presented in Table VII, indicate that our proposed framework that combines the MPNet tokenizer with the BERT layer, also known as Hybrid MPNet, has the highest F1 score among the other models. Hybrid MPNet achieves an F1-score of 99.6%. In addition, Hybrid DistilBERT and Hybrid RoBERTa are among the models proposed in this study that have the highest value relative to models used in previous research. The F1 scores produced by Hybrid DistilBERT and Hybrid RoBERTa were 99.5% and 99.3%, respectively.

The BERT model [47], [53], [65] produced the highest values among previous models. The F1-score for this BERT model was 99.2%. This F1-score is greater than that of one of the proposed models in this investigation, namely, Hybrid XLNet (98%), and is also greater than the F1-scores obtained by several models used in previous studies, including MPNet (95.2%) [52], DistilBERT (96.1%) [53], XLNet [50], and RoBERTa (96.3%) [50], [53], [65].

The results of this study indicate that Hybrid MPNet is a more accurate method than those used in previous research. This is because MPNet utilizes the dependencies between predicted tokens through permuted language modeling and enables the model to see supplementary position information to overcome the difference between pretraining and fine-tuning. In addition, Hybrid MPNet's better results compared to alternative methods can be attributed to the specific correspondence between the corpus trained in the MPNet tokenizer and the collected dataset. The corpus utilized by the MPNet tokenizer is comprised of words extracted from English Wikipedia and BookCorpus, as shown in Table II. DistilBERT tokenizer operations utilize the identical corpus. An important distinction is found in the fact that DistilBERT trains a lower percentage of tokens (85%) than MPNet Tokenizer. In comparison to alternative tokenizers, the MPNet tokenizer utilizes a greater quantity of training tokens. Experiments on various tasks demonstrate that MPNet outperforms MLM and PLM, as well as previously robust pre-trained models, including BERT, XLNet, and RoBERTa, by a substantial margin [59].

The findings derived from this study will influence the area of education. This system will improve the performance of teachers when evaluating student work. It will not be long before the students are informed of the assessment results. As a result, teachers can dedicate more time to planning and refining the learning process within the classroom. Instead of having to wait for the instructor to evaluate their work manually, this method enables students to obtain immediate feedback. Students will receive more objective grades as a consequence of the reduced subjectivity of the teacher caused by the implementation of this system. By establishing confidence among teachers and students, the experimental results indicate that utilizing AI to assess short-answer assessments produces reliable and objective outcomes. Aside from that, the implementation of this system's results demonstrates that artificial intelligence can be applied to the field of education. The opportunities for both educators and students to utilize AI are described by the Sustainable Development Goals (SDG4) of the UNESCO 2030 Agenda as they pertain to the impact of AI in education [68].

## VI. CONCLUSION

Teachers can select from various effective assessment methods for students, one of which is short answer questions. However, one of the most challenging aspects of teaching is evaluating student work in a limited amount of time. Consequently, the results of an assessment can be inconsistent if the teacher is pressured. Our study assists teachers in overcoming these inconsistencies by developing a system that automatically assigns grades to students' short answers. The goal is to construct a trustworthy system, so students believe the assessments are accurate. A method that can generate near-perfect system accuracy is required to achieve this objective. In addition, the system must be objective about student work. For the method proposed in this study, both objectives are met. We implement hybrid transfer learning as a novel technique for achieving high accuracy and generate a new training dataset containing students' short responses and feedback. We anticipate that the constructed system will be capable of objective evaluation with this dataset. Based on the results of the conducted experiments, the hybrid transfer learning method proposed in this study has the highest accuracy of 99.6%. Despite focusing solely on the

F1-score to assess accuracy, the test results for this system indicate a 99.6% accuracy rate, which signifies a highly optimistic implementation potential. Nevertheless, additional assessment utilizing additional matrices is necessary. There is no doubt that a more comprehensive assessment of the system's capability to evaluate student exams can be obtained by administering tests utilizing a broader variety of comprehensive matrices. The F1-score matrix, nevertheless, is considered satisfactory from the perspective of this study.

Future research may concentrate on implementing the proposed framework in disciplines other than information technology. In addition, other evaluation matrices can be applied to evaluate this mode. In the future, automated scoring will hopefully make administering assessments easier for teachers to concentrate on enhancing the quality of learning.

## REFERENCES

- [1] Q. Aini, A. E. Julianto, and D. Purbohadi, "Development of a Scoring Application for Indonesian Language Essay Questions," in Proceedings of the 2018 2nd International Conference on Education and E-Learning, 2018, pp. 6-10.
- [2] S. Burrows, I. Gurevych, and B. Stein, "The Eras and Trends of Automatic Short Answer Grading," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 60-117, Oct. 2014, doi: <https://doi.org/10.1007/s40593-014-0026-8>.
- [3] B. S. J. Kapoor et al., "An analysis of automated answer evaluation systems based on machine learning," in 2020 International Conference on Inventive Computation Technologies (ICICT), IEEE, 2020, pp. 439-443.
- [4] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artificial Intelligence Review*, Sep. 2021, doi: <https://doi.org/10.1007/s10462-021-10068-2>.
- [5] F. F. Lubis et al., "Automated Short-Answer Grading using Semantic Similarity based on Word Embedding," *International Journal of Technology*, vol. 12, no. 3, p. 571, Jul. 2021, doi: <https://doi.org/10.14716/ijtech.v12i3.4651>.
- [6] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 752-762.
- [7] A. Çınar, E. Ince, M. Gezer, and Ö. Yılmaz, "Machine learning algorithm for grading open-ended physics questions in Turkish," *Education and Information Technologies*, Mar. 2020, doi: <https://doi.org/10.1007/s10639-020-10128-0>.
- [8] A. Olowolayemo, S. D. Nawi, and T. Mantoro, "Short answer scoring in English grammar using text similarity measurement," in 2018 International Conference on Computing, Engineering, and Design (ICCED), IEEE, 2018, pp. 131-136.
- [9] G. De Gasperis et al., "Automated grading of short text answers: preliminary results in a course of health informatics," in Advances in Web-Based Learning-ICWL 2019: 18th International Conference, Magdeburg, Germany, September 23-25, 2019, Proceedings, Springer International Publishing, 2019, pp. 190-200.
- [10] S. Patil and K. P. Adhiya, "Automated Evaluation of Short Answers: a Systematic Review," in *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2021*, 2022, pp. 953-963.
- [11] Y.-H. Park, Y.-S. Choi, C.-Y. Park, and K.-J. Lee, "EssayGAN: Essay Data Augmentation Based on Generative Adversarial Networks for Automated Essay Scoring," *Applied Sciences*, vol. 12, no. 12, p. 5803, Jun. 2022, doi: <https://doi.org/10.3390/app12125803>.
- [12] M. J. Gierl, S. Latifi, H. Lai, A.-P. Boulais, and A. De Champlain, "Automated essay scoring and the future of educational assessment in medical education," *Medical Education*, vol. 48, no. 10, pp. 950-962, Sep. 2014, doi: <https://doi.org/10.1111/medu.12517>.
- [13] S. H. Mijbel and A. T. Sadiq, "Short Answers Assessment Approach based on Semantic Network," *Iraqi Journal of Science*, pp. 2702-2711, Jun. 2022, doi: <https://doi.org/10.24996/ij.s.2022.63.6.35>.
- [14] C. E. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer, "Scaling

- short-answer grading by combining peer assessment with algorithmic scoring,” in Proceedings of the first ACM conference on Learning@Scale Conference, 2014, pp. 99-108.
- [15] A. K. F. Lui, S. C. Ng, and S. W. N. Cheung, “A framework for effectively utilising human grading input in automated short answer grading,” *International Journal of Mobile Learning and Organisation*, vol. 16, no. 3, p. 266, 2022, doi: <https://doi.org/10.1504/ijmlo.2022.124160>.
- [16] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, “Automatic short answer grading and feedback using text mining methods,” *Procedia Computer Science*, vol. 169, pp. 726-743, 2020.
- [17] S. Bonthu, S. R. Sree, and M. H. M. K. Prasad, “Automated short answer grading using deep learning: A survey,” in *Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings*, vol. 5, 2021, pp. 61-78.
- [18] K. Anekboon, “Automated scoring for short answering subjective test in Thai’s language,” in *2018 International Conference on Image and Video Processing, and Artificial Intelligence*, vol. 10836, SPIE, 2018, pp. 324-329.
- [19] M. Beseiso, O. A. Alzubi, and H. Rashaideh, “A novel automated essay scoring approach for reliable higher educational assessments,” *Journal of Computing in Higher Education*, Jun. 2021, doi: <https://doi.org/10.1007/s12528-021-09283-1>.
- [20] J. Xiong, J. M. Wheeler, H. Choi, J. Lee, and A. S. Cohen, “An empirical study of developing automated scoring engine using supervised latent dirichlet allocation,” in *Quantitative Psychology: The 85th Annual Meeting of the Psychometric Society, Virtual, Springer International Publishing*, 2021, pp. 429-438.
- [21] P. Kudi, A. Manekar, K. Daware and T. Dhattrak, “Online Examination with short text matching,” *2014 IEEE Global Conference on Wireless Computing & Networking (GCWCN)*, 2014, pp. 56-60, doi: 10.1109/GCWCN.2014.6998787.
- [22] A. Conдор, “Exploring automatic short answer grading as a tool to assist in human rating,” in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II*, vol. 12164, Springer International Publishing, 2020, pp. 74-79.
- [23] S. Roy, Y. Narahari, and O. D. Deshmukh, “A perspective on computer assisted assessment techniques for short free-text answers,” in *Computer Assisted Assessment. Research into E-Assessment: 18th International Conference, CAA 2015, Zeist, The Netherlands, June 22–23, 2015, Proceedings*, vol. 18, Springer International Publishing, 2015, pp. 96-109.
- [24] X. Ye and S. Manoharan, “Machine Learning Techniques to Automate Scoring of Constructed-Response Type Assessments,” in *2018 28th EAEEIE Annual Conference (EAEEIE)*, IEEE, 2018, pp. 1-6.
- [25] Y. Kumar, S. Aggarwal, D. Mahata, R. R. Shah, P. Kumaraguru, and R. Zimmermann, “Get it scored using autosas—an automated system for scoring short answers,” in *Proceedings of the AAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9662-9669.
- [26] P. Shweta and K. Adhiya, “Comparative Study of Feature Engineering for Automated Short Answer Grading,” in *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*, IEEE, 2022, pp. 594-597.
- [27] C. N. Tulu, O. Ozkaya, and U. Orhan, “Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM,” *IEEE Access*, vol. 9, pp. 19270–19280, 2021, doi: <https://doi.org/10.1109/access.2021.3054346>.
- [28] T. Sato, H. Funayama, K. Hanawa, and K. Inui, “Plausibility and Faithfulness of Feature Attribution-Based Explanations in Automated Short Answer Scoring,” presented at the *International Conference on Artificial Intelligence in Education, 2022, Lecture Notes in Computer Science*, vol 13355. Springer, Cham.
- [29] M. Heilman and N. Madnani, “The impact of training data on automated short answer scoring performance,” presented at the *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, Denver, Colorado, 2015*, pp. 81-85.
- [30] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. Lee, “Investigating neural architectures for short answer scoring,” presented at the *Proceedings of the 12th workshop on innovative use of NLP for building educational applications, Copenhagen, Denmark, 2017*, pp. 159-168.
- [31] H. Funayama, T. Sato, Y. Matsubayashi, T. Mizumoto, J. Suzuki, and K. Inui, “Balancing Cost and Quality: An Exploration of Human-in-the-Loop Frameworks for Automated Short Answer Scoring,” presented at the *International Conference on Artificial Intelligence in Education, 2022, Lecture Notes in Computer Science*, vol 13355. Springer, Cham.
- [32] C. Leacock and M. Chodorow, “C-rater: Automated scoring of short-answer questions,” *Computers and the Humanities*, vol. 37, no. 4, pp. 389-405, 2003.
- [33] R. Siddiqi, C. J. Harrison, and R. Siddiqi, “Improving Teaching and Learning through Automated Short-Answer Marking,” *IEEE Transactions on Learning Technologies*, vol. 3, no. 3, pp. 237–249, Jul. 2010, doi: <https://doi.org/10.1109/tlt.2010.4>.
- [34] M. Mohler and R. Mihalcea, “Text-to-text semantic similarity for automatic short answer grading,” presented at the *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Greece, 2009, pp. 567-575.
- [35] F. S. Pribadi, A. B. Utomo, and A. Mulwinda, “Automated short essay scoring system using normalized Simpson methods,” in *Proceedings of the 6th International Conference on Education, Concept, and Application of Green Technology, Semarang, Indonesia, 2018*.
- [36] L. dela-Fuente-Valentín, E. Verdú, N. Padilla-Zea, C. Villalonga, X. P. Blanco Valencia, and S. M. Baldiris Navarro, “Semiautomatic Grading of Short Texts for Open Answers in Higher Education,” in *Higher Education Learning Methodologies and Technologies Online*, 2022, pp. 49-62.
- [37] L. Ramachandran, J. Cheng, and P. Foltz, “Identifying patterns for short answer scoring using graph-based lexico-semantic text matching,” in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, Denver, Colorado, 2015*, pp. 97-106.
- [38] I. G. Ndukwe, B. K. Daniel, and C. E. Amadi, “A Machine Learning Grading System Using Chatbots,” in *Artificial Intelligence in Education, 2019*, pp. 365-368.
- [39] H. Qi, Y. Wang, J. Dai, J. Li, and X. Di, “Attention-based hybrid model for automatic short answer scoring,” in *Simulation Tools and Techniques: 11th International Conference, SIMUtools 2019, Chengdu, China, July 8–10, 2019, Proceedings 11, 2019*.
- [40] M. Uto and Y. Uchida, “Automated Short-Answer Grading Using Deep Neural Networks and Item Response Theory,” in *Artificial Intelligence in Education, 2020*, pp. 334-339.
- [41] K. Sakaguchi, M. Heilman, and N. Madnani, “Effective feature integration for automated short answer scoring,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, 2015*, pp. 1049-1054.
- [42] N. LaVoie, J. Parker, P. J. Legree, S. Ardison, and R. N. Kilcullen, “Using Latent Semantic Analysis to Score Short Answer Constructed Responses: Automated Scoring of the Consequences Test,” *Educational and Psychological Measurement*, vol. 80, no. 2, pp. 399–414, Jul. 2019, doi: <https://doi.org/10.1177/0013164419860575>.
- [43] R. Agarwal, V. Khurana, K. Grover, M. Mohania and V. Goyal, “Multi-Relational Graph Transformer for Automatic Short Answer Grading,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, 2022*, pp. 2001-2012.
- [44] H. Oka, H. T. Nguyen, C. T. Nguyen, M. Nakagawa and T. Ishioka, “Fully Automated Short Answer Scoring of the Trial Tests for Common Entrance Examinations for Japanese University,” in *International Conference on Artificial Intelligence in Education, 2022, Lecture Notes in Computer Science*, vol 13355. Springer, Cham.
- [45] J. Sawatzki, T. Schlippe and M. Benner-Wickner, “Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers,” in *Artificial Intelligence in Education: Emerging Technologies, Models and Applications: Proceedings of 2021 2nd International Conference on Artificial Intelligence in Education Technology, 2022, Lecture Notes on Data Engineering and Communications Technologies*, vol 104. Springer, Singapore.
- [46] K. Steimel and B. Riordan, “Toward instance-based content scoring with pretrained transformer models,” in *Proceedings of the Thirty-Fourth AAI Conference on Artificial Intelligence*, 2020, vol. 34.
- [47] C. Sung, T. I. Dhamecha and N. Mukhi, “Improving short answer grading using transformer-based pretraining,” in *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I, 2019*.
- [48] S. Takano and O. Ichikawa, “Automatic scoring of short answers using justification cues estimated by BERT,” in *Proceedings of the*



- 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), Seattle, Washington, 2022, pp. 8-13.
- [49] H. A. Ghavidel, A. Zouaq and M. C. Desmarais, "Using BERT and XLNET for the Automatic Short Answer Grading Task," in Proceedings of the 12th International Conference on Computer Supported Education - Volume 1: CSEDU, 2020, pp 58-67.
- [50] R. Somers, S. Cunningham-Nelson, and W. Boles, "Applying natural language processing to automatically assess student conceptual understanding from textual responses," *Australasian Journal of Educational Technology*, vol. 37, no. 5, pp. 98-115, Dec. 2021, doi: <https://doi.org/10.14742/ajet.7121>.
- [51] J. Garg, J. Papreja, K. Apurva, and G. Jain, "Domain-Specific Hybrid BERT based System for Automatic Short Answer Grading," presented at the 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-6.
- [52] V. Ramnarain-Seetohul, V. Bassoo, and Y. Rosunally, "Work-in-Progress: Computing Sentence Similarity for Short Texts using Transformer models," presented at the 2022 IEEE Global Engineering Education Conference (EDUCON), Tunis, Tunisia, 2022, pp. 1765-1768.
- [53] M. H. Haidir and A. Purwarianti, "Short answer grading using contextual word embedding and linear regression," *Jurnal Linguistik Komputasional*, vol. 3, no. 2, pp. 54-61, 2020.
- [54] J. H. Clark, D. Garrette, I. Turc, and J. Wieting, "Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 73-91, 2022, doi: [https://doi.org/10.1162/tacl\\_a\\_00448](https://doi.org/10.1162/tacl_a_00448).
- [55] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
- [56] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers), 2019, pp. 4171-4186.
- [57] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016.
- [58] B. Rodrawangpai and W. Daungjaiboon, "Improving text classification with transformers and layer normalization," *Machine Learning with Applications*, vol. 10, p. 100403, Dec. 2022, doi: <https://doi.org/10.1016/j.mlwa.2022.100403>.
- [59] K. Song, X. Tan, T. Qin, J. Lu and T. Y. Liu, "Mpnnet: Masked and permuted pretraining for language understanding," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857-16867, 2020.
- [60] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [61] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66-71, 2018.
- [62] R. Sennrich, B. Haddow and A. Birch, "Neural Machine Translation of Rare Words with Subword Units," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1715-1725, 2016.
- [63] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
- [64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez et al., "Attention is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [65] L. Camus and A. Filighera, "Investigating transformers for automatic short answer grading," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6-10, 2020, Proceedings, Part II*, vol. 21, pp. 43-48, Springer International Publishing, 2020.
- [66] J. Seo, S. Lee, and L. Liu, "TA-SBERT: Token Attention Sentence-BERT for Improving Sentence Representation," *IEEE Access*, vol. 10, pp. 39119-

39128, Apr. 2022, doi: <https://doi.org/10.1109/ACCESS.2022.3164769>.

- [67] F. García-Peñalvo, & A. Vázquez-Ingelmo, "What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 7-16, 2023, <https://doi.org/10.9781/ijimai.2023.07.006>
- [68] J. M. Flores-Vivar, & F. J. García-Peñalvo, "Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4)." *Comunicar*, vol. 31, no. 74, pp. 37-47, 2023, <https://doi.org/10.3916/C74-2023-03>



Martinus Maslim

Martinus Maslim earned a master's in informatics engineering in 2012 from Universitas Atma Jaya Yogyakarta, Indonesia. His areas of interest are artificial intelligence, information systems, and data science. Since 2013, he has been a lecturer in the Department of Informatics at Universitas Atma Jaya Yogyakarta, and he is now studying for his PhD at National Cheng Kung University, Taiwan. He is part of the Web Knowledge Discovery Lab.



Hei-Chia Wang

Hei-Chia Wang is a Professor from the Institute of Information Management at National Cheng Kung University, Taiwan. He got a doctoral degree from The University of Manchester in 1999. His research interests in natural language processing, e-learning, bioinformatics, information retrieval, and knowledge discovery. Currently, he is the leader of the Knowledge Discovery Lab.



Cendra Devayana Putra

Cendra Devayana Putra earned a master's degree in the Institute of Information Management at National Cheng Kung University, Taiwan. His research areas include natural language processing, deep learning, and management science. He is currently a doctoral student at National Cheng Kung University, Taiwan. He is also part of the Web Knowledge Discovery Lab.



Yulius Denny Prabowo

Yulius Denny Prabowo is a lecturer at Computer Science Department, Bina Nusantara University Jakarta, Indonesia. He finishes his doctoral degree from Bina Nusantara University in 2022. His research interests are natural language processing, machine learning, and deep learning.

# Virtual Reality and Language Models, a New Frontier in Learning

Juan Izquierdo-Domenech, Jordi Linares-Pellicer, Isabel Ferri-Molla\*

Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV), València (Spain)

Received 14 September 2023 | Accepted 15 February 2024 | Published 21 February 2024



## ABSTRACT

The proposed research introduces an innovative Virtual Reality (VR) and Large Language Model (LLM) architecture to enhance the learning process across diverse educational contexts, ranging from school to industrial settings. Leveraging the capabilities of LLMs and Retrieval-Augmented Generation (RAG), the architecture centers around an immersive VR application. This application empowers students of all backgrounds to interactively engage with their environment by posing questions and receiving informative responses in text format and with visual hints in VR, thereby fostering a dynamic learning experience. LLMs with RAG act as the backbones of this architecture, facilitating the integration of private or domain-specific data into the learning process. By seamlessly connecting various data sources through data connectors, RAG overcomes the challenge of disparate and siloed information repositories, including APIs, PDFs, SQL databases, and more. The data indexes provided by RAG solutions further streamline this process by structuring the ingested data into formats optimized for consumption by LLMs. An empirical study was conducted to evaluate the effectiveness of this VR and LLM architecture. Twenty participants, divided into Experimental and Control groups, were selected to assess the impact on their learning process. The Experimental group utilized the immersive VR application, which allowed interactive engagement with the educational environment, while the Control group followed traditional learning methods. The study revealed significant improvements in learning outcomes for the Experimental group, demonstrating the potential of integrating VR and LLMs in enhancing comprehension and engagement in learning contexts. This study presents an innovative approach that capitalizes on the synergy between LLMs and immersive VR technology, opening avenues for a transformative learning experience that transcends traditional boundaries and empowers learners across a spectrum of educational landscapes.

## KEYWORDS

Large Language Models, RetrievalAugmented Generation, Virtual Reality.

DOI: 10.9781/ijimai.2024.02.007

## I. INTRODUCTION

TECHNOLOGY'S rapid expansion, especially in internet-related fields, has revolutionized learning. For today's students, born in this tech-savvy era, accessing information is effortless, but it has raised concerns about their attention and problem-solving abilities. In response, educators are adapting teaching methods, such as Massive Open Online Courses (MOOCs) and the flipped classroom [1]. Virtual reality (VR) has been a research focus within computer science and information technology, especially for its educational applications. Recent advancements have made VR more accessible and immersive, enhancing its potential as a learning tool. VR has been utilized in various educational contexts, from primary and secondary classrooms to professional training programs. It caters to diverse student profiles, including different age groups, learning abilities, and backgrounds. Studies have demonstrated VR's effectiveness across various disciplines like science, history, and medicine, emphasizing its role

in providing interactive and realistic learning experiences. Despite its promise, the implementation of VR in education faces challenges such as high development costs and the need for adequate technological infrastructure. Christian et al.'s systematic literature review on VR in superior education distance learning, especially during the COVID-19 pandemic, underscores VR's growing role in higher education. Their review reveals VR's effectiveness in enhancing learning experiences, motivation, and comprehension in fields like engineering and medicine, predominantly among university students, and that technological advancements have made diverse VR applications possible despite equipment issues and budget constraints [2]. Figueiredo et al. explore VR's impact on elementary education, emphasizing its capacity to create captivating learning experiences for young learners. Platforms like Google Expeditions and Nearpod VR have made complex subjects more accessible, promoting student engagement and empathy. The study reflects on the evolution of VR technology, its increasing affordability, and its potential to revolutionize traditional teaching methods despite content development and teacher training challenges [3]. In higher education, particularly in biomedical sciences, Fabris et al. discuss VR's role in enhancing the visual-spatial understanding of complex anatomical structures. The review presents mixed results from various studies regarding VR's effectiveness, highlighting the

\* Corresponding author.

E-mail addresses: juaizdom@upv.es (J. Izquierdo-Domenech), jlinares@dsc.upv.es (J. Linares-Pellicer), isfermol@upv.es (I. Ferri-Molla).

importance of interactivity in VR applications for effective learning. It also addresses scalability and cost considerations, pointing to the potential of VR as a valuable tool in education when appropriately integrated into curricula [4].

Numerous investigations have underscored VR's capacity to augment educational outcomes by furnishing learners with genuine and captivating learning settings. Ausburn contends that VR constitutes a potent innovative technology for pedagogy and research, facilitating deeper comprehension and reduced training durations [5]. Correspondingly, the works of Alshammari and Lee et al. scrutinize VR's support for collaborative learning, problem-centric pedagogy, and role-playing scenarios [6],[7]. Some inquiries delve into VR's unique ability to grant access to otherwise unreachable experiences. For example, Asad et al. discern that VR grants students first-hand experiences and amplifies experiential learning [8]. Zakaria et al. elucidate how VR affords simulations of remote and perilous locales [9], while Carruth posits that it permits students to interact with expensive equipment and explore intricate problem domains devoid of risk [10]. Additional investigations explore VR's potential to supplement or even supplant real-world experiences. Oiwake et al. introduce the groundbreaking idea of a "VR Classroom," where students experience the sensation of being in a physical classroom [11]. Similarly, Hunvik et al. have created a VR application tailored for a STEM course. Their research concludes that such an application holds potential as a precursor to conventional learning methods [12]. In a complementary fashion, Smutny et al. review VR applications spanning a wide array of academic disciplines, focusing on curricula including medicine, history, engineering, and music [13]. While VR displays considerable promise in enriching learning experiences, certain constraints persist. Asad et al. underscore the considerable implementation costs [8]. Lopez et al. coincide on the high cost of developing VR experiences, although highlighting that VR is an optimal tool for learning, even in professional contexts [14]. Yet, with these exciting advancements and ongoing inquiries, the future of education seems balanced for a transformative journey into the immersive realms of VR, offering both challenges and opportunities for educators and learners alike.

In this evolving educational landscape, the role of Artificial Intelligence (AI) is increasingly significant. As highlighted in "Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4)", AI brings a unique set of opportunities and challenges to the realm of education, with the potential to contribute significantly to achieving Sustainable Development Goal 4 (SDG4) of the UNESCO 2030 Agenda, which emphasizes quality education and lifelong learning opportunities for all. It also emphasizes the ethical considerations and the need for AI to be developed to benefit humanity and respect global norms and standards, making it particularly relevant in the educational context [15]. Empirical investigations conducted in this domain have consistently illuminated the manifold ways AI can be harnessed to ameliorate educational administration, instructional methodologies, and learning outcomes. Notably, AI systems have demonstrated their utility in alleviating the administrative burdens borne by educators. For instance, AI-driven tools have proven instrumental in automating tasks such as assignment grading or personalized teaching, allowing educators to redirect their efforts toward more individualized and engaging endeavors [16]. Furthermore, AI-powered adaptive learning systems have emerged as a pivotal mechanism for tailoring educational curricula and providing content to individual student requisites, thereby supporting student engagement and adapting to specific student needs [17]. The development of virtual classrooms and AI-driven chatbots is concurrently underway, seeking to provide autonomous instruction to students or to function as valuable adjuncts to human educators [18]. Generative AI is a branch of AI focused on

creating algorithms and models that produce human-like data or content. These systems use Deep Learning (DL) to learn patterns from large datasets, enabling them to generate contextually relevant and creative outputs, such as text, images, or music. Generative AI has wide-ranging applications, from text generation to creative arts and data synthesis, and is already being applied in education. Leiker et al. highlight that AI-generated synthetic videos can efficiently replace traditional instructional videos, facilitating the cost-effective and time-efficient production of high-quality educational content [19]. Bekeš et al. discovered that AI-generated content is favored by teachers over conventional materials, primarily due to its adaptability and flexibility [20]. However, it is imperative to acknowledge that the advent of AI has prompted discourse regarding its potential to redefine the role of educators. Some studies posit that AI's increasing integration into the educational milieu may gradually transition teachers from traditional lecturers into facilitators as AI assumes instructional responsibilities [21]. Du Boulay argues in favor of enhancing human educators with AI, suggesting that AI can serve as a personalized tutor when necessary, allowing human teachers to concentrate on the broader classroom context [22]. Yang predicts that AI and VR will significantly impact education in the coming years [23]. While AI can detect students' weaknesses and tailor instruction to their needs, VR can foster students' interest and social development.

The potential of technology in education is vast and encompasses a range of innovative tools and methods. While AI plays a crucial role in enhancing educational experiences, it is not the sole driving force. Alongside AI, emerging technologies like VR and generative AI are becoming transformative factors in the educational landscape.

While the application of VR in education has been extensively studied, its combination with Large Language Models (LLMs) represents a novel frontier that holds significant promise for further revolutionizing learning methodologies. The present proposal focuses on integrating VR technology and generative AI to tackle a significant educational challenge: providing rapid and contextually accurate access to information. The system empowers students to access information through Question Answering (QA) mechanisms, offering a unique approach to enhancing comprehension, even in complex laboratory settings. By harnessing the immersive capabilities of VR and the data synthesis abilities of generative AI, this proposal represents an exciting synergy of technological advancements that have the potential to revolutionize education. This research aims to bridge the gap between the immersive experiences provided by VR and the advanced capabilities of LLMs in processing and generating human-like text. The synergy between VR's interactive environments and LLMs' ability to understand and respond to Natural Language (NL) queries presents an unprecedented opportunity to create more engaging, personalized, and effective learning experiences. Our study is positioned at this intersection, exploring how the integration of VR with LLMs can enhance the learning process, particularly in settings where traditional educational methods may fall short, thus filling the gap of research on the combined use of VR and LLMs.

A critical aspect of this system's functionality is using generative AI techniques to generate responses based solely on contextual information, reducing the risk of producing inaccurate or fictitious information. This contextual information can take various forms, such as text documents or .pdf files, with the adaptable library providing access to a wide range of alternative data sources, including databases, spreadsheet files, and even Application Programming Interfaces (APIs). To overcome the challenges associated with physical laboratory access, including scheduling and logistical constraints, VR technology, combined with 360° photos, has been chosen to represent complex environments like laboratories and shopfloors, each containing diverse points of interest. With this approach, the evaluated system empowers



users to articulate queries in NL, allowing them to receive responses to their original questions. Moreover, as these answers are derived from contextual knowledge, the application seamlessly guides the users' attention to the relevant elements of interest in the VR environment associated with their questions.

The article is structured into distinct sections, each addressing specific aspects of the research. First, in Section II, the article explores the potential impact of LLMs and generative AI in education. Next, Section III delves into Retrieval-Augmented Generation (RAG) methods and their significance in contextual information retrieval. Subsequently, Section IV explains the system's implementation, including server-side and client-side components. The critical phases of system evaluation are covered in Section V, and a comprehensive examination of limitations is presented in Section VI. Finally, the conclusions are drawn in Section VII.

## II. LLM IN EDUCATION

Large Language Models (LLMs) and generative AI are emerging as transformative educational tools, automating and enhancing various educational processes. While they offer significant advantages in generating high-quality educational content and analyzing student responses, they also present challenges. LLMs can exhibit biases inherited from their training data, leading to ethical concerns. Their lack of deep understanding can result in superficial or inaccurate content, and there is a risk of student overreliance on these models, which may impede the development of critical thinking skills. Additionally, their operation requires considerable computational resources, posing a barrier in some educational settings. Numerous studies have delved into the utilization of LLMs for the generation of high-quality educational content at scale, ranging from programming exercises and code explanations [24] to the creation of comprehensive multimedia course materials [25]. Through techniques like clustering and summarization, LLMs facilitate the rapid and accurate identification of underlying themes and patterns within student responses, surpassing the capabilities of manual analysis alone [26]. Nevertheless, it remains imperative to incorporate human oversight and review mechanisms to ensure the accuracy and reliability of these AI-generated resources before they are made available to students [27]. While the automated generation of educational materials promises to reduce instructors' workload significantly, addressing practical and ethical concerns associated with integrating LLMs into educational settings is essential. A comprehensive analysis of 118 research papers revealed that LLMs have been applied across 53 distinct educational use cases, encompassing tasks such as grading, teaching support, content generation, and recommendation [27]. Although LLMs exhibit the potential to automate and enhance these educational functions, their performance, transparency, privacy implications, commitment to equality, and ethical considerations must be evaluated to ascertain their suitability for educational contexts. Furthermore, LLMs offer a promising avenue for gaining insights into student learning by conducting in-depth analyses of student-generated artifacts, such as essays.

## III. RAG FOR CONTEXTUAL INFORMATION RETRIEVAL

RAG methods have recently gained significant interest since they allow to combine neural generation models (i.e., parametric memory) with contextual information (i.e., non-parametric memory), as depicted in Fig. 1. RAG is an approach in Natural Language Processing (NLP) that combines the power of language models with information retrieval, enabling the generation of more informed and contextually relevant responses by dynamically fetching and integrating external knowledge sources during the generation process. Numerous articles

have explored RAG models for open-domain question answering and found that they can achieve state-of-the-art performance. Lewis et al. introduced a general RAG recipe, showing RAG models outperform parametric seq2seq models and task-specific architectures on knowledge-intensive NLP tasks like open-domain QA [28]. Ranjit et al. built on this work, proposing a RAG model for radiology report generation that achieved the best metrics [29]. While early RAG work focused on retrieving text, recent papers have expanded to multimodal knowledge. Yu discussed obstacles to single-source retrieval and provided solutions for RAG over heterogeneous knowledge [30]. Chen et al. introduced the first multimodal RAG, accessing images and text to answer questions [31]. Zhao et al. surveyed RAG methods across modalities, reviewing image, code, table, graph, and audio retrieval for generation [32]. Some work has aimed to improve RAG domain adaptation. Siriwardhana et al. proposed an end-to-end trained RAG variant with an auxiliary loss for reconstructing sentences from retrieved knowledge [33]. They showed significant gains in adapting RAG to COVID-19, news, and conversation domains. Finally, Mao et al. presented an alternative approach: Generation-Augmented Retrieval (GAR) [34]. GAR uses generation to expand queries before retrieving relevant passages. On open-domain QA, GAR with sparse retrieval matched or outperformed dense retrieval methods, achieving state-of-the-art extractive QA performance.

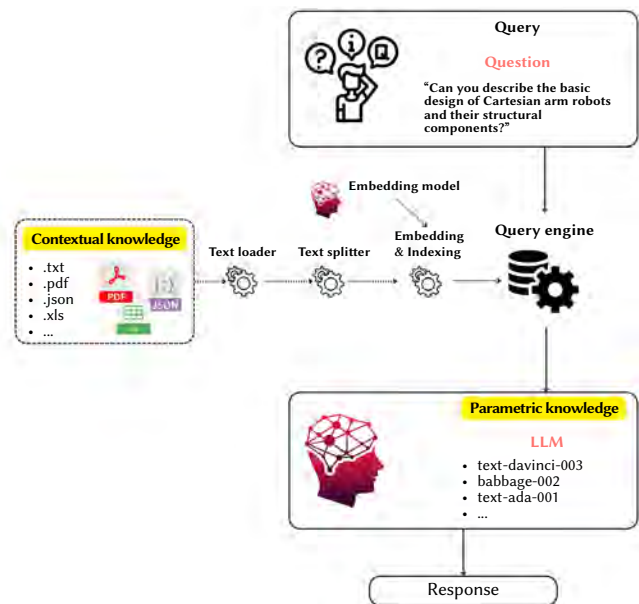


Fig. 1. RAG complements parametric knowledge with contextual knowledge.

The RAG methodology is a powerful tool for tailoring NLP and generation to specific domains, such as education. It offers a unique advantage over techniques such as few-shot learning by constraining generated responses to verified information, effectively reducing the risk of "hallucination" or generating incorrect or irrelevant answers. This feature ensures that students receive accurate and consistent information, enhancing the overall learning experience.

Few-shot learning, while valuable, has limitations, particularly concerning the maximum token size for prompts. This constraint can lead to less accurate or incomplete answers. RAG addresses this issue by using semantic similarity to pull the most relevant information from a given context, thereby creating a more precise final prompt. This ensures the generated answer is accurate and relevant to the student's query.

However, it is essential to consider the limitations of fine-tuning, especially in QA environments. While fine-tuning offers granular

control over model behavior, it often requires substantial training data and computational resources. More critically, fine-tuning can be less reliable in generating precise answers to specific questions, as it does not inherently constrain the model's responses to verified information. This makes it less suitable for applications where the accuracy of each individual answer is paramount, such as educational settings where incorrect information could have enduring impacts. RAG's efficiency and focus on accuracy offer a more reliable alternative in these contexts.

#### IV. SYSTEM IMPLEMENTATION

This section is dedicated to explaining the specific implementation that was carried out for the evaluated system. The architecture presented here is based on a client-server model, with the server responsible for processing LLMs and contextual data to respond to user queries and the client serving as a VR interface for users to explore educational environments and pose NL questions.

##### A. Server-Side Implementation

The server-side implementation is responsible for tasks and actions related to the processing of LLMs and, crucially, the use of RAG for extracting specific contextual information. In this implementation, the LlamaIndex library has been employed due to its ability to provide an interface enabling developers to work with various LLMs, such as gpt-3.5 or text-davinci-003 [35]. Furthermore, LlamaIndex facilitates the execution of RAG, which means that it is possible to enhance the parametric knowledge of the LLM with contextual information. Fig. 1 details how contextual information is accessed to enable RAG.

In the analyzed context, the contextual information is based on text documents in formats such as .pdf and .txt, which describe various elements present in a classroom (e.g., an Angular arm robot or a Cartesian arm robot). However, the flexibility of LlamaIndex allows access to a broad spectrum of alternative data sources, encompassing databases, spreadsheet files, and even Application Programming Interfaces (APIs). Besides, a .json file is utilized to define the relationships between the elements in a classroom and the VR scenes located on the client side, as shown in the appendix in Listing 1. Further details can be found in subsection IV.B.

To facilitate RAG's utilization of contextual knowledge exclusively, LlamaIndex incorporates the concept of an Index. Illustrated in Fig. 1, the indexing stage assumes responsibility for allowing rapid access to relevant context for a user query. These generated indexes streamline the retrieval process, automating vector embedding calculations. While the VectorStoreIndex is a prevalent index type, the system's preference in this instance is the KeywordTableIndex. This choice aligns with the system's approach, wherein each node (i.e., each textual chunk produced during the text splitting task) additionally factors in specific keywords. During a query operation, the node selection containing the relevant text chunk is determined based on keywords extracted from the query, enhancing answer reliability.

The requests that the server handles are summarized in the following routes:

1. GET /summaries This route returns a list of summaries organized by scene, briefly explaining the elements present in each of them, as illustrated in Fig. 2.
2. POST /query When requesting this route, based on the user's query, a response is returned based solely on the available contextual information. In addition to the response, a unique identifier of the queried element, the name of the element of interest, and the VR scene in which it is located are provided.

When a user submits a question through the immersive application (refer to Section IV.B), the query is transmitted to the POST /query

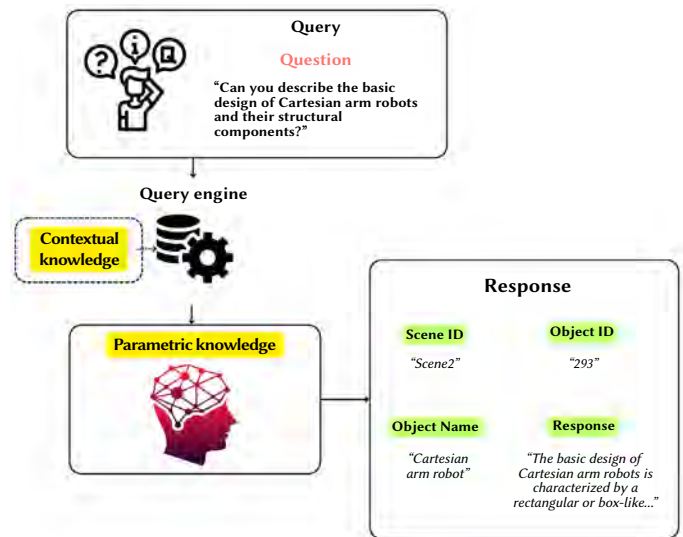


Fig. 2. Example of a GET /query request diagram.

endpoint. Here, a two-step process is employed to enhance the system's contextual understanding and ensure meaningful responses.

First, we use prompting to integrate the user's question, leading to a format like this:

Answer the question using ONLY THE CONTEXT, and if you're not TOTALLY sure of the answer, say 'Sorry, I don't know'. Q: {question} A:

This approach compels the system to rely on contextual knowledge to answer the question effectively. To pinpoint the specific index or chunk of text where the answer resides, the SubQuestionQueryEngine, available in LlamaIndex, is used.

Once a valid response (i.e., an answer different from 'Sorry, I don't know') is obtained, the system follows up with another prompt:

Based on the question "{original\_question}" and its answer "{query\_answer}", please return an answer in the format "scene\_id:\_,object\_id:\_,object\_name:\_" If you don't know the answer, respond "scene\_id:N/A,object\_id:N/A,object\_name:N/A" A:

The success of this query to the LLM dramatically depends on the .json file that establishes relationships between scenes and elements of interest. In this regard, the results achieved have been quite promising.

Eventually, in response to the user's query, they receive the answer and identifiers for the scenes and objects in question, which they can utilize in the client-side application.

Although RAG techniques, as used in the project, significantly improve LLM performance in front of users' questions, RAG cannot improve the current well-known limitations in reasoning questions or multi-hop questions on documents that are still part of current LLM solutions [36].

The underlying technology in the server development is based on Python 3.10.11 as the primary programming language, with FastAPI and Uvicorn for implementing the REST server. To access the content of contextual information in PDF format, the PyPDF library is utilized.

##### B. Client-Side Implementation

Unity was chosen as the development engine for the client-side component due to its outstanding capabilities in creating cross-platform applications. Specifically, it has enabled the efficient and effective development of VR applications. In order to optimize the cost associated with VR application development, the decision was made to employ an accessible yet entirely valid technique for exploring

a specific environment, namely, an educational laboratory. This technique uses 360° photographs to circumvent the complexities associated with 3D modeling and physical laboratory access, including scheduling and logistical constraints. The VR device used to deploy the evaluated system was the Meta Quest 2; nevertheless, Unity's cross-platform architecture enables executing the same application to similar devices, such as the Pico VR or the HTC Vive.

The VR application comprises several scenes, each hosting various points of interest. In the example of the scene depicted in Fig. 3, the most prominent element is a Cartesian robotic arm; however, it is essential to note that the system is adaptable and can accommodate different points of interest in the same scene, and spread among different scenes. All this information must be explicitly detailed in the .json file described in Subsection IV.A (Listing 1 in the appendix).



Fig. 3. VR scene with a Cartesian arm robot in the middle.

Users can pose questions in NL using speech recognition during the virtual environment exploration. These questions are transmitted to the server through the GET /query request. Suppose the sought-after information is part of the contextual knowledge (described in Section III). In that case, the provided response will include the requested information and metadata related to the point of interest and the user's current scene. Consequently, students are not obliged to be in the VR scene containing the point of interest they inquire about; they can ask about any point of interest encompassed by the contextual knowledge. Fig. 4 visually represents the user's post-response perspective. Notably, alongside the textual answer, a visual cue is strategically employed to direct the user's attention towards the specific element relevant to the initial question.

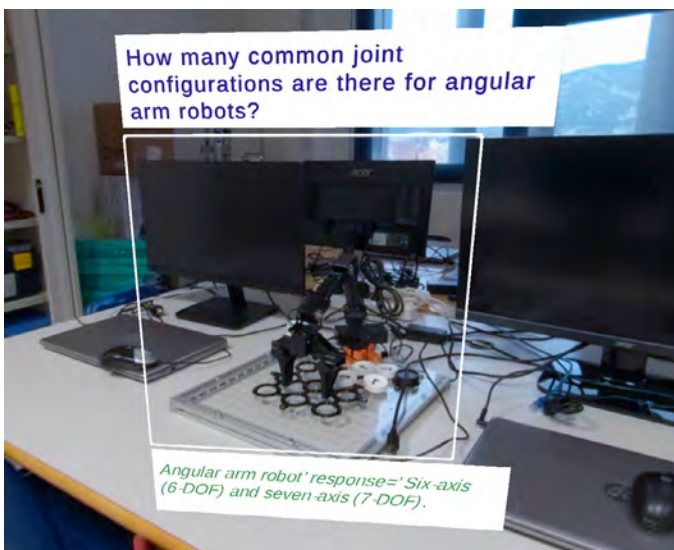


Fig. 4. In VR, user questions trigger dual feedback - textual responses and visual cues, so students get textual answer and VR interaction.

## V. SYSTEM EVALUATION

In this study, 20 participants divided into two distinct groups were selected to evaluate the effectiveness of our architecture in enhancing the learning process. This sample size was determined based on the available resources, the innovative nature of the technology involved, and the need for in-depth interaction with each participant. While a larger sample could provide more generalizable results, this exploratory study's specific constraints and focus guided this decision. The first group, referred to as the 'Experimental group,' had access to the immersive VR application, which empowered participants to interact with their educational environment, pose questions, and receive informative responses in text and as visual cues in VR. The second group, termed the 'Control group,' followed a more traditional learning approach devoid of VR technology, where participants relied on conventional methods to access educational content and resources, such as PDF files. By juxtaposing these two groups, this study aims to discern the transformative potential of our VR and LLM architecture compared to established learning practices, thus providing a comprehensive evaluation of its impact across diverse educational contexts.

In this study, Robotics was selected as the primary subject for system evaluation due to its relevance in modern education and the potential to benefit from VR and LLM technologies. The practical nature of robotics, involving theoretical knowledge and hands-on skills, makes it an ideal candidate for the presented educational architecture. The application was pivotal in bridging the gap between theory and practice. It enabled students to understand complex robotics concepts and later apply them to manipulating and controlling the robot arms. Integrating immersive VR experiences with enriched theoretical insights by LLMs illustrates the system's capability to offer a comprehensive learning experience, particularly in subjects where practical skills are as crucial as theoretical knowledge. In class, students have access to a variety of robotic arms, including angular and cartesian types, which they were required to manipulate after completing the necessary learning modules.

To ensure the accuracy and trustworthiness of the presented system within the limitations of RAG systems, as explained in section III, a set consisting of 10 questions per type of robot was defined. These questions were designed to cover an overall spectrum of topics, including definitions, historical backgrounds, design features, and applications. Using prompting strategies to reduce hallucinations and given that all the answers to the questions could be found within the provided contextual knowledge, the system presented a high accuracy rate in delivering correct responses. This testing protocol ensured a comprehensive evaluation of the system's capacity to handle varied inquiries and affirmed its effectiveness in providing precise and relevant information.

The participants in this study had some prior familiarity with the subject matter as they had been students in a course that involved working with robots; however, they had not previously interacted with the specific robots featured in the VR setting. The system evaluation took place during four 2-hour sessions outside of regular class hours, during which participants received training on how to use the robots (both groups), familiarized themselves with the system (Experimental group), and completed the tests (both groups). It is important to remark that the primary purpose of this experiment was to learn how to operate the robots and to compare the Control group with the Experimental group.

Before engaging with the VR and LLM application or the traditional learning method, participants completed pre-tests to establish their baseline knowledge and skills in the subject matter. While it was acknowledged that some participants might have had limited prior exposure to the subject, the pre-tests captured their initial understanding. Subsequently, post-tests were administered after



TABLE I. DESCRIPTIVE AND STATISTICAL CONTRASTS

Group	Score		Mean difference ( <i>p</i> -value)	Intra-subject Effects Tests	
	Pre-test	Post-test		Score	Score*Group
	Mean (Sd)	Mean (Sd)		<i>F</i> (d.f.); <i>p</i> -value ( $\eta^2$ )	<i>F</i> (d.f.); <i>p</i> -value ( $\eta^2$ )
Control	1.50 (1.08)	5.80 (1.23)	-4.3 (<0.001)	<i>F</i> (1;18) = 239.63; <i>p</i> < 0.001 (0.93)	<i>F</i> (1;18) = 8.53; <i>p</i> < 0.009 (0.322)
Experimental	1.30 (0.95)	7.60 (0.97)	-6.3 (<0.001)		
Mean difference ( <i>p</i> -value)	0.2 (0.455)	-1.8 (0.002)			

d.f.: degrees of freedom.  $\eta^2$ : partial eta-squared (effect size)

participants had interacted with their respective learning methods. The post-tests allowed to measure the extent of learning gains and the overall impact of this educational approach. By comparing pre-test and post-test results, it was possible to evaluate the system's effectiveness in fostering learning and comprehension, even among those with little prior knowledge.

To ensure that participants in the Experimental group engage effectively with the application, participants were encouraged to explore the application at their own pace while highlighting the significance of thorough knowledge acquisition. They were informed about the availability of a diverse range of learning resources within the application. They were guided on how navigating and asking questions was performed and the contents they needed to review for the subsequent assessment. A standardized VR experience across all participants in the Experimental group was ensured. Each participant used the same VR hardware and software configurations to minimize variability in the quality of the VR experience. Additionally, the technological background of each participant was assessed through a pre-study questionnaire. This assessment helped understand the participants' familiarity and comfort with VR and other digital technologies, which could influence their interaction with the VR environment.

In Table I, the result of the two-way repeated measures ANOVA test is displayed to determine whether the methodology influences the scores obtained in the test. The result shows a statistically significant difference between the pre-test and post-test, regardless of the group. However, the interaction between the group and the score was significant, indicating that the test scores depend on the group. Thus, the students in the Experimental group significantly increased their scores in the post-test compared to the pre-test, as did the Control group, although to a lesser extent. In the beginning, no differences were observed between the groups. In contrast, at the end of the study (i.e., post-test), the scores of the students in the Experimental group were significantly higher than those of the Control group. Fig. 5 displays the evolution of the scores of the groups.

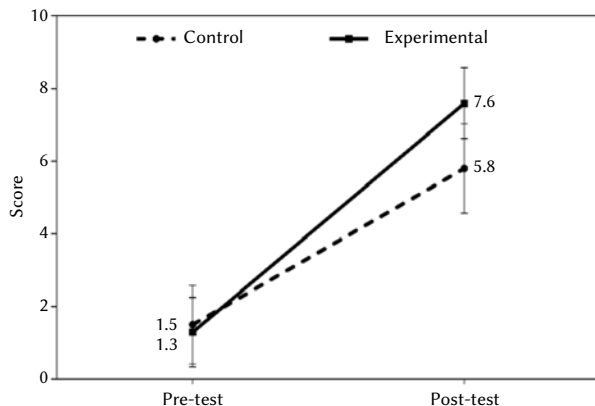


Fig. 5. Score evolution per group.

A Likert-scale questionnaire was also administered to evaluate user satisfaction, chosen for its versatility and effectiveness in capturing nuanced user sentiments. The Likert-scale questionnaire provides a structured, user-friendly format that encourages participants to express their opinions across various dimensions, accommodating diverse user backgrounds and preferences. This inclusivity makes it a valuable tool for assessing user satisfaction in the context of our transformative educational technology. The questionnaire results can be found in the appendix, specifically in Table III, along with the questionnaire questions in Table II. In terms of satisfaction, the median score for students in the Control group was 1.5 (*IQR* = 1-2), while students in the Experimental group scored 5 points (*IQR* = 4-5). The Mann-Whitney U test for independent samples revealed that the difference in satisfaction between students in the Experimental group was significantly higher than that of students in the Control group ( $U = 0$ ,  $p < 0.001$ ).

TABLE II. USER EXPERIENCE QUESTIONNAIRE

Index	Question
Q1	I found the learning experience engaging and immersive.
Q2	The learning materials/methods provided me with valuable information and learning opportunities.
Q3 (Experimental)	Using the VR application improved my understanding of the subject.
Q3 (Control)	The traditional learning materials/methods improved my understanding of the subject matter.
Q4 (Experimental)	I felt more confident in applying the knowledge gained through the system.
Q4 (Control)	I felt more confident in applying the knowledge gained through the traditional learning methods.
Q5	Overall, I am satisfied with my learning experience.

## VI. DISCUSSION AND LIMITATIONS

In exploring the integration of AI in educational contexts, it is crucial to consider both the transformative potential and the challenges posed by these technologies. As García et al. highlight in their study, the emergence of tools like ChatGPT has significantly influenced teaching and learning processes, raising important questions about AI's biases, ethical considerations, and social implications in education. Their work underscores the need for a nuanced understanding of AI's role in education, balancing its benefits with a critical awareness of its limitations and potential risks [37].

The presented architecture, which integrates VR technology with LLMs for educational purposes, has shown promising results in enhancing the learning process. The significant improvement is evident from the post-test scores of the Experimental group compared to the

TABLE III. LIKERT-SCALE QUESTIONNAIRE

Participant	Group	Q1	Q2	Q3	Q4	Q5
1	Experimental	5	4	5	4	5
2	Experimental	5	3	5	5	4
3	Experimental	5	5	5	5	5
4	Experimental	4	5	5	5	5
5	Experimental	5	5	5	4	5
6	Experimental	3	4	4	4	4
7	Experimental	4	4	5	4	4
8	Experimental	5	3	5	5	5
9	Experimental	5	5	4	4	5
10	Experimental	4	5	5	4	5
11	Control	1	3	4	2	2
12	Control	1	2	3	4	1
13	Control	2	3	4	3	3
14	Control	1	3	4	2	1
15	Control	3	4	4	3	2
16	Control	2	3	3	3	1
17	Control	3	2	3	3	2
18	Control	1	4	2	2	2
19	Control	1	4	4	2	1
20	Control	2	2	4	3	1

Control group. The substantial increase in post-test scores among the Experimental group suggests that an immersive learning experience, coupled with the assistance of LLMs, can effectively foster knowledge acquisition and comprehension, even among participants with limited prior knowledge of the subject matter; however, several noteworthy limitations must be acknowledged to provide a more comprehensive understanding of the system's potential and its impact on education.

- The effectiveness of this approach heavily depends on the quality and comprehensiveness of the contextual knowledge provided to the LLM. Incomplete or inaccurate contextual knowledge may lead to suboptimal responses to user queries. Ensuring the accuracy and relevance of the information fed into the system is critical for its success.
- The success of this system also depends on the availability of appropriate 360° photographs and the accurate mapping of points of interest within VR scenes.
- The study design involved two distinct groups: the Experimental group and the Control group. While the Experimental group experienced the immersive VR and LLM-based learning environment, the Control group followed traditional learning methods. This design might introduce biases related to individual learning preferences and engagement levels. Some participants in the Control group may have needed more motivation due to the absence of the novel VR experience, potentially affecting their post-test performance. The novelty of the VR and LLM integration in the Experimental group might have influenced the motivation and engagement levels, which could affect learning outcomes. Future iterations of the research will aim to equalize engagement potential between the groups. Therefore, it is essential to consider the potential impact of participant motivation and engagement as a limitation when interpreting the study's results.
- The limited number of participants could influence the generalizability of our findings. However, it is noteworthy that similar exploratory studies in VR and LLM integration have also operated with small sample sizes. Future studies with larger samples are necessary to confirm these initial observations and to understand the broader implications of VR and LLMs in educational settings.

- Users unfamiliar with VR technology may face a learning curve when using the system, necessitating adequate training and guidance to ensure a smooth educational experience. Although measures were taken to standardize the VR experience and assess the participants' technological backgrounds, the potential impact of technological issues must be considered. Variations in individual comfort levels and familiarity with VR technology may have influenced the results. Although no significant technological issues were reported during the study, future research should further explore the role of technological familiarity in the effectiveness of VR-based educational interventions.
- While 360° photographs offer a cost-effective means of environment representation, they might not capture all aspects of a physical laboratory, potentially limiting users' tactile and spatial experiences.
- Response times for user queries may fluctuate depending on query complexity and contextual data volume, leading to occasional delays in receiving responses.
- Scalability becomes a crucial consideration in terms of time allocation, the availability of physical resources such as VR headsets, and the number of students, pointing out that as long as these essential materials and resources are accessible, the proposed system can effectively accommodate a growing number of participants linearly without experiencing a substantial decline in performance. This attribute holds significant importance as it ensures that educational institutions can seamlessly expand the adoption of immersive VR and LLM technologies to reach a broader student audience, enhancing the scalability and widespread applicability of innovative educational methodologies.

These limitations underscore the need for ongoing refinement, quality assurance, user support, and research efforts to optimize the system's educational utility and ensure its effectiveness in diverse educational settings. Future research should also explore strategies to mitigate biases in study designs and improve the overall user experience within this innovative educational framework.

## VII. CONCLUSIONS

Integrating VR technology and generative AI provided by LLMs within the educational landscape represents a transformative approach to learning. This research has explored the potential of combining VR and generative AI to address the challenge of providing rapid and contextually accurate access to information. Through developing an immersive VR application and using RAG, this proposal has demonstrated the ability to enhance comprehension and learning outcomes. Indeed, as highlighted by Garcia et al., the ongoing evolution and refinement of generative AI technologies, including those used in our VR application, are rapidly shaping the future of education, promising transformative changes and novel approaches in teaching and learning methodologies [38].

The findings from the system evaluation suggest a clear advantage for the Experimental group, which had access to the immersive VR and LLM-based learning environment, over the Control group that followed traditional learning methods. The significant improvement in post-test scores among the Experimental group highlights the effectiveness of this innovative approach in fostering knowledge acquisition and comprehension, even when the participants had limited prior knowledge of the subject matter. Moreover, the user satisfaction scores from the Experimental group were significantly higher, underlining the appeal and user-friendliness of this novel educational system.

Nevertheless, it is essential to acknowledge the limitations identified in this research. The system's success is contingent on the accuracy and comprehensiveness of the contextual knowledge provided to the AI models. Incomplete or inaccurate information may lead to suboptimal responses. Besides, the availability of appropriate 360° photographs and accurate establishment of points of interest within VR scenes are critical for the system's effectiveness.

While this study provides insightful initial findings on the integration of VR and generative AI in education, it is important to note that the use of RAG was limited to handling text-based data without multimodal integration from the VR environment. This focus is due to the scope of the current research. However, exploring integrating multimodal data, such as visual inputs from VR, into RAG is a promising direction for future work. The small sample size in this study limits generalizability. Hence, future research with larger, more diverse participant groups needs to be done. Such studies would validate and expand upon the findings and explore the full potential of multimodal VR and LLM systems in educational contexts. As technology advances, addressing these limitations, education is poised for transformation through immersive VR, presenting both challenges and opportunities. Continued research and development are essential for realizing this transformative potential in global education.

#### APPENDIX

**Listing 1:** .json file linking scenes with elements

```
{
  "scenes": [
    {
      "scene_id": "Scene1",
      "objects": [
        {
          "object_id": 776,
          "object_name": "Angular arm robot"
        },
        // Other objects in Scene 1
      ]
    },
    {
      "scene_id": "Scene2",
      "objects": [
        {
          "object_id": 293,
          "object_name": "Cartesian arm robot"
        }
      ]
    },
    // Other scenes , with their objects
  ]
}
```

#### REFERENCES

- [1] A. Cordero, C. J. Lluch, E. Sanabria Codesal, J. Torregrosa, "Towards a Better Learning Models Through OCWs and MOOCs," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 3, no. 4, pp. 26–30, 2015, doi: 10.9781/ijimai.2015.345.
- [2] B. Christian, C. Salvador, G. Christian, "Virtual Reality (VR) in Superior Education Distance Learning: A Systematic Literature Review," *JOIV: International Journal on Informatics Visualization*, vol. 5, no. 3, pp. 264–270, 2021.
- [3] M. Figueiredo, R. Mafalda, A. Kamensky, "Virtual reality as an educational tool for elementary school," in *Smart Innovation, Systems and Technologies*, vol. 198 SIST, 2021, pp. 261–267, Springer. doi: 10.1007/978-3-030-55374-6\_26.
- [4] C. P. Fabris, J. A. Rathner, A. Y. Fong, C. P. Sevigny, "Virtual Reality in Higher Education," *International Journal of Innovation in Science and Mathematics Education*, vol. 27, no. 8, pp. 69–80, 2019.
- [5] L. J. Ausburn, F. B. Ausburn, "Desktop virtual reality: A powerful new technology for teaching and research in industrial teacher education," *Journal of Industrial Teacher Education*, vol. 41, no. 4, pp. 1–16, 2004.
- [6] S. H. Alshammari, "The Role of Virtual Reality in Enhancing Students' Learning," *International Journal of Educational Technology and Learning*, vol. 7, no. 1, pp. 1–6, 2019, doi: 10.20448/2003.71.1.6.
- [7] E. Ai-Lim Lee, K. Wai Wong, "A Review of Using Virtual Reality for Learning," *Transactions on edutainment I*, pp. 231–241, 2008.
- [8] M. M. Asad, A. Naz, P. Churi, M. M. Tahanzadeh, "Virtual Reality as Pedagogical Tool to Enhance Experiential Learning: A Systematic Literature Review," *Education Research International*, vol. 2021, pp. 1–17, 2021, doi: 10.1155/2021/7061623.
- [9] G. Zakaria, S. Wilkie, "Applications for virtual reality experiences in tertiary education," *ASCILITE Publications*, pp. 186–193, 2020.
- [10] D. W. Carruth, "Virtual reality for education and workforce training," in *2017 15th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, 2017, pp. 1–6.
- [11] K. Oiwake, K. Komiya, H. Akasaki, T. Nakajima, "VR classroom: enhancing learning experience with virtual classrooms," in *2018 Eleventh International Conference on Mobile Computing and Ubiquitous Network (ICMU)*, 2018, pp. 1–6.
- [12] S. R. B. Hunvik, F. Lindseth, "Making Use of Virtual Reality for Artificial Intelligence Education," in *Bridges and Mediation in Higher Distance Education: Second International Workshop, HELMeTO 2020, Bari, BA, Italy, September 17–18, 2020, Revised Selected Papers 2*, vol. 1344 of *Communications in Computer and Information Science*, Cham, 2021, pp. 56–70, Springer International Publishing. doi: 10.1007/978-3-030-67435-9.
- [13] P. Smutny, M. Babiuch, P. Foltynek, "A review of the virtual reality applications in education and training," in *2019 20th International Carpathian Control Conference (ICCC)*, 2019, pp. 1–4.
- [14] M. A. Lopez, S. Terron, J. M. Lombardo, R. Gonzalez-Crespo, "Towards a solution to create, test and publish mixed reality experiences for occupational safety and health learning: Training-MR," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 212–223, 2021, doi: 10.9781/ijimai.2021.07.003.
- [15] J. M. Flores-Vivar, F. J. García-Peñalvo, "Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4)," *Comunicar: Revista científica de comunicacion y educacion*, vol. 31, no. 74, pp. 37–47, 2023, doi: 10.3916/C74-2023-03.
- [16] L. Chen, P. Chen, Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.
- [17] M. Fahimirad, S. S. Kotamjani, "A Review on Application of Artificial Intelligence in Teaching and Learning in Educational Contexts," *International Journal of Learning and Development*, vol. 8, p. 106, 12 2018, doi: 10.5296/ijld.v8i4.14057.
- [18] J. Huang, S. Saleh, Y. Liu, "A review on artificial intelligence in education," *Academic Journal of Interdisciplinary Studies*, vol. 10, pp. 206–217, 5 2021, doi: 10.36941/AJIS-2021-0077.
- [19] D. Leiker, A. R. Gyllen, I. Eldesouky, M. Cukurova, "Generative AI for learning: Investigating the potential of synthetic learning videos," *arXiv preprint arXiv:2304.03784*, 2023.
- [20] E. R. Bekeš, V. Galzina, "Utilizing smart digital technology and artificial intelligence in education for transforming the way content is delivered," in *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2022, pp. 573–578, IEEE.
- [21] K. Mondal, "A Synergy of Artificial Intelligence and Education in the 21st Century Classrooms," in *2019 International Conference on Digitization (ICD)*, 2019, pp. 68–70.
- [22] B. Du Boulay, "Artificial intelligence as an effective classroom assistant," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 76–81, 2016.
- [23] H. Yang, "Explore How Artificial Intelligence and VR Technology



will Change the Development of Future Education,” *Journal of Physics: Conference Series*, vol. 1744, no. 4, 2021, doi: 10.1088/1742-6596/1744/4/042146.

- [24] S. Sarsa, P. Denny, A. Hellas, J. Leinonen, “Automatic Generation of Programming Exercises and Code Explanations Using Large Language Models,” in *ICER 2022 - Proceedings of the 2022 ACM Conference on International Computing Education Research*, vol. 1, 8 2022, pp. 27–43, Association for Computing Machinery, Inc. doi: 10.1145/3501385.3543957.
- [25] D. Leiker, S. Finnigan, A. Ricker Gyllen, M. Cukurova, “Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale,” *arXiv preprint arXiv:2306.01815*, 2023.
- [26] A. Katz, U. Shakir, B. Chambers, “The Utility of Large Language Models and Generative AI for Education Research,” *arXiv preprint arXiv:2305.18125*, 5 2023.
- [27] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, D. Gašević, “Practical and Ethical Challenges of Large Language Models in Education: A Systematic Literature Review,” *British Journal of Educational Technology*, 3 2023, doi: 10.1111/bjet.13370.
- [28] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [29] M. Ranjit, G. Ganapathy, R. Manuel, T. Ganu, “Retrieval Augmented Chest X-Ray Report Generation using OpenAI GPT models,” *arXiv preprint arXiv:2305.03660*, 5 2023.
- [30] W. Yu, “Retrieval-augmented generation across heterogeneous knowledge,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 2022, pp. 52–58.
- [31] W. Chen, H. Hu, X. Chen, P. Verga, W. W. Cohen, “MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text,” *arXiv preprint arXiv:2210.02928*, 10 2022.
- [32] R. Zhao, H. Chen, W. Wang, F. Jiao, X. L. Do, C. Qin, B. Ding, X. Guo, M. Li, X. Li, S. Joty, “Retrieving Multimodal Information for Augmented Generation: A Survey,” *arXiv preprint arXiv:2303.10868*, 3 2023.
- [33] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rajib, S. Nanayakkara, “Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023, doi: 10.1162/tacl.
- [34] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, W. Chen, “Generation-Augmented Retrieval for Open-domain Question Answering,” *arXiv preprint arXiv:2009.08553*, 9 2020.
- [35] J. Liu, “LlamaIndex,” 11 2022. [Online]. Available: [https://github.com/jerryliu/llama\\_index](https://github.com/jerryliu/llama_index).
- [36] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” *arXiv preprint arXiv:2309.15217*, 9 2023.
- [37] F. J. García Peñalvo, F. Llorens-Largo, J. Vidal, “The new reality of education in the face of advances in generative artificial intelligence,” *RIED-Revista Iberoamericana de Educación a Distancia*, vol. 27, pp. 9– 39, 7 2023, doi: 10.5944/ried.27.1.37716.
- [38] F. García-Peñalvo, A. Vázquez-Ingelmo, “What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 7–16, 2023, doi: 10.9781/ijimai.2023.07.006.



Jordi Linares-Pellicer

Jordi Linares Pellicer is an Associate Professor at Universitat Politècnica de València (UPV, Spain), where he leads the VertexLit research group at the Valencian Research Institute for Artificial Intelligence (VRRAIN). He received his Ph.D. in Computer Science from UPV and holds a Master’s degree in Artificial Intelligence from Universidad Internacional de La Rioja (UNIR, Spain).



Isabel Ferri-Molla

Isabel Ferri Mollá is currently pursuing her Master's Degree in Artificial Intelligence, Pattern Recognition, and Digital Imaging at Universitat Politècnica de València (UPV, Spain). She received her Bachelor's Degree in Computer Science Engineering from UPV in 2022. Her research interests include areas of artificial intelligence, augmented reality, and human-computer interaction.



Juan Izquierdo-Domenech

Juan Jesús Izquierdo Doménech is an Adjunct Professor of Computer Science in Universitat Politècnica de València. He received his Bachelor's degree in Computer Science Engineering from Universitat Politècnica de València (UPV, Spain) and holds a Master’s degree in Multimedia Applications from Universitat Oberta de Catalunya (UOC, Spain). He is currently performing his

Ph.D. studies in UPV in the field of Human-Computer Interaction, Mixed Reality, and Artificial Intelligence.

# Generative Artificial Intelligence in Product Design Education: Navigating Concerns of Originality and Ethics

Kristin A. Bartlett<sup>1\*</sup>, Jorge D. Camba<sup>2</sup>

<sup>1</sup> University of Kentucky, Lexington, KY (USA)

<sup>2</sup> Purdue University, West Lafayette, IN (USA)

Received 21 August 2023 | Accepted 13 February 2024 | Published 20 February 2024



## ABSTRACT

Image-generative artificial intelligence (AI) is increasingly being used in the product design process. In this paper, we present examples of how it is being used and discuss the possibilities of how applications may evolve in the future. We discuss the legal and ethical implications of image-generative AI, including concerns about bias, hidden labor, theft from artists, lack of originality in the outputs, and lack of copyright protection. We discuss how these concerns apply to design education and provide recommendations to educators about how AI should be addressed in the design classroom. We recommend that educators introduce AI as one tool among many in the designer's toolkit and encourage it to be used as a process tool rather than for generating final design deliverables. We also provide guidance for how educators might engage students in discussions about AI to enhance their learning.

## KEYWORDS

Education, Generative Artificial Intelligence, Industrial Design, Product Design, Text-to-Image.

DOI: 10.9781/ijimai.2024.02.006

## I. INTRODUCTION

**P**RODUCT design, or industrial design, is the design of objects for mass-manufacturing. Product designers are responsible for the design of products in a wide range of industries including housewares, sporting equipment, medical devices, consumer electronics, and more. Recent articles have discussed applications of machine learning, big data, and artificial intelligence (AI) to product design [1], [2]. However, little has been written in academic literature about the application of AI text-to-image or text-to-3D generators to the industrial design process, though the discussion is well underway in the faster-moving world of social media. Titles of YouTube videos published in the past year illustrate the growing relevance of AI to the product design discipline. For example: *AI Designed this Product: These Tools are the Future of Design* [3], *Using AI in Your Design Process (MidJourney, Stable Diffusion, Vizcom) – AI For Industrial Design* [4], *How to Design with AI #ai #midjourney #vizcom #chatgpt* [5], *Hyper-Real Prototyping for Speed and Control – AI for Industrial Design* [6], *A.I. Product Design – THIS Will Change Everything!* [7], *A.I. vs Pro Car Designer! Is There Still a Future for Us?* [8]. Videos like these discuss AI applications that generate digital images from textual descriptions, reference images, or reference sketches, as can be done in popular software programs including DALL-E, Stable Diffusion, Midjourney, Adobe Firefly, and more. The videos illustrate that AI is being presented to viewers, many of whom

are design students, as a “must-have skill” for future designers. Thus, educators need to be aware of the capabilities of these AI tools and the accompanying pitfalls and advantages for design education.

As we will illustrate with examples in the Section III, most industrial designers are utilizing generative AI in the front-end of the design process for inspiration, to create mood boards, and for ideation, in other words, to come up with design concept ideas. Some are also using generative AI for refinement or to broaden their ideas. In a few cases, designers have begun to output images generated by AI as underlays for manual 3D modeling, or to create 3D displacement maps which are used to directly texture 3D models. While text-to-3D software applications are not yet widely available, some websites claim to provide these services and student designers may be misled to believe that they are paying for an automated software application to create a 3D model from their sketches, when there is really a human modeler working behind the scenes [9].

The possibility of AI-facilitated plagiarism is a common concern for educators. However, the onset of generative AI introduces issues besides plagiarism, as the training data for the models utilizes work from photographers and visual artists who did not give consent for its use and were not compensated. Furthermore, as with other AI models, generative image AI often exhibits harmful biases. There are also legal and intellectual property-related issues to contend with. The purpose of this paper is to discuss these issues and provide general recommendations for educators regarding generative AI in design education. The primary domains which are publishing about the application of generative AI are medicine and computer vision [10], and very little has been written on the application of generative AI

\* Corresponding author.

E-mail address: kristibartlett@uky.edu

in design education. Therefore, this paper addresses the audience of design educators, who must determine how best to address generative AI in the classroom as they prepare their students for the workforce.

Because the AI landscape is fast-moving, and because product design professionals often do not publish their work in academic journals, this paper draws material from multiple areas, including conversations with practicing designers and literature from the fields of computer science and design education. First, we describe related literature. After that, we provide examples of ways that generative AI is currently being used by professionals in the product design field. Then, we delve into the ethical and legal issues through a discussion of recent academic literature and news articles. We conclude with recommendations for design educators. While the examples we are focusing on in this paper are from the context of industrial or product design, our discussion should also be relevant to other areas of art and design education.

## II. LITERATURE REVIEW

Many authors have argued that instead of replacing human designers, AI will become a powerful partner for human designers and enhance their capabilities. For example, Verganti and colleagues claimed that artificial intelligence has the capacity to reinforce the fundamental principles of design thinking, rather than displacing them [11]. Seidel et al. said that the emergence of autonomous design tools indicates that the role of human designers is changing [12], and Koch advocated the belief that systems leveraging AI can become collaborative partners in the design process [13].

Human-AI collaboration has been investigated in various stages of the design process, including early ideation and concept evaluation [14], later-stage ideation [15], management of the design team [16], aiding teams in design problem-solving [17], and aiding teams in the design of complex systems [18]. The addition of AI in the design process is not always found to be helpful. In one case, AI enabled broader and more efficient exploration of potential solutions [18]. In another case, however, AI was seen to hinder the performance of high-performing teams, though it did help low performing teams [17].

While the aforementioned studies all explored the application of AI to the design process, AI-human interaction in the engineering design process remains an understudied area [17]. This paper focuses specifically on image-generative AI applied to the design process. Generative AI is defined as the “production of previously unseen synthetic content, in any form and to support any task, through generative modeling,” where generative modeling means “modeling the joint distribution of inputs and outputs” [10]. Image-generative AI models are trained on large datasets of images paired with textual descriptions and work through a process called diffusion. Diffusion models add noise to data (image data, in this case) in a series of steps. Then, the process is run in reverse, and each step gradually denoises the image, leaving behind what the model predicts will be an image of the user’s prompted input [19].

While few authors discuss the application of image-generative AI to industrial design or product design, image-generative AI has been investigated in other related fields such as fashion design [20]. Researchers found that the majority of their generated fashion design images were thought to be created by human designers rather than being computer-generated [20]. Another team investigated image-generative AI which uses sketch-based input in the context of architectural design [21]. They commonly encountered a problem of the AI generating images that would be impossible to construct [21].

Rather than generating images, generative AI has been explored in the context of mechanical design to generate 3D geometry. A case study examined the use of generative design in a computer-aided design (CAD) software to perform structural optimization [22]. This

process involves inputting a set of design requirements in the form of numerical constraints relating to materials and manufacturing, as well as defining some basic geometric constraints in the CAD software [22]. Generative AI is also being applied to larger-scale structural design problems, such as building structures [23].

Cai et al. introduced a generative AI tool which creates inspiration mood boards of generated images based on a text prompt [24]. They had a group of participants with experience in art or design compare the outputs to results of an image search on Pinterest. Participants found the generative AI tools to be more useful, inspirational, and enjoyable than the traditional image search. Having a larger diversity of images generated was only slightly favored by the participants, and in the search condition, the lower diversity of images was preferred [24]. This study did not demonstrate or investigate how effectively or ethically participants might then use those inspiration images, but these are important aspects to consider. While designers perceived the AI-generated images to be more useful, what does this really mean? Might AI-generated inspiration images limit someone’s thinking, or lead them to unintentionally plagiarize?

While clear-cut rules about plagiarism and citing sources exist in nearly every university regarding written work, the concept of plagiarism for visual design work is already quite murky. In the postmodern design context, there is no consensus of where to draw the lines between borrowing, referencing, and plagiarizing [25]. Writing in the context of the year 2011, Economou described a “remix” realm in which design students are operating [25]. How much truer is this today, when the “remix” realm has given rise to what is essentially an automated remix machine in image generative AI? Writing in 1994, Saffo asked, “will the act of creativity be reduced to assembling old ideas like so much digital clip art, as the once-sustaining web of tradition becomes a suffocating blanket of electronic recall?” [26]. These examples from earlier writings demonstrate that concerns about lack of originality in design work were around long before the introduction of generative AI, and generative AI is just the latest iteration of technology which may facilitate design plagiarism. Educators and employers alike have concerns about plagiarism, both for the integrity of learning and to protect businesses from a legal standpoint.

This review of the literature indicates that many researchers see the value in applying generative AI to design. While researchers are increasingly investigating applications of AI to the design process, there is a need for more work that focuses specifically on design education. Other reviews have focused on classifying and categorizing generative AI systems and outlining the technical requirements, without discussing ethics [27]. Writing on the related topic of text-generative AI argues that following “responsible practices to uphold academic integrity and ensure ethical use” is crucial [28]. Thus, this paper discusses the ethics of image-generative AI applied to product design, as well as concerns about plagiarism and originality.

## III. USING AI TO GENERATE DESIGNS

### A. Example 1: Using AI to Generate Inspiration Images

Most of the popular image-generative AI software products allow the users to input text or other images to “prompt” the AI and tell it what kinds of images to generate. For example, designer Caterina Rizzoni of Kaleidoscope input the following text prompt into Midjourney V3: “Light fixture lighting a brilliant, elegant light and airy crystalline patterns of light dancing photorealistic detailed plants greenery daytime bright modern beautiful balcony patio trees natural colors outdoors.” From this prompt, Midjourney generated multiple images, some of which are shown in Fig. 1.





Fig. 1. Design inspiration images generated by Midjourney V3.

In many cases, users may simply use these AI output images as they are, if their goal was to generate an image. However, in the case of product design, the end goal is to come up with a product idea. In many cases, the images generated by AI are not a perfect match with the design requirements and are instead used as inspiration material. Taking the generated images in Fig. 1 as inspiration material, designer Tom Gernetzke sketched various lamp concepts. These sketches are shown in Fig. 2. (The Kaleidoscope innovation team's process is described in further detail in [29]). The inspiration from the generated images is clear, but the human designer added other details such as structurally supportive bases and electrical cords which are critical to the feasibility of the final lighting design.

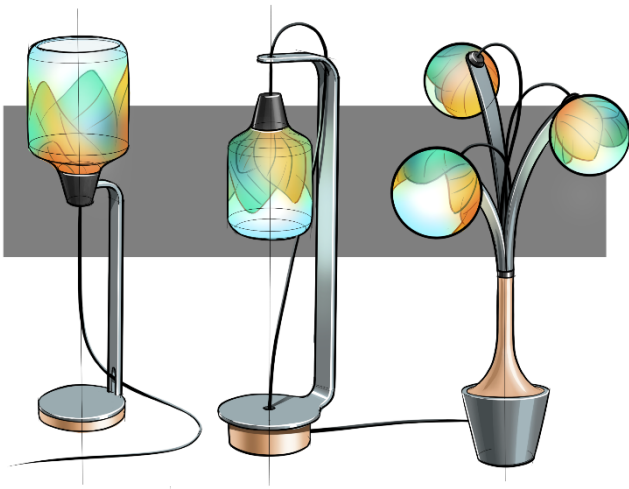


Fig. 2. Design concepts drawn by the design team, taking inspiration from the Midjourney images.

### B. Example 2: Using AI to Generate 2D Images of 3D Topology

As Example 1 illustrates, using an output image generated by AI to get a production-ready design is currently a highly manual process, and most often the images are used as a jumping-off point for inspiration but differ largely from the final design. However, designers are increasingly pushing the use of AI to complete more steps in the design process. In Example 2, the designer moves directly from an AI-generated image to a 3D model. Designer Kedar Benjamin used the AI image generation software Dall-E to create an image of a shoe, shown in Fig. 3.



Fig. 3. Shoe image generated with Dall-E software by designer Kedar Benjamin.

A 3D model of the shoe was then built collaboratively by two designers, Benjamin and Svet Abjo, using the software packages Blender and Maya. The designers drew topology over the original image in the 3D modeling software to create the overall shape of the shoe (Fig. 4).



Fig. 4. Topology drawn over the original image to create 3D model of the shoe.



Fig. 5. Final 3D-printed shoes produced by Zellerfeld. The shoe is printed in fused 3D lattice.

The designers interpreted the shoe's topology from the AI-generated image but also made changes based on what they thought

would be most appropriate for the final product, as well as adding their own designs for the parts of the shoe that are not visible in the generated 2D image. The final shoe, based on their manually built 3D model, is now in production from a made-to-order 3D printed shoe company, Zellerfeld (Fig. 5). While the shape of the final shoe is similar to the original image generated by AI, the creation of the final product required significant manual input from the designers in the creation of the detailed 3D model and the selection of a manufacturing method and material.

### C. Example 3: Using AI to Generate 3D Topology

In Example 3, designers added even more automation into the process of going from a generated image to a final product. Designer Marina Aperribay used Dalle-2 to create inspiration images for a shoe. Once she found the best prompt to create the desired outcome, she used this same prompt in Stable Diffusion 2.0, which has the capability to create a displacement map, or a texture, based on the image using the tool “depth2img.” The final shoe image used to generate the texture is shown in Fig. 6.



Fig. 6. Shoe image generated in Stable Diffusion 2.0 used to create displacement map [30].

Designer Kedar Benjamin created an automated workflow using the software Houdini that allowed the displacement map to be wrapped around a basic shoe model. This final model was then 3D printed. The 3D model and 3D printed shoe are shown in Fig. 7.



Fig. 7. (L) 3D model created by wrapping displacement map around a basic shoe model. (R) 3D print of final shoe [30].

Unlike the previous example which required manual 3D modeling, this workflow was both fast and suitable for people with limited 3D modeling skills. However, the designers anticipated this workflow would soon be obsolete with the arrival of “text-to-3d mesh” AI applications [30]. At the present time, software that uses AI to generate 3D models does not appear to be available to product designers, yet many designers hope that such programs will soon be developed. The recent breakthroughs in text-to-image AI are dependent on datasets which include billions of image-text pairs. The same approach cannot be taken for generating 3D models because large scale datasets of labeled 3D data do not exist, and neither do efficient architectures for

denoising 3D data [31]. However, researchers are working on other approaches that can generate 3D models from various combinations of 2D images, text prompts, and 3D priors [31], [32], [33], [34], [35], [36], [37].

## IV. ETHICAL AND LEGAL CONCERNS

While generative AI is a powerful tool that can create impressive images, the new field is fraught with ethical and legal concerns that designers cannot ignore. These issues are discussed in this section.

### A. Biased Outputs

One problem with image-generative AI is that the images it creates often reproduce biases, particularly when depicting humans. For example, a research group studying DALL-E and Stable Diffusion found that the models “learn specific gender/skin tone biases from web image-text pairs” [38]. The creators of DALL-E 2 were aware of the biases in their output images, stating that the images produced overrepresent people who are “White-passing.” Their model also overrepresents people who appear female for female-stereotyped jobs, such as a flight attendant, and overrepresents people who appear male for male-stereotyped jobs, such as a builder [39]. The DALL-E 2 team also found that their initial approach to filtering sexual content reduced the overall quantity of generated images of women, including images that did not contain sexual content [39]. Ultimately, these biases probably are not the largest concern for design students, since they are designing objects rather than people. Regardless, students should be careful to avoid representing people in biased ways in their imagery, and students need to be aware that AI tools often reproduce harmful societal biases.

Beyond images of people, the DALL-E 2 model also overrepresented “Western concepts” [39]. This certainly has implications for product design, as styles of design differ by culture and region. If AI models overrepresent Western styles of buildings, products, fashion, etc., then uncritical use of these AI tools could further perpetuate a Eurocentric bias in design. Students should be informed about cultural variations in design styles and be trained to recognize how AI may not present a very well-rounded sampling of styles from around the world. Design students are being trained to create the most appropriate design for the brief, and if AI generates a very narrow set of inspiration material, then the students need to recognize this fact and take their own steps to broaden their inspiration sources. (We note that although DALL-E was the primary example used in this section, the issue of bias is not limited to any single AI application.)

### B. No Guarantee of Originality

Another serious issue with image-generative AI is the possibility that the output is not unique. The output that one user gets from a generative AI software may be the same as what other users get, or it may be very similar to the images used in the training data. Researchers ran an experiment on Stable Diffusion and found while most generated images did not contain copied content, “a non-trivial amount of copying does occur” [40]. They focused on object-level similarity because it could potentially be the subject of future intellectual property disputes. An example from their research is shown in Fig. 6. The left image was generated by Stable Diffusion, the right image is a nearly equivalent shoe image found in the LAION-Aesthetics v2 6+ dataset.

The DALL-E team also stated that lack of unique outputs is a concern for their software, though they focused on the possibility of the software generating the same output for multiple users, rather than generating something that is very similar to the training data. They said, “due to the nature of machine learning, output may not be





Fig. 6. (L) Image of athletic shoe generated by Stable Diffusion, (R) image of similar athletic shoe found in LAION-Aesthetics v2 6+ dataset.

unique across users and the Services may generate the same or similar output for OpenAI or a third party” [41]. Both the issues of copying the original input images or giving the same output to multiple users raise the concern that anyone using image-generating AI cannot be sure that they have generated something unique. The lack of assurance that a generated design is unique poses a problem for designers whose primary goal is to create a novel design.

Designer Benjamin from Example 2 shared that his team took special care when creating the 3D model to avoid a “swoosh” shape that would be reminiscent of the brand Nike. Through his experiments with text-to-image AI, Benjamin observed that, “text to image tends to generate swooshes [reminiscent of Nike] or three stripes [reminiscent of Adidas] on a lot of shoes, and even when it doesn’t generate swooshes it sometimes makes design elements which resemble them. We have to be very careful about this.”

Most of the terms of service for image generative software applications that we reviewed for this paper put the responsibility of ensuring that there are no intellectual property violations onto the users (for example, [41], [42]). This means that even though the AI companies trained their models on images containing other companies’ intellectual property (IP), it would be the user’s responsibility if the software generated an output that was too similar to the IP of those companies and the user decided to use this output as their own “original” design. Thus, designers who are using generative AI have to be knowledgeable about the brand language and IP of other brands and take care that they are not generating designs that infringe on that intellectual property.

### C. Theft From Artists

Data is a fundamental building block of AI and machine learning models [43]. Many argue that the image generation programs like Midjourney are damaging to artists and photographers in that the training data contains millions of artworks and images without the creators’ consent [44]. Stability AI, the company behind Stable Diffusion, is being sued by Getty Images, who argues that more than 12 million of Getty Images’ stock photos were used to train Stable Diffusion’s algorithm without permission or compensation [45].

Midjourney admittedly did not seek consent from living artists or work still under copyright because, according to their CEO, “there isn’t really a way to get a hundred million images and know where they’re coming from.” [46] Midjourney’s training dataset was built from “a big scrape of the internet” and they train across multiple published open data sets [46]. Artists cannot opt out of being named in prompts and none have had their work taken out of the training dataset [46]. This problem is not isolated to Midjourney, as the CEO further stated, “our training data is pretty much from the same place as everybody else’s — which is pretty much the internet. Pretty much every big AI model just pulls off all the data it can, all the text it can, all the images it can. Scientifically speaking, we’re at an early point in the space, where everyone grabs everything they can, they dump it in a huge file, and

they kind of set it on fire to train some huge thing, and no one really knows yet what data in the pile actually matters” [47]. Stable Diffusion was trained on the 2b English language label subset of LAION 5b, “a general crawl of the internet created by the German charity LAION” [48]. There was also no opportunity for artists to opt-out of having their work included in the LAION 5b model data [48].

One example of a different approach by an AI company is that of Adobe, who has trained their initial Firefly model on “Adobe Stock images, openly licensed content, and public domain content where copyright has expired” [49]. Many designers we have spoken with expressed excitement about using Adobe’s Firefly software (currently in Beta), or other future software that takes a similar approach to not training on artists’ work without permission, because many in the creative community have ethical concerns about creative work being used to train AI without permission. Adobe is also exploring ways for future creators to be able to train the model with their own assets so that they can generate content in their own unique design style or brand language without using other creators’ content as source material [49]. This and other similar future solutions would also open future possibilities for AI as a design tool that could remediate some of the ethical concerns discussed in this section.

### D. Other Hidden Labor: Annotators

While the artists and photographers whose work is fed into AI models represent one invisible labor group in the AI ecosystem, they are not the only such group. In discussions of AI taking away human jobs, we often overlook the fact that AI creates jobs as well, although as the case of the annotators illustrates, many of the jobs created are not high-paying, not highly skilled, and not necessarily desirable. Annotators, or people who label, caption, and characterize text, images, or other data to create training data for AI models, are another group whose human labor is often unacknowledged in discussions of AI.

Many AI companies outsource the job of annotation to overseas companies. One author interviewed Kenyan annotators who were making somewhere between \$1 and \$3 per hour. The work was not consistent and came in waves, and they could not always count on having tasks [50]. A study of annotators working in India found that the work practices served the interests of the companies and requesters rather than the workers. Many of the annotators had entered the workforce under the guise that this was an entry point to a career as an AI/ML engineer when this was not the reality for most workers. The work was tedious and repetitive, the workers sometimes had to work overtime hours which were not compensated, and the work was project dependent, if a project ended, so did the work [43].

In the world of product design, many designers are on the lookout for services which will save them time in their workflow, and building 3D models from images is a time-consuming task for many designers. Kaedim is a 2D image to 3D model service which many believe is misleading users by selling their technology as AI, but instead using human workers behind the scenes build the 3D models in real-time. Users thought that the way the 3D models were simplified from the images looked like something that would be difficult to train an AI model to do. Furthermore, Kaedim had previously posted a job advertisement looking for workers who could “produce low quality 3D assets from 2D images 15 minutes after they are requested” [51]. In response to the criticism of falsely advertising their services as AI, their CEO said, “We have a product that’s starting to produce some exciting results — but it’s far from perfect” [9]. She said that although images pass through their AI algorithm for reconstruction as 3D files, a quality control engineer takes a look at each output and improves it where necessary [9].

Humans manually working on processes that are advertised as AI brings in a serious concern when it comes to student work. While a



professor might be fine with a student using software to automatically generate a file output, they would probably not be okay with a student paying another person to create the file for them, which is what is happening in cases of falsely advertised AI. Students need to be informed about the existence of such misleading services and advised by their instructors about what kinds of software and services are acceptable to use for class assignments.

### *E. Ownership of AI Outputs*

The U.S. Copyright Office has determined that AI-generated images are not protectable under current copyright law because they are “not the product of human authorship.” They said that Midjourney users have very little control over the final images in comparison to a human artist or photographer [52]. They also said, “the fact that Midjourney’s specific output cannot be predicted by users makes Midjourney different for copyright purposes than other tools used by artists.” [53]. At present, Midjourney only allows users to input text or images. It is not clear how copyright possibilities will evolve for cases such as Vizcom, an AI application which allows manual drawing in the input box, or a case where an artist could train a model on their own work. But at present, it is safest for design students and designers to assume that any images they create using image-generative AI software like Midjourney will not be able to receive copyright protection.

Furthermore, using free versions of AI programs often yields images that are explicitly open source. This is the case for Stable Diffusion Online [48] and for the free version of Midjourney [42]. Midjourney’s terms of service state, “Midjourney is an open community which allows others to use and remix Your images and prompts whenever they are posted in a public setting. By default, your images are publically [sic] viewable and remixable. As described above, You grant Midjourney a license to allow this” [42]. Students must understand that AI-generated images are often not protectable as their own work, and depending on what software they are using, the outputs they generate may be considered open-source.

## **V. RECOMMENDATIONS FOR EDUCATORS**

### *A. Engage Students in Discussions About the Ethical and Legal Implications of Using Generative AI*

The use of AI is often marketed to design students as a “must-have skill” for them to stay up to date with the latest and greatest technologies. However, as the issues discussed in the previous section illustrate, AI is fraught with ethical and legal concerns that are not relevant to other design technologies like computer-aided design or rendering software. Design students must be aware of the ethical and legal issues. In-class discussions are one way that students could be engaged with these issues. Discussions could focus on the five issues we introduced in the previous section: bias, lack of originality/copying, theft from artists, hidden labor, and ownership of outputs. Portions of this paper could also be used as jumping off points for the class discussion.

Our recommendation would be to encourage students to choose AI applications that train on images that they own, rather than on scraps of the internet, and that allow artists to opt-out of having their work included in training data. We would also recommend that students not be permitted to use an AI-generated output as a final deliverable for an assignment, since what is generated may not be unique and may infringe on the IP of others. We recommend putting a clear policy in place that does not discourage the use of AI as a part of the design process. Our examples in section III illustrate that designers are using AI in creative ways to come up with unique design solutions. However, the final outputs of generative AI are not copyright protectable for the

reason that the user does not have enough control over the output. For this same reason, an AI-generated output is likely not going to be a perfect design solution anyway, as human input is most likely needed to make sure that the output best meets the requirements of the design brief. Thus, students should be encouraged to keep going and keep refining their design solutions as much as possible, and not rely solely on what they can produce using AI tools. While these policies are our recommendation, students should be engaged in a discussion to help create a class policy regarding the use of AI, and that the policy can adapt and change as the AI landscape changes.

### *B. Concerns About Plagiarism*

#### *1. Should It Be Considered Plagiarism for a Student to Turn in an AI-Generated Design as Their Final Deliverable for an Assignment?*

The general definition of plagiarism is presenting the work of someone else as if it were your own [54]. What constitutes plagiarism in creative design disciplines is far less clear-cut than in disciplines that ask for written solutions, where there are clear guidelines that can be taught to students for quoting, attributing, paraphrasing, and citing. In design, there are no such guidelines [25].

In courses where craft is the focus, for example, a sketching course or a CAD course, the students need to create their own sketches or CAD for the deliverables. Thus, using AI to create these deliverables and being dishonest about the origins of the work would certainly constitute plagiarism. However, in studio courses where the design outcome, rather than a specific design skill, is the main focus, the use of AI as part of the process should be acceptable, as in the examples shared in section III. Using AI to directly create final deliverables could still be problematic.

Based on the fact that generated designs are not necessarily original, as in they might have copied heavily from the training data and might be extremely similar to an output given to someone else, we do not think it is wise for students or designers to claim a generated design as their own original design at the present time. That being said, presenting an AI-generated solution alongside substantial background research that provides a robust justification for the novelty and suitability of the solution could be valuable. Perhaps future iterations of generative design software will offer features that can make a stronger guarantee of originality of the outputs.

The fact that the original outputs of image generative AI are not protectable by copyright is another argument against allowing students to turn in AI outputs as part of their final design deliverables. AI outputs may be presented in process books and certainly should be documented if they played a part in the student’s design process, but the final product should be crafted by the student. Take, for example, the sketches in Fig. 2 or the 3D prints in Fig. 5 and Fig. 7. These would be acceptable outputs of projects which used generative AI in the early stages of ideation, as the designers added their own creative hand in creating models and sketches of the final product.

#### *2. How Would an Instructor Know if a student Was Trying to Pass Off an AI-generated Design as Their Own Original Work?*

This question is not inherently different than asking how an instructor would know if the student was using a file they found on the internet and trying to pass it off as their own unique work. Thus, we will review the recommendations that are already in place for combatting plagiarism in classrooms of creative disciplines like industrial design.

Prior to the introduction of generative AI, it was already common practice for designers to reference inspiration images they find on the internet [25]. Eighty-five percent of design students reported that

their first step in beginning an assignment is to conduct a Google image search or create an inspiration board using Pinterest, and they continue to reference these things throughout the design process [55]. In fact, many design instructors even encourage their students to collect a broad range of visual samples to draw inspiration from in their design process [56]. The problem comes when the inspiration sources are too similar to the final design submission, and design students face growing difficulty in navigating the lines between plagiarism, appropriation, homage, inspiration, and referencing others' work [25].

Educators have proposed various solutions to combat plagiarism in design education. Pedagogical approaches that discourage plagiarism are preferred over detection approaches [57]. For example, project-based learning has a lower risk of plagiarism because the instructors closely supervise students' work and students keep a logbook of their individual contributions to team projects [58]. Coorey argues that training students to engage in their own design process is the most important method of discouraging plagiarism [55]. Studio projects naturally lend themselves to this as there are many milestones along the way where students perform the different steps to develop their projects [59]. Process work should be emphasized in the assessment practice in order to place focus on the designer's role in developing the final solutions [25]. A process book, in which students show their process of ideation and revisions which led them to the final design outcome, can serve as an assessment tool for the instructors [55].

Design programs should provide lectures on visual plagiarism and appropriation theory, studio practice should include visual referencing systems to provide students a method to indicate their source material which they referenced to build to their final design [25]. One approach called "Beyond Style" guided students through a process of how to be inspired by creative precedents without plagiarizing, with the idea that this would also help students to respect the creative works of others [57]. An alternative option is to train students to write a statement of novelty, which may serve as a useful exercise in the context of design education where students may want to protect their IP in the future with patents [60]. Ultimately, art and design programs need plagiarism policy documents relevant to their disciplines [25]. Design instructors today need to ensure that their plagiarism policies address the use of AI and what is and is not acceptable in their classroom.

### C. Ensuring That Students Build the Skills Needed to Be Successful in Industry

Can students expect to be allowed to use AI in their jobs upon graduation? We spoke with multiple design professionals about this question. Many designers who work in US-based consumer products companies were given restrictions by their legal departments about how they could use AI software at work. One company's training on AI said that using AI to make images can generate content that infringes on others' intellectual property rights, which could open the company up to lawsuits. They forbid inputting company data into AI software as prompts, as this could expose the company's own IP. Thus, they placed heavy restrictions on their design teams using AI.

A designer at another company was provided with a Pro license for the software Midjourney, however, the designers were only allowed to use Midjourney to generate images for storytelling or background material to explain the context or intentions behind a design and could not use Midjourney to generate actual design concepts. They were also forbidden to use any brand names in the text prompts. Another designer said that her team did not feel comfortable using generative AI for ethical reasons. They were specifically concerned about the ethical issue of AI using the work of artists without the artists' consent.

In contrast to the previous examples, a designer who works at a large tech company said that her company encourages the use of AI in their work since the company is in the business of creating AI

applications themselves. Another designer pointed out that larger companies like hers were working with tech companies to develop proprietary AI applications that would not expose them to legal and IP concerns.

From these examples, it is clear that design students today cannot count on entering the workforce and being encouraged or allowed to freely use AI as part of their design process, especially if they enter in an industrial design role in a large consumer goods company (individuals who end up working for tech or small design consultancies without legal teams will likely face different policies regarding AI). While students should know the capabilities of generative AI, they should also be well-versed in the legal and ethical issues surrounding AI so that they will be able to make informed decisions that do not violate the guidance from their employers. They should also be fully capable of creating excellent designs without the aid of generative AI in the event that they work for an employer who does not permit its use. Students could end up in a situation where they use AI freely during their education, become reliant on it during their design process, and then graduate and are not allowed to use it in the workplace, which would not be ideal.

### D. Design Competitions

While some companies are hesitant about adopting AI, design competitions appear to be taking a different stance. The iF Design Award considers that many winners already involve "AI" as they are smart products such as fridges or smart phones. So, they did not plan to differentiate entries that involved AI. The Red Dot Award focuses on the end results, and if AI plays a part in leading to an award-worthy physical product, then that product would still be eligible to win the award [61]. Thus, students who want to enter their work into competitions probably do not need to be concerned that using AI in their design process would disqualify them. That said, the students should still be transparent in their process books and portfolios about how and where AI was leveraged. Of course, students and educators should also check the policies of any design competition that they plan to enter to see if the policies place any restrictions on the use of AI.

### E. Summary of Recommendations for Educators

Section V has consisted of an in-depth discussion of our recommendations for educators who are faced with the choice of introducing generative-AI in design classrooms. Table I provides a summary of these recommendations and the reasoning behind them.

## VI. CONCLUSION

Image-generative AI is a promising new tool for product designers to use in their design process. In this paper, we presented three examples of projects which used AI-generated images as an inspiration source for design sketches, as an underlay for a 3D modeled design, and to automatically generate a texture. Image generative AI is still a new technology, and future iterations will be even more advanced. Product designers are increasingly looking for tools to help them generate 3D designs more quickly and efficiently and with increased control over the final outcome.

To help ensure that students are graduating with the most up-to-date software skills, educators would do well to introduce generative AI as one tool among the many tools in which they train their students. However, AI differs in many ways from traditional technologies, and should not be introduced without a clear discussion of the ethical and legal implications, and clear guidelines about the instructor's policies for how AI can be used in projects and final deliverables. Even if an instructor does not plan to introduce AI, these guidelines should be provided as part of the plagiarism policy given in a syllabus.

TABLE I. SUMMARY OF RECOMMENDATIONS FOR EDUCATORS

Recommendation	Reason
Introduce AI as a tool to be used in conjunction with other design tools.	Students should be familiar with the capabilities of AI but should not get the idea that AI replaces fundamental skills of designers at this time.
Do not allow students to turn in raw AI-generated content as a final product.	At this time, raw outputs of image-generative AI are not as refined as what is needed for professional design work. Students who use AI tools should build on the outputs of AI and refine them manually using their own critical thinking.
Encourage use of AI tools that give a high degree of control, such as tools trained on one's own work or tools that use sketch-based inputs rather than text-based.	Tools which give the designer a higher degree of control are more likely to result in original outputs.
Require students to document use of AI in projects, process books, and portfolios.	Students should make a clear distinction between their own work and AI-generated or AI-assisted work in order to avoid plagiarism concerns.
Ensure that students can still complete required tasks without AI.	Some companies do not permit design teams to use AI, so students cannot count on being able to use AI in all future jobs.
Engage students in discussions of ethical issues surrounding AI (eg. Bias in outputs, theft from artists, hidden labor, inadvertent copying, copyright and ownership).	Students should be aware of the many ethical and legal issues surrounding image generative AI to help them make informed decisions of how they might or might not want to use AI in their work.

We do not recommend students be allowed to turn in fully AI-generated files as their final project artifacts. It is unlikely that this will be permitted in their future jobs due to copyright and IP concerns. Furthermore, allowing students to use AI for final artifacts could hinder their skills development, as generative-AI does not currently allow for the same level of control over the final design outcome that other tools do. However, using AI alongside the traditional design tools could be an asset to helping students work more efficiently and could lead to new creative insights.

Educators should not naively cling to traditional techniques and methods but must remain open to the possibility that certain hand skills in design may decrease in importance in the future. No doubt educators in the past were afraid to introduce CAD, 3D rendering, and digital sketching for fear that students would lose hand sculpting and hand rendering skills. Both industry and education evolve as new technologies change designers' workflow and clients' expectations.

Saffo (1994) argued that originality was increasingly rare, and originality would eventually cease to be the true litmus test of creativity. Instead, value would be placed on passion, surprise, and insight [26]. As illustrated by the examples in section III, the designer's creativity is still critical to transforming the outputs of generative AI into a viable final design solution. At present, generative AI is not going to output a manufacturable final product. The designer must be the one to curate the best solution, taking into consideration the user needs, market appropriateness, and IP space. The designer can certainly leverage generative AI to help get to the final viable outcome, but designer's human skills are still of critical importance. Trend research, user research, understanding of branding and brand identity, and manufacturability knowledge may become increasingly valuable skills in the age of generative AI.

Although this paper focuses on product design education, and the examples that we presented are all from the product design field, we believe that our recommendations for educators apply to other design disciplines which have a visual emphasis, such as graphic design, architecture, engineering design, fashion design, interior design, and fine art. The fact that we only spoke with individuals from the discipline of product design is a limitation of this paper. However, our review of ethical and legal concerns was not discipline-specific, and we drew from a range of sources to write this section.

In conclusion, educators must take notice of image-generative AI, because their students are certainly aware of it and will be experimenting with the technology regardless of whether the educators address it or not. At present, the raw outputs of AI are likely

not suitable for use as final deliverables in design education due to their lack of copyright protection and the possibility of copying and IP infringement. However, future AI tools are likely to offer more control over the final solution, and a stronger guarantee of originality. Future tools should also address the ethical issues surrounding bias and theft from artists. Generative AI offers exciting possibilities when used as part of a comprehensive design process, and engaging students in discussions about AI in design can help them think critically about their role as designers in the face of technological change.

#### ACKNOWLEDGMENT

The authors would like to thank Kedar Benjamin and Caterina Rizzoni for sharing insights into their design processes and images of their teams' work. The authors would also like to thank Linda Bui, Lindsay Malatesta, and Lea Stewart for sharing their insights on the use of AI in design workplaces.

#### REFERENCES

- [1] P. Fournier-Viger, M. S. Nawaz, W. Song, and W. Gan, "Machine Learning for Intelligent Industrial Design," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, vol. 1525, M. Kamp, I. Koprincka, A. Bibal, T. Bouadi, B. Frénay, L. Galárraga, J. Oramas, L. Adilova, G. Graça, et al., Eds., in Communications in Computer and Information Science, vol. 1525, Cham: Springer International Publishing, 2021, pp. 158–172. doi: 10.1007/978-3-030-93733-1\_11.
- [2] T. Guo, R. Eckert, and M. Li, "Application of Big Data and Artificial Intelligence Technology in Industrial Design," vol. 5, no. 1, 2020.
- [3] *AI Designed this Product: These Tools are the Future of Design*, (Aug. 12, 2021). Accessed: Jun. 22, 2023. [Online Video]. Available: [https://www.youtube.com/watch?v=sy\\_lq2yq9U](https://www.youtube.com/watch?v=sy_lq2yq9U)
- [4] *Using AI in Your Design Process (MidJourney, Stable Diffusion, Vizcom) – AI For Industrial Design*, (Feb. 12, 2023). Accessed: Jun. 22, 2023. [Online Video]. Available: <https://www.youtube.com/watch?v=2y10zBnH2bk>
- [5] *How to Design with AI #ai #midjourney #vizcom #chatgpt*, (Feb. 08, 2023). Accessed: Jun. 22, 2023. [Online Video]. Available: <https://www.youtube.com/watch?v=zzYdWkgGgUM>
- [6] *Hyper-Real Prototyping for Speed and Control – AI for Industrial Design*, (Apr. 10, 2023). Accessed: Jun. 22, 2023. [Online Video]. Available: [https://www.youtube.com/watch?v=0Xo\\_LjsAc4](https://www.youtube.com/watch?v=0Xo_LjsAc4)
- [7] *A.I. Product Design – THIS Will Change Everything!*, (Mar. 08, 2023). Accessed: Jun. 22, 2023. [Online Video]. Available: <https://www.youtube.com/watch?v=bAQv1Sbx1-U>
- [8] *A.I. vs Pro Car Designer! Is There Still a Future for Us?*, (Sep. 15, 2022). Accessed: Jun. 22, 2023. [Online Video]. Available: <https://www.youtube.com/watch?v=bAQv1Sbx1-U>



- com/watch?v=v-ZkVsjmXUk
- [9] K. Psoma, "Response from Kaedim re our AI," Medium. Accessed: Jul. 17, 2023. [Online]. Available: <https://medium.com/kaedim/response-from-kaedim-re-our-ai-931d3ef39c33>
- [10] F. García-Peñalvo and A. Vázquez-Ingelmo, "What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, p. 7, 2023, doi: 10.9781/ijimai.2023.07.006.
- [11] R. Verganti, L. Vendraminelli, and M. Iansiti, "Innovation and Design in the Age of Artificial Intelligence," *Journal of Product Innovation Management*, vol. 37, no. 3, pp. 212–227, May 2020, doi: 10.1111/jpim.12523.
- [12] S. Seidel, N. Berente, A. Lindberg, K. Lyytinen, and J. V. Nickerson, "Autonomous tools and design: a triple-loop approach to human-machine learning," *Communications of the ACM*, vol. 62, no. 1, pp. 50–57, Dec. 2018, doi: 10.1145/3210753.
- [13] J. Koch, "Design implications for Designing with a Collaborative AI," presented at the The AAAI 2017 Spring Symposium on Designing the User Experience of Machine Learning Systems, 2017, pp. 415–418.
- [14] B. Camburn, Y. He, S. Raviselvam, J. Luo, and K. Wood, "Machine Learning-Based Design Concept Evaluation," *Journal of Mechanical Design*, vol. 142, no. 3, p. 031113, Mar. 2020, doi: 10.1115/1.4045126.
- [15] C. Yuan and M. Moghaddam, "Attribute-Aware Generative Design With Generative Adversarial Networks," *IEEE Access*, vol. 8, pp. 190710–190721, 2020, doi: 10.1109/ACCESS.2020.3032280.
- [16] J. T. Gyory *et al.*, "Human Versus Artificial Intelligence: A Data-Driven Approach to Real-Time Process Management During Complex Engineering Design," *Journal of Mechanical Design*, vol. 144, no. 2, p. 021405, Feb. 2022, doi: 10.1115/1.4052488.
- [17] G. Zhang, A. Raina, J. Cagan, and C. McComb, "A cautionary tale about the impact of AI on human design teams," *Design Studies*, vol. 72, p. 100990, Jan. 2021, doi: 10.1016/j.destud.2021.100990.
- [18] B. Song, N. F. Soria Zurita, H. Nolte, H. Singh, J. Cagan, and C. McComb, "When Faced With Increasing Complexity: The Effectiveness of Artificial Intelligence Assistance for Drone Design," *Journal of Mechanical Design*, vol. 144, no. 2, p. 021701, Feb. 2022, doi: 10.1115/1.4051871.
- [19] L. Yang *et al.*, "Diffusion Models: A Comprehensive Survey of Methods and Applications," arXiv, Mar. 23, 2023. Accessed: Jun. 24, 2023. [Online]. Available: <http://arxiv.org/abs/2209.00796>
- [20] O. Sbai, M. Elhoseiny, A. Bordes, Y. LeCun, and C. Couprie, "DesIGN: Design Inspiration from Generative Networks," in *Computer Vision – ECCV 2018 Workshops*, vol. 11131, L. Leal-Taixé and S. Roth, Eds., in Lecture Notes in Computer Science, vol. 11131. Cham: Springer International Publishing, 2019, pp. 37–44. doi: 10.1007/978-3-030-11015-4\_5.
- [21] C. Zhang, W. Wang, P. Pangaro, N. Martelaro, and D. Byrne, "Generative Image AI Using Design Sketches as input: Opportunities and Challenges," in *Creativity and Cognition*, Virtual Event USA: ACM, Jun. 2023, pp. 254–261. doi: 10.1145/3591196.3596820.
- [22] F. Buonamici, M. Carfagni, R. Furferi, Y. Volpe, and L. Governi, "Generative Design: An Explorative Study," *Computer-Aided Design and Applications*, vol. 18, no. 1, pp. 144–155, May 2020, doi: 10.14733/cadaps.2021.144-155.
- [23] W. Liao, X. Lu, Y. Fei, Y. Gu, and Y. Huang, "Generative AI design for building structures," *Automation in Construction*, vol. 157, p. 105187, Jan. 2024, doi: 10.1016/j.autcon.2023.105187.
- [24] A. Cai *et al.*, "DesignAID: Using Generative AI and Semantic Diversity for Design Inspiration," in *Proceedings of The ACM Collective Intelligence Conference*, Delft Netherlands: ACM, Nov. 2023, pp. 1–11. doi: 10.1145/3582269.3615596.
- [25] I. Economou, "The Problem with Plagiarism," in *Conference Proceedings of the Sixth International DFSA Conference*, Design Education Forum of South Africa, 2011, pp. 79–86.
- [26] P. Saffo, "The Place of Originality in the Information Age," *Journal of Graphic Design*, vol. 12, no. 1, 1994.
- [27] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, "The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges," *Future Internet*, vol. 15, no. 8, p. 260, Jul. 2023, doi: 10.3390/fi15080260.
- [28] A. M. Jarrah, Y. Wardat, and P. Fidalgo, "Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say?," *Online Journal of Communication and Media Technologies*, vol. 13, no. 4, p. e202346, Oct. 2023, doi: 10.30935/ojcm/13572.
- [29] T. Siebel, "An Experiment in Collaboration," *Innovation: Quarterly of the Industrial Designers Society of America*, vol. 42, no. 1, pp. 34–37, 2023.
- [30] footwearology\_lab, "We are super proud to present the world's first AI-generated, 3D printed, wearable sneaker!" Instagram. Accessed: Jun. 22, 2023. [Online]. Available: <https://www.instagram.com/p/Cl138IXo3mq/?igshid=MzRlODBiNWFlZA%3D%3D>
- [31] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D Diffusion," arXiv, Sep. 29, 2022. Accessed: Jun. 12, 2023. [Online]. Available: <http://arxiv.org/abs/2209.14988>
- [32] J. Xu *et al.*, "Dream3D: Zero-Shot Text-to-3D Synthesis Using 3D Shape Prior and Text-to-Image Diffusion Models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20908–20918.
- [33] Q. Shen, X. Yang, and X. Wang, "Anything-3D: Towards Single-view Anything Reconstruction in the Wild," arXiv, Apr. 19, 2023. Accessed: Jun. 12, 2023. [Online]. Available: <http://arxiv.org/abs/2304.10261>
- [34] R. Chen, Y. Chen, N. Jiao, and K. Jia, "Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation," arXiv, Mar. 28, 2023. Accessed: Jun. 12, 2023. [Online]. Available: <http://arxiv.org/abs/2303.13873>
- [35] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, and D. Li, "GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images," presented at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, 2022.
- [36] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, "Zero-Shot Text-Guided Object Generation with Dream Fields," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 857–866. doi: 10.1109/CVPR52688.2022.00094.
- [37] C.-H. Lin *et al.*, "Magic3D: High-Resolution Text-to-3D Content Creation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 300–309.
- [38] J. Cho, A. Zala, and M. Bansal, "DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Models," arXiv, Nov. 14, 2022. Accessed: Jun. 03, 2023. [Online]. Available: <http://arxiv.org/abs/2202.04053>
- [39] OpenAI, "DALL-E 2 Preview - Risks and Limitations," OpenAI Github. Accessed: Jun. 05, 2023. [Online]. Available: <https://github.com/openai/dalle-2-preview/blob/main/system-card.md>
- [40] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6048–6058.
- [41] OpenAI, "Terms of Use," OpenAI. Accessed: Jun. 05, 2023. [Online]. Available: <https://openai.com/policies/terms-of-use>
- [42] Midjourney, "Terms of Service," Midjourney. Accessed: Jun. 05, 2023. [Online]. Available: <https://docs.midjourney.com/docs/terms-of-service>
- [43] D. Wang, S. Prabhat, and N. Sambasivan, "Whose AI Dream? In search of the aspiration in data annotation.," in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–16. doi: 10.1145/3491102.3502121.
- [44] B. Nicholls, "Midjourney founder basically admits to copyright breaching and artists are angry," Digital Camera World. Accessed: Jun. 05, 2023. [Online]. Available: <https://www.digitalcameraworld.com/news/midjourney-founder-basically-admits-to-copyright-breaching-and-artists-are-angry>
- [45] B. Brittain, "Getty Images lawsuit says Stability AI misused photos to train AI," Reuters. Accessed: Jun. 05, 2023. [Online]. Available: <https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/>
- [46] R. Salkowitz, "Midjourney Founder David Holz On The Impact Of AI On Art, Imagination And The Creative Economy," Forbes. Accessed: Jun. 05, 2023. [Online]. Available: <https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy/?sh=2c41eb42d2b8>
- [47] J. Vincent, "'An engine for the imagination': the rise of AI image generators - An interview with Midjourney founder David Holz," The Verge. Accessed: Apr. 12, 2023. [Online]. Available: <https://www.>

theverge.com/2022/8/2/23287173/ai-image-generation-art-midjourney-multiverse-interview-david-holz

- [48] Stable Diffusion, "Frequently asked questions," Stable Diffusion Online. Accessed: Jun. 05, 2023. [Online]. Available: <https://stablediffusionweb.com/#faq>
- [49] Adobe, "Adobe Firefly," Adobe. Accessed: Jun. 22, 2023. [Online]. Available: <https://www.adobe.com/sensei/generative-ai/firefly.html#faqs>
- [50] J. Dzieza, "AI Is a Lot of Work," The Verge. Accessed: Jun. 23, 2023. [Online]. Available: <https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots>
- [51] Partnerkin, "'AI Startups,' where Indians Freelancers Work Instead of AI, and a Selection of 15 AI Tools that You Can Use to Make Money Online," Partnerkin. Accessed: Jun. 22, 2023. [Online]. Available: <https://partnerkin.com/en/blog/publications/ai-startups>
- [52] J. Lang, "U.S. Copyright Office Says AI-Generated Images Do Not Qualify For Copyright Protection," Cartoon Brew. Accessed: Jun. 05, 2023. [Online]. Available: <https://www.cartoonbrew.com/law/midjourney-ai-images-us-copyright-office-226437.html>
- [53] B. Brittain, "AI-created images lose U.S. copyrights in test for new technology," Reuters. Accessed: Jun. 05, 2023. [Online]. Available: <https://www.reuters.com/legal/ai-created-images-lose-us-copyrights-test-new-technology-2023-02-22/>
- [54] B. Rosamond, "Plagiarism, Academic Norms and the Governance of the Profession," *Politics*, vol. 22, no. 3, pp. 167–174, Sep. 2002, doi: 10.1111/1467-9256.00172.
- [55] J. Coorey, "Removing Plagiarism from the Design Process: Stimulating Creativity and Originality in the Design Classroom," *The International Journal of Design Education*, vol. 12, no. 1, pp. 11–19, 2018.
- [56] S. Izadpanah, "Evaluating the Role of Pinterest in Education and the Profession of Interior Architecture," *Idil*, vol. 10, no. 87, Nov. 2021, doi: 10.7816/idil-10-87-01.
- [57] M. Mostafa, "Inspiration versus Plagiarism: Academic Integrity in Architectural Education," *The International Journal of the Constructed Environment*, vol. 1, no. 3, pp. 85–100, 2011, doi: 10.18848/2154-8587/CGP/v01i03/37482.
- [58] A. Shekar, "Project-based Learning in Engineering Design Education: Sharing Best Practices," in *2014 ASEE Annual Conference & Exposition Proceedings*, Indianapolis, Indiana: ASEE Conferences, Jun. 2014, p. 24.1016.1-24.1016.18. doi: 10.18260/1-2--22949.
- [59] J. Carroll, "Assessment, learning and judgement in higher education," G. Joughin, Ed., Dordrecht: Springer Netherlands, 2008, pp. 115–131.
- [60] J. Everitt and A. M. Holmes, "Assessment of Novelty for Intellectual Property with Implications for Design Education: A Case Study," *The International Journal of Design Education*, vol. 10, no. 4, 2016.
- [61] S. Sheth, "How are Design Competitions and Awards Planning on Dealing with Ai-Based Submissions?," Yanko Design. Accessed: Jul. 17, 2023. [Online]. Available: [https://www.yankodesign.com/2023/06/29/how-are-design-competitions-and-awards-planning-on-dealing-with-ai-based-submissions/?fbclid=PAAabwNRK9KEg68snKy-db1LaUQ-9RORayGnDEmrXmTMB5scm7Yu7UagAd7OQ\\_aem\\_AQyHCdJF3\\_pbb6tnYc5jm879kQySSAgfWym5HWpDiaj47T2kIuPjCsEp29Kw2lbD0Q](https://www.yankodesign.com/2023/06/29/how-are-design-competitions-and-awards-planning-on-dealing-with-ai-based-submissions/?fbclid=PAAabwNRK9KEg68snKy-db1LaUQ-9RORayGnDEmrXmTMB5scm7Yu7UagAd7OQ_aem_AQyHCdJF3_pbb6tnYc5jm879kQySSAgfWym5HWpDiaj47T2kIuPjCsEp29Kw2lbD0Q)



Jorge D. Camba

Jorge D. Camba is an Associate Professor in the School of Engineering Technology at Purdue University in West Lafayette, IN. Prior to joining Purdue, he taught at the University of Houston and at Texas A&M University. Dr. Camba is the author of 100+ peer-reviewed publications and nine books. His research focuses on 3D model complexity and optimization, design automation and reusability, human-computer interaction, and Model-Based Engineering.



Kristin A. Bartlett

Kristin A. Bartlett is an Assistant Professor in the Department of Product Design at the University of Kentucky in Lexington, KY. She holds a PhD in Technology from Purdue University in West Lafayette, IN, USA, a Master of Science in Industrial Design from the University of Houston in Houston, TX, USA, and a Bachelor's in Civil & Environmental Engineering from Rice University in Houston, TX, USA. Kristin's research interests include design education, gender equity in STEM, and applied research in the area of medical device development.

# Can Generative AI Solve Geometry Problems? Strengths and Weaknesses of LLMs for Geometric Reasoning in Spanish

Verónica Parra<sup>1,3</sup>, Patricia Sureda<sup>1,3</sup>, Ana Corica<sup>1,3</sup>, Silvia Schiaffino<sup>2,3</sup>, Daniela Godoy<sup>2,3</sup> \*

<sup>1</sup> Universidad Nacional del Centro de la Provincia de Buenos Aires, Facultad de Ciencias Exactas, NIEM, Tandil, Buenos Aires (Argentina)

<sup>2</sup> Universidad Nacional del Centro de la Provincia de Buenos Aires, Facultad de Ciencias Exactas, ISISTAN, Tandil, Buenos Aires (Argentina)

<sup>3</sup> CONICET, Buenos Aires (Argentina)

Received 18 November 2023 | Accepted 15 February 2024 | Published 28 February 2024



## ABSTRACT

Generative Artificial Intelligence (AI) has emerged as a disruptive technology that is challenging traditional teaching and learning practices. Question-answering in natural language fosters the use of chatbots, such as ChatGPT, Bard and others, that generate text based on pre-trained Large Language Models (LLMs). The performance of these models in certain areas, like Math problem solving is receiving a crescent attention as it directly impacts on its potential use in educational settings. Most of these evaluations, however, concentrate on the construction and use of benchmarks comprising diverse Math problems in English. In this work, we discuss the capabilities of most used LLMs within the subfield of Geometry, in view of the relevance of this subject in high-school curricula and the difficulties exhibited by even most advanced multimodal LLMs to deal with geometric notions. This work focuses on Spanish, which is additionally a less resourced language. The answers of three major chatbots, based on different LLMs, were analyzed not only to determine their capacity to provide correct solutions, but also to categorize the errors found in the reasoning processes described. Understanding LLMs strengths and weaknesses in a field like Geometry can be a first step towards the design of more informed methodological proposals to include these technologies in classrooms as well as the development of more powerful automatic assistance tools based on generative AI.

## KEYWORDS

Chatbots, Generative AI, Geometry, LLMs, Math Problem-Solving.

DOI: 10.9781/ijimai.2024.02.009

## I. INTRODUCTION

THE emergence and fast adoption of natural-language chatbots, such as OpenAI ChatGPT<sup>1</sup>, or Google Bard<sup>2</sup>, leveraging Large Language Models (LLMs) to question-answering, is a phenomenon having a growing impact in several daily activities. Education is among the most heavily impacted areas by the irruption of these tools as the interaction between generative AI with both students and teachers allows to envision promising applications in pedagogical scenarios, but also unveils potential risks.

Mathematics is a valuable testbed for evaluating problem-solving capabilities of LLMs as it involves the ability to analyze and comprehend the problem stated, select viable heuristics from a potentially large set of strategies, and combine them into a chain-of-thought leading to a solution. Each of these high-level abilities poses complex challenges for AI-based technologies, in general, and generative AI models, in particular.

The incorporation of generative AI in educational settings requires a deep understanding of both the capabilities and limitations of LLMs to provide solutions to Math problems as well as step-by-step explanations at different levels. Novel AI-based techniques can be built upon this knowledge and exploit LLMs potential for the development of more powerful tools, including Math teaching assistants interacting with students during their learning process and potentially offering individualized instruction.

Studies oriented to evaluate the performance of LLMs on mathematical reasoning have been mostly concerned with the construction of appropriate benchmarks and the quantitative analysis of a given model results with respect to them [1]–[5]. Although their findings can provide an overall view of LLMs performance in the Math domain, there is still a lack of understanding of their strengths and weaknesses in general terms and in specific Math areas, such as Geometry.

<sup>1</sup> <https://chat.openai.com/>

<sup>2</sup> <https://bard.google.com/>

### \* Corresponding author.

E-mail addresses: vparra@niem.exa.unicen.edu.ar (V. Parra), psureda@niem.exa.unicen.edu.ar (P. Sureda), acorica@niem.exa.unicen.edu.ar (A. Corica), silvia.schiaffino@isistan.unicen.edu.ar (S. Schiaffino), daniela.godoy@isistan.unicen.edu.ar (D. Godoy).



Finding solutions for Geometry problems might result in a specially challenging task for generative AI based on multimodal LLMs as it not only involves the knowledge of fundamental concepts (theorems) and its correct application, but specially the use of spatial reasoning skills. At the same time, Geometry has a preeminent place in high-school curricula in many countries. Because of this, it becomes essential to better understand the potential and pitfalls of chatbots in solving Geometry problems as an essential step towards the construction of more powerful teaching assistance tools as well as pedagogical strategies integrating available general-purpose chatbots.

In addition, current studies are concentrated on English texts, while the performance of LLMs in less represented languages, such as Spanish, remains to be investigated. The quality of answers of models for different languages is directly related to the amount of training data available for each language, performing better for languages with larger representation like English and exhibiting an inferior performance for languages like Spanish.

This work presents a study tending to shed some light on the abilities of chatbots to provide accurate solutions to Geometry problems in Spanish. We carried out an analysis of the answers provided by three available chatbots, namely OpenAI ChatGPT, Microsoft Bing Chat (BingChat)<sup>3</sup>, and Google Bard, using a case study of Geometry high-school problem. The three major chatbots covered, leveraging versions of GPT-3.5 [6], GPT-4 [7] and PaLM-2 [4] models, were chosen because they are accessible and currently being used by students in everyday activities and schools. The problem analyzed corresponds to an Iberoamerican Math competition<sup>4</sup> oriented to high school students, and it is targeted to students under 13 years old. As a result of this study, we propose a categorization of errors made by chatbots in Geometry reasoning that can be used as input towards the construction of methodological proposals fostering the use of generative AI for learning and skill acquisition.

The structure of this document is as follows: section II discusses related works in the area, section III introduces the material and methods used in this study, section IV discusses the results obtained and, finally, section V presents the conclusions and devises promising avenues for further research.

## II. BACKGROUND & RELATED WORKS

In this section we first summarize some aspects regarding the use of LLMs in education (subsection A), then we discuss research on the performance of these models in Math problem-solving (subsection B) and finally we introduce some context and background concepts related to Geometry teaching (subsection C).

### A. LLMs in Education

Since the launching of ChatGPT by OpenAI in 2022, there has been an intensive discussion about the integration of generative AI in several fields, particularly in education [8],[9], as well as about the ethical aspects of using artificial intelligence (AI) systems in educational contexts [10], [11]. ChatGPT was trained on a large volume of text data, using the Generative Pre-trained Transformer (GPT) deep learning architecture. Immediately, the friendly, human-like responses in natural language conversations lead ChatGPT to be one of the technologies of fastest adoption.

The irruption of generative AI and the widespread adoption of ChatGPT opened the discussion on both challenges and concerns regarding its use in educational settings. On one side, there is a pressing need of harnessing the power of these tools for enhancing teaching

and learning practices. Among other benefits, LLMs can be used in the development of personalized learning tutors for students and being of assistance to teachers in the creation of educational resources (e.g. syllabus and class planning, course material and exercises) as well as the assessment of students capabilities (e.g. generating tests and evaluation scenarios), among many other applications. On the other side, LLMs potential uses rise concerns in relation to their accuracy and reliability as well as other threats such as misuses, plagiarism, the presence of biases and hallucinations and other ethical considerations. In [12] it had even found that risks also encompass the potential to limit critical thinking and creativity and impede a deep understanding of subject matter, and foster passivity.

General purpose chatbots, such as ChatGPT or Bard, are trained for dealing with question-answering in diverse domains as they are trained with large portions of the Web. However, recent studies have shown that chatbots perform differently in different subject areas including finance, coding, maths, and general public queries [13]. In [14], for example, it was found that ChatGPT performance varied across subject domains, ranging from outstanding (e.g., economics) and satisfactory (e.g., programming) to unsatisfactory (e.g., mathematics). Fine-tuning LLMs in specific domains to build educational applications upon these trained models can circumvent this issue, examples include ChemBERTa [15] or MathChat [16]. However, training for downstream tasks requires specialized data corpora and the final product is tied to the language of such data. Understanding the capabilities of most accessed, general-purpose chatbots is relevant to both introduce them as a pedagogical tool in classrooms, but also counteract inaccuracies students and teachers are exposed to while interacting with generative AI.

### B. LLMs in Math and Geometry Problem-Solving

Although the entire scholar curricula is affected, the presence of AI impacts differently according to the competences and skills to be acquired by students, depending on whether they involve, for example, language abilities, communication, problem-solving capabilities, researching factual information or critical thinking.

Given its current level of adoption by students, it becomes increasingly important to evaluate LLMs performance on specific tasks, such as in this case Geometry problem-solving. It is worth noticing that, as pointed out by [17], autoregressive language models are trained for predicting the next word given a previous sequence of words. The mismatch between the problem the model was developed to solve and the task that is being given, can have significant consequences. In fact, the authors highlight the importance of viewing LLMs not as a “Math problem solver” but rather as a “statistical next-word prediction system” being used to solve Math problems. Then, failures can be understood directly in terms of a conflict between next-word prediction and the task at hand.

Different LLMs have been tested on multiple mathematical reasoning datasets showing how these models struggle to solve problems even at the level of a graduate student. In [1] a new natural-language dataset, named GHOSTS<sup>5</sup>, was introduced. This dataset that covers graduate-level Mathematics and was curated by researchers working in Mathematics includes a subset, named Olympiad-Problem-Solving, consisting of a selection of exercises often used to prepare for Mathematics competitions. The study over this dataset concluded that ChatGPT cannot get through a university Math class, but for undergraduate Mathematics, GPT-4 can offer sufficient (but not perfect) performance. In a quantitative comparison of GPT versions in different subsets of GHOSTS it was shown that Olympiad problem solving was the subset proving to be the more difficult for these models, obtaining lower scores in such problems than even for symbolic integration.

<sup>3</sup> <https://www.bing.com/chat>

<sup>4</sup> <https://www.oma.org.ar/internacional/may.htm>

<sup>5</sup> <https://github.com/xyfrieder/science-GHOSTS>

GPT-2 and GPT-3 were tested in the Mathematics Aptitude Test of Heuristics (MATH) dataset [2] consisting of problems from high school Math competitions classified in different subjects and levels. GPT-2 accuracy reached an average of 6.9%, being better at problems of Pre-Calculus and Geometry and worse for problems related to Number Theory. GPT-3, in turn, reaches an average accuracy of 5.2%, being better at pre-Algebra and worse at Geometry. In [3], an study on the performance of ChatGPT on Math word problems (MWP) from the dataset DRAW-1K<sup>6</sup> found that it changes dramatically if it is asked to provide explanations of the answer instead of simply being asked for the answer without further text. PaLM [4] version of 540-billion parameters reported to solve 58% of the problems in GSM8K<sup>7</sup>, a benchmark of thousands of challenging grade school level Math questions, with 8-shot chain-of-thought prompting in combination with an external calculator. In turn, this result outperforms the prior top score of 55% achieved by fine-tuning the GPT-3 175B model with a training set of 7500 problems and combining it with an external calculator and verifier [5].

A few studies can be found comparing multiple available chatbots answers for Math problems. In [18] an evaluation of the Mathematics performance of Google Bard in solving Mathematics problems commonly found in the Vietnamese curricula was presented. The work findings indicate that in this regard Google Bard's performance falls behind its counterparts (Bing Chat and ChatGPT). For these experiments, a Vietnamese dataset was translated into English since Bard lacks support for Vietnamese at the moment the study was carried out. A comparison between three chatbots like ChatGPT-3.5, ChatGPT-4 and Google Bard was presented in [19], focusing on their ability to give correct answers to Mathematics and Logic problems. For a set of 30 questions, it was found that for straightforward arithmetic, algebraic expressions, or basic logic puzzles, chatbots may provide accurate solutions, although not in every attempt. For more complex Mathematics problems or advanced logic tasks, their answers were unreliable.

Mechanisms to improve the ability of LLMs to complex reasoning are based on generating a chain of thought, i.e. a series of intermediate reasoning steps. Chain-of-thought prompting (CoT) [20] leverages intermediate natural language rationales as prompts to enable LLMs to first generate reasoning chains and then predict an answer for an input question. On the GSM8K benchmark of Math word problems, for example, chain-of-thought prompting with PaLM 540B outperforms standard prompting by a large margin and achieves new state-of-the-art performance, surpassing even finetuned GPT-3 with a verifier. In the same direction, an evaluation on difficult high school competition problems from the MATH dataset was presented in [16] and MathChat, a conversational problem-solving framework was proposed. It simulates a mock conversation between an LLM assistant using GPT-4 and a user proxy agent working together to solve the Math problem. On the problem with the highest level of difficulty from MATH, MathCat improves the accuracy from 28% of GPT-4 to 44% and has competitive performance across all the categories of problems.

Multimodal LLMs (MLLMs) seem to be the most appropriate option to complement reasoning capabilities with the spatial thinking needed to Geometry problem-solving. However, even the most advanced MLLMs still exhibit limitations in addressing geometric problems due to challenges in accurately comprehending geometric figures [21]. Specifically, the model struggles with understanding the relationships between fundamental elements like points and lines, and in accurately interpreting elements such as the degree of an angle. It has been argued [21] that the inaccurate descriptions for geometric shapes produced by models such as GPT4-V (GPT4 with vision) reside on the fact that the model struggles with understanding the relationships between fundamental elements like points and lines, and in accurately

interpreting elements such as the degree of an angle. Current solutions like G-LLaVA [21], built upon LLaVA (Large Language and Vision Assistant) model [22], involve enriching the training data and creating augmented datasets (Geo170K) for improving model training. As mentioned before, the resulting models are less accessible than general-purpose ones and available a mainstream language as English.

With large language models rapidly evolving, there is a pressing need to understand their capabilities and limitations in the context of mathematical reasoning and, particularly, in specific fields like Geometry. Current studies have been centered on measuring the performance of LLMs on benchmarks of broad sets of Mathematical problems in English. To the best of our knowledge, this is the first work focusing on understanding question-answering capabilities of the widely available chatbots regarding Geometry in Spanish language.

### C. Geometry in the Classroom

Geometry is one of the basic subjects of Mathematics. For analyzing Geometry in the context of Argentine educational system, in which the present study takes place, three edges need to be considered: curricular design, actual work in classrooms and the Argentine Mathematics Olympiads (OMA<sup>8</sup>). In the first case, one of the four priority learning blocks proposed by the Argentine Ministry of Education is Geometry [23]. Thus, the vast majority of the curricular designs of each jurisdiction prescribe studying Geometry throughout the secondary education (both in the basic and higher levels). The curricular relevance of Geometry derives from its close relationship with various fields, including Natural and Social Sciences, as well as everyday life [24]–[26]. However, even though Geometry continues to be present in secondary school curricular designs, various researchers highlight the absence of Geometry in the classroom [24],[27]. The third edge corresponds to a competition that has been taking place in Argentina for more than 30 years: the Argentine Mathematics Olympiads [28]. The fundamental objective of these Olympiads is to stimulate mathematical activity among young people and develop the ability to solve problems (OMA, regulations, art 2.). The OMA proposes the resolution of problems, which can be grouped into two large types: arithmetic-algebraic and geometric.

In summary, the official curricular guidelines propose studying Geometry in secondary school, however, this guideline is not materialized in the classrooms (or it is, but weakly). Moreover, Geometry is one of the two types of problems that are used to assess mathematical skills of the students who participate in the OMA. We highlight the importance given to OMA because it is not only promoted by educational centers, but also by provincial governments (as it can be seen in their official site), motivating students to participate actively. In this work we explore how various resources from generative AI can be used to study geometric problems.

## III. MATERIALS AND METHODS

The goal of the analysis carried out in this work is to explore the performance of chatbots when dealing with a problem involving Geometry notions at the level of second and third year of high-school curricular design. The assessment of chatbots capacity of providing accurate answers and, or in the case of failure, the common mistakes and deficiencies found in the described solutions, can serve as basis for the creation of more efficient teaching methodologies involving generative AI.

For the purpose of this study, an Olympiad problem was selected, as described in section A, and the answers of three chatbots, enumerated in section B, to its formulation were collected. The methodology used for analyzing these answers is described in section C.

<sup>6</sup> <https://paperswithcode.com/dataset/draw-1k>

<sup>7</sup> <https://paperswithcode.com/dataset/gsm8k>

<sup>8</sup> <https://www.oma.org.ar/>

TABLE I. SUMMARY OF ERRORS FOUND IN THE ANSWERS OF CHATBOTS

Error type	ChatGPT 3.5				Bing Chat				Bard			
	#1	#2	#3	Total	Precise	Balanced	Creative	Total	#1	#2	#3	Total
Construction	2	0	2	4	-	0	3	3	3	0	1	4
Conceptual	2	3	0	5	-	3	0	3	0	2	0	2
Contradiction	0	0	1	1	-	0	1	1	0	0	0	0
Total	4	3	3	10	-*	3	4	7	3	2	1	6

\* This is a case in which the chatbot did not provide a solution to the problem.

A. Geometry Problem

The problem used in this work belongs to the May Olympiads, an Iberoamerican Mathematics contest. This competition has 2 levels, the first level is for students who, in the year previous to the contest, are under 13 years old at December 31st, and the second level is for students under 15 years old at December 31st. In each level the test is unique, and it consists of 5 problems that students must solve within 3 hours. From these problems, a Geometry problem of level 1 proposed at May Olympiads in 2018<sup>9</sup> [29] was considered.

The problem selected is characterized by not having an immediate and unique solution. In fact, reaching a solution requires knowledge about regular polygons and their properties, circumference and its properties, similarity between polygons, the Pythagorean theorem, trigonometric ratios, among other concepts. Therefore, it is necessary to know and understand a variety of geometrical notions to decide which is the most appropriate to reach a solution.

The geometric problem was selected in such a way that both the mathematical concepts involved and the procedures for its resolution correspond to what is indicated in the official curricular design for Argentine secondary schools [23]. In these designs, the Ministry of Education proposes the minimum knowledge that must be taught in each discipline for each year of the Argentine secondary level. In particular, in Mathematics and in the Geometry area, for students aged 12–13 years old, the study of figures is proposed, arguing about the analysis of properties. In correspondence with the selected problem, students are encouraged to: determine points that meet conditions related to distances and construct circumferences, circles, bisectors and perpendicular bisectors as geometric spaces; explore different constructions of triangles and argue about necessary and sufficient conditions for their congruence; construct similar figures from different information and identify necessary and sufficient conditions of similarity between triangles; analyze claims about properties of figures and argue about their validity, recognizing the limits of empirical evidence; formulate conjectures about properties of figures (in relation to interior angles, bisectors, diagonals, among others) and produce arguments that allow them to be validated. Therefore, the problem analyzed in this work, although it may not be a typical high-school task, involves the concepts that should be addressed at school according to what is prescribed by the Argentinian curricular design.

The problem statement is as follows:

**Problem Statement**

Sea ABCDEFGHIJ un polígono regular de 10 lados que tiene todos sus vértices en una circunferencia de centro O y radio 5. Las diagonales AD y BE se cortan en P y las diagonales AH y BI se cortan en Q. Calcular la medida del segmento PQ.

**English translation:** Let ABCDEFGHIJ be a regular 10-sided polygon that has all its vertices in a circumference with center O and radius 5. The diagonals AD and BE intersect at P and the diagonals AH and BI intersect at Q. Calculate the length of segment PQ.

<sup>9</sup> [https://www.oma.org.ar/enunciados/enunciados\\_Mayo2018.pdf](https://www.oma.org.ar/enunciados/enunciados_Mayo2018.pdf)

The solution proposed by the OMA [29] is based on the graphic representation of the decagon and the identification of the segment that needs to be calculated (PQ). The suggested strategy for reaching the solution consists in drawing segments that join the vertices of the decagon with its center and diagonals. The analysis of the triangles and trapezoids that result from the constructions allows to infer that the triangles are isosceles. From this analysis it is concluded that the requested segment has the same length as the radius of the circumference in which the decagon is inscribed. This resolution enables to find the exact value of the length of the segment PQ, which is 5 cm.

B. Chatbots and LLMs

The three major, freely accessible chatbots available at the time of this article were used for collecting answers for the previous problem. Each of these chatbots rely on its own large language model, an AI model designed to understand and generate human-like text based on deep learning techniques, learned on different corpus using also different learning strategies. LLMs have a large number of parameters and are trained over a massive amount of text data from different sources to capture complex language patterns and relationships. Specifically, the chatbots used for this study were:

**ChatGPT:** ChatGPT (September 25 version) trained over GPT-3.5 language model is the original chatbot launched by OpenAI in November, 2022.

**Bing Chat:** the chatbot accessible through Microsoft Bing search engine and running on GPT-4. This chat offers answers in three modes: (1) More Creative: responses are original and imaginative, creating surprise and entertainment; (2) More Precise: responses are factual and concise, prioritizing accuracy and relevancy; and (3) More Balanced: responses are reasonable and coherent, balancing accuracy and creativity in conversation.

**Bard:** the chatbot developed by Google AI and powered by PaLM-2 large language model.

For this analysis, zero-shot learning was employed. This is, LLMs were asked to answer the question directly, without any prior data or example questions. The prompt was the problem statement in Spanish exactly as in the original text of the Olympiad competition. For each model, 3 answers were obtained by regenerating the responses in order to account for the randomness in text generation.

C. Methodology

Beyond the correctness of the solution itself, the answers provided by chatbots were scanned for identifying reasoning mistakes and inaccuracies in the generated chain-of-thought, individual steps and operations. Basically, it was checked if the appropriate notions were recalled and correctly applied and if the chatbot was able to generate a coherent answer with an accurate solution.

In the process of analyzing the answers of chatbots to the stated Geometry problem, several mistakes of different types were identified. After grouping these mistakes according to their nature, we propose a general categorization of errors. Mistakes made in solving the problem were classified into three main types or categories:



- **Construction:** in this category we find errors originated on the representation made on the plane of the geometric elements indicated in the text answer given by a chatbot. In other words, a construction error is a mismatch between the textual response and the actual geometric figures and their graphical representation. For example, the chatbot ensures that a central angle has  $72^\circ$  when the actual amplitude according to the description given of the figure's elements is necessarily a different one.

Construction errors denote a lack of comprehension of the LLMs of the spatial relationships among elements like points, lines and angles. As the description of the geometric problem reasoning advances, it starts to lose correlation with the actual graph that materializes such description. More likely, there errors stem from the inability of generative AI to understand the semantics behind these geometric notions at the level required for geometric reasoning.

- **Conceptual:** errors in this category relate to incorrect definitions, the application of properties without guaranteeing the necessary conditions or mixing measurement units (e.g. units of length with those of amplitude). An example of conceptual error can be applying the Pythagorean theorem to a not right-angled triangle. The possible causes of these mistakes can be varied. Language generation tools based on AI are capable of producing text using geometric vocabulary, which allows them, for example, to give a reasonable explanation of the Pythagorean theorem. However, as a consequence of an inadequate knowledge and representation of geometric shapes, they are also likely to offer solutions that apply the theorem incorrectly or make inaccurate calculations. LLMs can also suffer from a deficient context description, which in a next-word mechanism is the previous sequence of words. Then, the omission of relevant information reduces the precision in text prediction. The deficient description of the context includes simply missing some piece of information (e.g. the amplitude of a given angle), but also well-known properties (e.g. that the angles of a triangle must sum to 180 degrees) and common assumptions. Furthermore, LLMs are data-driven models trained on data that might include generalized mistakes and misconceptions. Due to their probabilistic nature, LLMs are then prone to reproduce them.
- **Contradiction:** in a number of reasoning steps, contradictions arise as an inconsistency between a deduction and either information involved in the following reasoning steps or the representation on the plane. In other words, the chain-of-thought contains contradictory knowledge, which invalidates the whole reasoning. For example, a contradiction can be inferring that an angle is acute while the graphical representation built starting from this deduction depicts a straight angle.

The mentioned categories groups a number of mistakes found in the solutions provided by chatbots. In a single answer, one or more of these mistakes were identified, leading to a conjunction of errors that ended up in a wrong answer to the problem. This general classification of mistakes found in the collected answers enables to reach a better comprehension about the failures on geometric reasoning of LLM generated texts.

#### IV. RESULTS & DISCUSSION

In order to compare the performance of chatbots according to the provided responses, which due to space limitations are not detailed here, Table 1 summarizes the total number of errors found within each category. For ChatGPT 3.5 three responses were generated, Bard also offers three versions of the answer through its interface, and Bing Chat provides three answers in the form of the more precise, the more balanced and the more creative one.

From the 9 answers (3 for each model) extracted from ChatGPT 3.5, Bing Chat and Bard, only one of them indicated the correct value of the PQ segment length, i.e. only one provided the correct solution to the stated problem, this corresponds to the Bard response #2. However, the model arrived at the result through a method having conceptual errors, thereby it cannot be considered a satisfactory solution either. In addition, there was a case in which the chatbot did not provide a solution at all, this is the case of Bard when it is asked for the More Precise answer to the question. The answer pointed out some decagon properties, but ends up saying (translated from Spanish): "*However, this calculation can be quite complicated and would require in-depth knowledge of the Geometry of the decagon. I would recommend that you consult a Geometry textbook or online resource for a detailed explanation of how to perform these calculations.*"

Overall, the general performance of LLMs in generating a text for answering the Geometry problem stated was disappointing, completely failing at providing an accurate answer to the problem at hand and making a considerable number of mistakes of different types along the reasoning process. This is a concerning finding, considering that the problem presented is a high-school level one, designed for students under 13 years old, which are likely to access chatbots looking for help and would receive not only unreliable answers, but possible introducing or reaffirming Geometry misconceptions.

Considering the type of errors made by each chatbot, ChatGPT 3.5 and Bard were the ones exhibiting more errors belonging to the *Construction* type. Additionally, ChatGPT 3.5 contains a greater number of errors of the *Conceptual* category. Less frequent in all answers are the errors in the "Contradiction" category, accounting for one error of ChatGPT 3.5 and one of Bing Chat, but none in Bard.

For illustrating the different types of errors found in the analyzed answers, Tables II, III and IV provide examples of each type of the errors existing in the actual answers from the model. The tables include a fragment of the response (2nd column) generated by a chatbot (indicated in the 1st column) based on the corresponding LLM when queried using the problem statement and a description of the mistake made (3rd column). In the last column, observations related to the error detected are commented accompanied by a graph, made by the authors of this paper, based on the indications provided in the response.

In the first of them, Table II, the errors refer to the construction of angles (ChatGPT3.5), the construction of right triangles (Bing Chat) and supplementary angles (Bard). Then, in Table III, the errors that are exemplified refer to units of length and amplitude (ChatGPT3.5), to lengths of diagonals of the decagon (Bing Chat) and to heights of triangles (Bard). Table IV contains prototypes of statements about the equality of segments of different lengths (ChatGPT3.5), and mismatch between exterior and interior angles (Bing Chat). In this case, Bard does not contain errors of the *Contradiction* type.

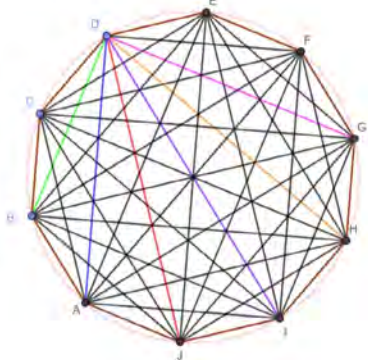
From a qualitative point of view, responses of the different chatbots powered by the corresponding LLM, exhibit some distinctive characteristics.

In the case of the ChatGPT 3.5, some observations can be made about the provided answers. First, the generated text do not justify the mathematical procedures. Justification and reasoning about solutions are important goals in mathematical work in general, and for students who solve math problems in particular. Research shows that this is an issue that needs attention as there are important differences in the way math problems are approached and how justification supports the proposed solutions [30], [31]. Second, the notions used are not explicitly identified and mentioned as denoted in the field. Third, the response does not offer definitions and/or clarifications that could contribute to the resolution.

TABLE II. EXAMPLES OF CONSTRUCTION ERRORS PRODUCED BY THE MODELS

Chatbot	Fragment of the response	Description
ChatGPT 3.5	<p>Ahora, consideremos el triángulo OPQ formado por las diagonales AD y BE que se cruzan en P y las diagonales AH y BI que se cruzan en Q. Este triángulo tiene un ángulo en el centro O que subtende el arco AB (36 grados), y los otros dos ángulos en O son iguales a <math>\alpha</math> (los ángulos en los vértices A y B).</p> <p>Usaremos la propiedad de la suma de los ángulos de un triángulo, que nos dice que la suma de los ángulos en un triángulo es siempre igual a 180 grados. Entonces:</p> $36^\circ + \alpha + \alpha = 180^\circ$ $2\alpha + 36^\circ = 180^\circ$ $2\alpha = 180^\circ - 36^\circ$ <p><b>English translation:</b></p> <p>Now, consider the triangle OPQ formed by the diagonals AD and BE that intersect at P and the diagonals AH and BI that intersect at Q. This triangle has a central angle O that subtends the arc AB (36 degrees), and the other two angles at O are equal to <math>\alpha</math> (the angles at vertices A and B).</p> <p>We will use the property of the sum of the angles of a triangle, which tells us that the sum of the angles in a triangle is always equal to 180 degrees. Then:</p> $36^\circ + \alpha + \alpha = 180^\circ$ $2\alpha + 36^\circ = 180^\circ$ $2\alpha = 180^\circ - 36^\circ$	<p>Fig. 1. Graphical representation in GeoGebra of ChatGPT 3.5 response.</p> <p>Considering the OPQ triangle, the angle with center O is <math>108^\circ(3\alpha</math> with <math>\alpha=36^\circ</math>) as can be observed in Fig. 1. The model states erroneously that the angle with center O is <math>72^\circ(2\alpha</math> with <math>\alpha=36^\circ</math>)</p>
Bing Chat (More creative)	<p>3. Aplicando el teorema del ángulo exterior, hallé que el ángulo BPD mide <math>72^\circ + 36^\circ + 108^\circ</math> y el ángulo AQH mide <math>72^\circ + 72^\circ = 144^\circ</math>.</p> <p>4. Usando el teorema de los ángulos opuestos por el vértice, deduje que el ángulo APD también mide <math>108^\circ</math> y el ángulo BQH también mide <math>144^\circ</math>.</p> <p>5. Por último, aplicando el teorema de Pitágoras en el triángulo rectángulo OPQ, encontré que la medida del segmento PQ es:</p> $PQ^2 = OP^2 + OQ^2 - 2 \cdot OP \cdot OQ \cdot \cos(\widehat{POQ})$ $PQ^2 = 5^2 + 5^2 - 255 \cdot \cos(108^\circ + 144^\circ)$ $PQ^2 = 50 - 50 \cdot \cos(252^\circ)$ $PQ^2 = 50 + 50 \cdot 0.9511$ $PQ^2 = 97.555$ $PQ = \sqrt{97.555}$ <p><b>English translation:</b></p> <p>3. Applying the exterior angle theorem, I found that angle BPD measures <math>72^\circ + 36^\circ + 108^\circ</math> and angle AQH measures <math>72^\circ + 72^\circ = 144^\circ</math>.</p> <p>4. Using the vertex opposite angles theorem, I deduced that angle APD also measures <math>108^\circ</math> and angle BQH also measures <math>144^\circ</math>.</p> <p>5. Finally, applying the Pythagorean theorem to the right triangle OPQ, I found that the measure of the segment PQ is:</p> $PQ^2 = OP^2 + OQ^2 - 2 \cdot OP \cdot OQ \cdot \cos(\widehat{POQ})$ $PQ^2 = 5^2 + 5^2 - 255 \cdot \cos(108^\circ + 144^\circ)$ $PQ^2 = 50 - 50 \cdot \cos(252^\circ)$ $PQ^2 = 50 + 50 \cdot 0.9511$ $PQ^2 = 97.555$ $PQ = \sqrt{97.555}$	<p>Fig. 2. Graphical representation in GeoGebra of Bing Chat response.</p> <p>The model states that the OPQ triangle is a right one, pink triangle in Fig. 2, but it is not. The POQ angle is <math>108^\circ</math> because it is the sum of 3 central angles corresponding to the 3 sides of the decagon involved and each of them is <math>36^\circ</math>. Thus, each of the other 2 angles is <math>(180^\circ - 108^\circ)/2 = 36^\circ</math>.</p>
Bard	<p>➤ Solución:</p> <p>Dado que el polígono es regular, todos los ángulos interiores son iguales. En un polígono de 10 lados, la suma de los ángulos interiores es <math>(10 - 2) \cdot 180^\circ = 1440^\circ</math>. Como cada ángulo interior es igual, cada uno mide <math>1440^\circ / 10 = 144^\circ</math>.</p> <p>Los ángulos <math>\angle APD</math> y <math>\angle BQE</math> son suplementarios, por lo que <math>\angle APD + \angle BQE = 180^\circ</math>. Sustituyendo los valores de los ángulos, obtenemos <math>180^\circ = 144^\circ + \angle BQE</math>. De esta ecuación, se deduce que <math>\angle BQE = 36^\circ</math>.</p> <p>Los ángulos <math>\angle AOB</math> y <math>\angle BOC</math> son iguales, ya que son opuestos por el vértice O. Por lo tanto, <math>\angle AOB = \angle BOC = \frac{180^\circ - 144^\circ}{2} = 18^\circ</math>.</p> <p><b>English translation:</b></p> <p>Since the polygon is regular, all interior angles are equal. In a 10-sided polygon, the sum of the interior angles is <math>(10-2)*180^\circ=1440^\circ</math>. Since each interior angle is equal, each one measures <math>1440^\circ/10=144^\circ</math>.</p> <p>The angles <math>\angle APD</math> and <math>\angle BQE</math> are supplementary, so <math>\angle APD + \angle BQE = 180^\circ</math>. Substituting the values of the angles, we obtain <math>180^\circ = 144^\circ + \angle BQE</math>. From this equation, it follows that <math>\angle BQE = 36^\circ</math>.</p> <p>The angles <math>\angle AOB</math> and <math>\angle BOC</math> are equal, since they are opposite by the vertex O. Therefore, <math>\angle AOB = \angle BOC = \frac{180^\circ - 144^\circ}{2} = 18^\circ</math></p>	<p>Fig. 3. Graphical representation in GeoGebra of Bard response.</p> <p>The model considers that the APD angle is supplementary to BQE, but APD is already <math>180^\circ</math> as can be observed in Fig. 3.</p>

TABLE III. EXAMPLES OF CONCEPTUAL ERRORS PRODUCED BY THE MODELS

Chatbot	Fragment of the response	Description
ChatGPT 3.5	<p>La longitud del segmento PQ es igual a la longitud del arco PQ, que es <math>(5/4)\pi</math> unidades.</p> <p><b>English translation:</b> The length of the segment PQ is equal to the length of the arc PQ, which is <math>(5/4)\pi</math> units.</p>	<p>A radian is a unit of measurement for angles. The model gives the length of the segment PQ in radians, therefore, it confuses the units of length with those of amplitude.</p>
Bing Chat (More creative)	<p>Este problema se puede resolver utilizando la fórmula de la longitud de la diagonal de un polígono regular. La longitud de la diagonal de un polígono regular de n lados es igual a:</p> $d = \frac{s^2(n-2)}{2}$ <p><b>English translation:</b> This problem can be solved using the formula for the length of the diagonal of a regular polygon. The length of the diagonal of a regular polygon with n sides is equal to:</p> $d = \frac{s^2(n-2)}{2}$	 <p>Fig. 4. Graphical representation in GeoGebra of Bing Chat response.</p> <p>The model proposes a formula to calculate the length of the diagonals of the decagon. It assumes that all diagonals that can be drawn have the same length, which, as shown in Fig. 4, is incorrect.</p>
Bard	<p>Por lo tanto, el segmento <math>\overline{PQ}</math> es la altura del triángulo isósceles <math>\triangle POQ</math>.</p> <p>Si <math>x</math> es la medida del segmento <math>\overline{PQ}</math>, entonces <math>QO = 5 - x</math>. Por el teorema de Pitágoras en <math>\triangle POQ</math>, tenemos que</p> $x^2 = (5 - x)^2 + 5^2$ $x^2 = 25 - 10x + x^2 + 25$ $2x^2 - 10x = 0$ $x(2x - 10) = 0$ $x = 0 \text{ o } x = \frac{10}{2} = 5$ <p>Como <math>x</math> es la medida de un segmento, <math>x &gt; 0</math>. Por lo tanto, la medida del segmento <math>\overline{PQ}</math> es <math>\boxed{5}</math>.</p> <p><b>English translation:</b> Therefore, the segment <math>\overline{PQ}</math> is the height of the isosceles triangle <math>\triangle POQ</math>. If <math>x</math> is the measure of the segment <math>\overline{PQ}</math>, then <math>QO = 5 - x</math>. By the Pythagorean theorem in <math>\triangle POQ</math>, we have that</p> $x^2 = (5 - x)^2 + 5^2$ $x^2 = 25 - 10x + x^2 + 25$ $2x^2 - 10x = 0$ $x(2x - 10) = 0$ $x = 0 \text{ or } x = \frac{10}{2} = 5$ <p>Since <math>x</math> is the measure of a segment, <math>x &gt; 0</math>. Therefore, the measure of segment <math>\overline{PQ}</math> is 5.</p>	<p>The answer considers that the triangle <math>POQ</math> is isosceles. It also considers <math>PQ</math> to be the height of the triangle, but this is not the case. Based on this, using the Pythagorean theorem, it proposes to calculate the segment <math>PQ</math>. It is not noticed that the triangle <math>POQ</math> is not a right angle one, because the angle <math>POQ</math> measures <math>108^\circ</math> as explained above. The correct value of the segment <math>PQ</math> is found in this response, but the procedure is incorrect.</p>

In terms of this general characterization of responses, in the first response, Bing Chat explains the characteristics of the decagons, the properties of the angles and the sides, but it does not solve the problem at all. Instead, the chatbot limits itself to suggest consulting a Geometry book or online resources. On the contrary, in the second and third answers, it uses an equation editor (instead of equation written in a textual manner) and suggests resorting to external websites (which are linked) either to reference figures or schemes included in the answer (answer #2) or to refer the reader to more examples of solved Geometry problems (answer #3). Both characteristics, using proper notation to better display equations and linking to external sources, are beneficial for students as they can resort to some extra help

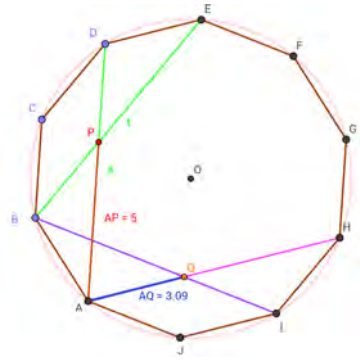
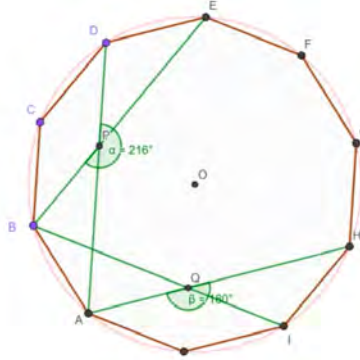
beyond the generated text. Finally, Bard mentions the decagons, the properties of angles and sides, and tries to solve the problem, but the justification of the procedure is incorrect. For answer #2, a justification is developed and, as previously mentioned, the chatbot arrives at the correct answer, but through a procedure containing *Conceptual* errors.

## V. CONCLUSIONS

In this work, we have presented an analysis and comparison of resolutions formulated by three major chatbots such as ChatGPT 3.5, Bing Chat and Bard, to a Geometry problem extracted from the first level of the May Olympiads competition (for students under 13). The



TABLE IV. EXAMPLES OF CONTRADICTION ERRORS PRODUCED BY THE MODELS

Chatbot	Fragment of the response	Description
ChatGPT 3.5	<p>Dado que <math>AP = AQ</math> (ya que ambos son radios de la circunferencia de radio 5), podemos simplificar la fórmula:</p> $PQ^2 = 2(AP^2)(1 - \cos(\angle PQB))$ $PQ^2 = 2(5^2)(1 - \cos(108 \text{ grados}))$ $PQ^2 = 2(25)(1 - \cos(108 \text{ grados}))$ $PQ^2 = 50(1 - \cos(108 \text{ grados}))$ <p>Ahora, podemos calcular el valor de <math>\cos(108 \text{ grados})</math>, que es aproximadamente igual a <math>-0.5</math> (puedes usar una calculadora para obtener un valor más preciso). Luego, podemos sustituir este valor en la ecuación:</p> $PQ^2 = 50(1 - (-0.5))$ $PQ^2 = 50(1 + 0.5)$ <p><b>English translation:</b></p> <p>Since <math>AP = AQ</math> (since both are radii of the circle of radius 5), we can simplify the formula:</p> $PQ^2 = 2(AP^2) (1 - \cos(\angle PQB))$ $PQ^2 = 2(5^2) (1 - \cos(108^\circ))$ $PQ^2 = 2(25) (1 - \cos(108^\circ))$ $PQ^2 = 50 (1 - \cos(108^\circ))$ <p>Now, we can calculate the value of <math>\cos(108^\circ)</math>, which is approximately equal to <math>-0.5</math> (you can use a calculator to get a more precise value). Then, we can substitute this value into the equation:</p> $PQ^2 = 50 (1 - (-0.5))$ $PQ^2 = 50 (1 + 0.5)$	 <p>Fig. 5. Graphical representation in GeoGebra of ChatGPT 3.5 response.</p> <p>It states that AP (5 u) is equal to AQ (3,09 u), both are depicted in Fig. 5.</p>
Bing Chat (More creative)	<p>3. Aplicando el teorema del ángulo exterior, hallé que el ángulo BPD mide <math>72^\circ + 36^\circ = 108^\circ</math> y el ángulo AQH mide <math>72^\circ + 72^\circ = 144^\circ</math>.</p> <p>4. Usando el teorema de los ángulos opuestos por el vértice, deduje que el ángulo APD también mide <math>108^\circ</math> y el ángulo BQH también mide <math>144^\circ</math>.</p> <p>5. Por último, aplicando el teorema de Pitágoras en el triángulo rectángulo OPQ, encontré que la medida del segmento PQ es:</p> $PQ^2 = OP^2 + OQ^2 - 2OP \cdot OQ \cdot \cos(\angle POQ)$ $PQ^2 = 5^2 + 5^2 - 2 \cdot 5 \cdot 5 \cdot \cos(108^\circ + 144^\circ)$ $PQ^2 = 50 - 50 \cdot \cos(252^\circ)$ $PQ^2 = 50 + 50 \cdot 0.9511$ $PQ^2 = 97.555$ $PQ = \sqrt{97.555}$ <p><b>English translation:</b></p> <p>Idem Table II, row 2.</p>	 <p>Fig. 6. Graphical representation in GeoGebra of Bing Chat response.</p> <p>It proposes to calculate the angle BPD by identifying an exterior angle and proposes that the angle measures <math>108^\circ</math> and <math>AQH=144^\circ</math>, this cannot be right because, as the points are aligned, the angle is straight (<math>180^\circ</math>) as can be observed in Fig. 6.</p>
Bard	No <i>Contradiction</i> errors were identified in this model answers	

three chatbots leverage different LLMs, namely GPT-3.5, GPT-4 and PaLM-2, to generate textual responses to natural language queries. In particular, the problem statement as originally presented to students in Spanish was used as a prompt for the chatbots so that three answers were collected from each in order to account for the random components of content generation.

In terms of correctness of the obtained solutions, chatbots had a disappointing performance. Only one answer, provided by Bard, reached the number that was expected ( $\overline{PQ} = 5$ ). However, even when it arrives to the right answer, the described reasoning contains conceptual errors. On the other side, the first response given by Bing Chat does not offer a solution, it only refers the user to consult a Geometry book or some online resource.

In a more detailed analysis of the answers, we found that all of the responses given by the different chatbots contained several types of errors. In a further inspection of these different errors we were able to define a classification encompassing three main categories: construction, conceptual and contradiction. Construction errors correspond to a mismatch between the text description and its geometric representation, conceptual errors involve the incorrect use of geometric concepts and misconceptions, while the last type of error refers to contradictions appearing within the textual description or with respect to the graphical representation.

According to the proposed categorization of errors, ChatGPT 3.5 and Bard made most mistakes within the Construction category. This is an issue related specifically to Geometry as it has to do with the

translation of a geometric specification given in text to a graphical representation. Additionally, ChatGPT 3.5 responses contain a greater number of errors in the Conceptual category, this is, in the application of geometric notions. The Contradiction category is the less frequent one, appearing once in ChatGPT 3.5 answers and once in the ones from Bing Chat, but never in Bard answers.

Most failures observed in the answers to the proposed problem are related to two common criticisms of LLMs [32], the lack of symbolic structure and the lack of grounding. Both questions their capacity to provide human language representation and understanding in spite of their human-like language abilities. The lack of symbolic structure prevents the model to perform formal reasoning and verify reasoning steps, whereas the lack of grounding leads to the misinterpretation of geometric notions and their visual representations. In other words, the fact of being language models poses some limitations for solving more formal problems, such as Geometry ones.

The proposed classification contributes to a better understanding of the failures of LLMs in math-problem solving and, more specifically, those related to spatial representations involved in Geometry problems (e.g. construction errors refers to the relation between the text and its graphical interpretation). The knowledge and recognition of these issues represent also an opportunity to see errors as a valuable educational tool [33]. This categorization can serve as the basis for the construction of methodologies that include the interaction with chatbots in the classroom leveraging on errors to foster their identification, critical thinking of reasoning steps and operations, and reflection on alternative problem solutions.

Although the disappointing results provided by chatbots cannot be directly attributed to the language used, training data in Spanish is known to be smaller than in English. Consequently, next-word prediction performed by LLMs can be assumed to be less precise, thereby the generated lower-quality content. In fact, the reported evaluations of LLMs on different benchmarks including Geometry problems in English, as discussed in section II, showed a better performance than the one achieved with this particular problem. Even though an example is clearly not sufficient to draw conclusions, the language can be considered a source of additional difficulties for LLMs.

Findings of the analysis carried out in this work are specially concerning, considering that the problem presented is a high-school level one, designed for students under 13 years old (although being an Olympiad problem may be beyond the capabilities of a typical of student of that age), which have easy access and are likely to resort to chatbots looking for help to solve similar problems. In this context, they not only will receive unreliable answers in terms of the correctness of the solution to a stated problem, but what is even more serious, they will be also exposed to inaccurate applications of mathematical notions, possibly introducing new misconceptions or reaffirming existing ones. This is also a warning sign for teachers using chatbots to generate course material or exam questions, as they can inadvertently introduce some mistakes.

According to the results obtained in solving the problem stated and taking into account the general characterization of the interface of these tools, it can be concluded that the use of chatbots (and the models behind them) for solving Geometry problems is not appropriate without a critical analysis from teachers as well as the students. The inclusion of these technologies in the classroom must follow a careful methodological approach. Potentially valuable applications of these models in the classroom could be the critically enhanced analysis, supported by teachers, of the responses obtained by chatbots, such as the one presented in this work. This would allow students to discuss and learn Geometry concepts (properties, characteristics, constructions in the plane, etc.) in a practical way. For example, it

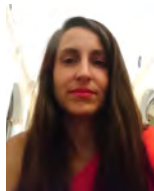
would be useful to distinguish when it is possible (or not) to apply a theorem (lemma, corollary, etc.).

In view of the current wide adoption of chatbot technologies in the classroom and by students of different ages, future work is envisioned to expand the categorization of errors in Geometry problems through the analysis of more problems in different levels. The analysis of a wider variety of problems would likely allow a finer-grained categorization of errors and the emergence of more types, less frequent types of mistakes. Ultimately, systematic evaluations of LLMs performance as the one carried out in this work contributes to the ongoing development of more advanced, capable AI chatbot systems that can be fully integrated in teaching practices to enhance learning processes.

## REFERENCES

- [1] S. Frieder, L. Pinchetti, A. Chevalier, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, J. Berner, "Mathematical capabilities of ChatGPT," 2023.
- [2] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, "Measuring mathematical problem solving with the MATH dataset," in *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [3] P. Shakarian, A. Koyyalamudi, N. Ngu, L. Mareedu, "An independent evaluation of ChatGPT on mathematical word problems (MWP)," in *Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023)*, 2023.
- [4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, "PaLM: Scaling language modeling with pathways," 2022.
- [5] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, J. Schulman, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.
- [6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*, Vancouver, BC, Canada, 2020.
- [7] OpenAI, "GPT-4 technical report," *ArXiv*, vol. abs/2303.08774, 2023.
- [8] F. J. García Pen" alvo, F. Llorens-Largo, J. Vidal, "La nueva realidad de la educación ante los avances de la inteligencia artificial generativa," *RIED-Revista Iberoamericana de Educación a Distancia*, vol. 27, p. 9–39, ene. 2024, doi: 10.5944/ried.27.1.37716.
- [9] B. Memarian, T. Doleck, "ChatGPT in education: Methods, potentials, and limitations," *Computers in Human Behavior: Artificial Humans*, vol. 1, no. 2, p. 100022, 2023, doi: 10.1016/j.chbah.2023.100022.
- [10] B. Han, S. Nawaz, G. Buchanan, D. McKay, "Ethical and pedagogical impacts of AI in education," in *Artificial Intelligence in Education*, Tokyo, Japan, 2023, pp. 667–673.
- [11] J. Flores-Vivar, F. García-Pen" alvo, "Reflections on the ethics, potential, and challenges of artificial intelligence in the framework of quality education (SDG4)," *Comunicar*, 2023, doi: 10.3916/C74-2023-03.
- [12] R. Hadi Mogavi, C. Deng, J. Juho Kim, P. Zhou, Y. D. Kwon, A. Hosny Saleh Metwally, A. Tlili, S. Bassanelli, A. Bucchiarone, S. Gujar, L. E. Nacke, P. Hui, "ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, p. 100027, 2024, doi: 10.1016/j.chbah.2023.100027.

- [13] S. S. Gill, M. Xu, P. Patros, H. Wu, R. Kaur, K. Kaur, S. Fuller, M. Singh, P. Arora, A. K. Parlikad, V. Stankovski, A. Abraham, S. K. Ghosh, H. Lutfiyya, S. S. Kanhere, R. Bahsoon, O. Rana, S. Dustdar, R. Sakellariou, S. Uhlig, R. Buyya, "Transformative effects of ChatGPT on modern education: Emerging era of AI chatbots," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 19–23, 2024, doi: 10.1016/j.iotcps.2023.06.002.
- [14] C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," *Education Sciences*, vol. 13, no. 4, 2023, doi: 10.3390/educsci13040410.
- [15] S. Chithrananda, G. Grand, B. Ramsundar, "ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction," *ArXiv*, vol. abs/2010.09885, 2020.
- [16] Y. Wu, F. Jia, S. Zhang, H. Li, E. Zhu, Y. Wang, Y. T. Lee, R. Peng, Q. Wu, C. Wang, "An empirical study on challenging math problem solving with GPT-4," 2023.
- [17] R. T. McCoy, S. Yao, D. Friedman, M. Hardy, T. L. Griffiths, "Embers of autoregression: Understanding large language models through the problem they are trained to solve," 2023.
- [18] P. Nguyen, P. Nguyen, Bruneau, L. Cao, Wang, H. Truong, "Evaluation of mathematics performance of Google Bard on the mathematics test of the vietnamese national high school graduation examination," 07 2023. doi: 10.36227/techrxiv.23691876.v1.
- [19] V. Plevris, G. Papazafeiropoulos, A. Jiménez Rios, "Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT- 3.5, ChatGPT-4, and Google Bard," 2023.
- [20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.
- [21] J. Gao, R. Pi, J. Zhang, J. Ye, W. Zhong, Y. Wang, L. Hong, J. Han, H. Xu, Z. Li, L. Kong, "G-LLaVA: Solving geometric problem with multi-modal large language model," 2023.
- [22] H. Liu, C. Li, Q. Wu, Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.
- [23] Ministerio de Educación, Argentina, *Núcleos de Aprendizajes Prioritarios. Matemática. Ciclo Básico Educación Secundaria 1° y 2° / 2° y 3° An°* os. 2006.
- [24] R. S. Abrate, G. I. Delgado, M. D. Pochulu, "Caracterización de las actividades de geometría que proponen los textos de matemática," *Revista Iberoamericana de Educación*, vol. 39, pp. 1–9, jun. 2006, doi: 10.35362/rie3912598.
- [25] M. B. López, I. B. Fernández, "Tendencias actuales de la enseñanza-aprendizaje de la geometría en educación secundaria," *Revista Internacional de Investigación en Ciencias Sociales*, vol. 8, no. 1, pp. 25–42, 2012.
- [26] A. M. Bressan, K. Crego, B. Bogisic, *Razones para enseñar geometría en la educación básica: mirar, construir, decir y pensar (1a. ed.)*. Novedades educativas, 2000.
- [27] C. R. Suárez, T. Ángel Sierra Delgado, "Spatial problems: An alternative proposal to teach geometry in compulsory secondary education," *Educação Matemática Pesquisa*, vol. 22, ago. 2021, doi: 10.23925/1983-3156.2020v22i4p593-602.
- [28] L. Santalo, "Olimpiadas matemáticas," *Revista de Educación Matemática*, vol. 6, ago. 2021, doi: 10.33044/revem.11101.
- [29] P. Fauring, F. Gutierrez Eds., *Olimpiadas de Mayo - XVII a XXIV*. Buenos Aires, Argentina: Red Olimpica, 2020.
- [30] B. Glass, C. Maher, "Students problem solving and justification," in *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education*, vol. 2, 2004, pp. 463–470.
- [31] Y. S. Eko, S. Prabawanto, A. Jupri, "The role of writing justification in mathematics concept: the case of trigonometry," *Journal of Physics: Conference Series*, vol. 1097, p. 012146, sep 2018, doi: 10.1088/1742-6596/1097/1/012146.
- [32] E. Pavlick, "Symbols and grounding in large language models," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 381, no. 2251, p. 20220041, 2023, doi: 10.1098/rsta.2022.0041.
- [33] G. M. Zunzarren, "The error as a problem or as teaching strategy," *Procedia - Social and Behavioral Sciences*, vol. 46, pp. 3209–3214, 2012, doi: 10.1016/j.sbspro.2012.06.038.



Verónica Parra

PhD in Mathematics Education from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), 2012. Associate professor in the Teacher Training Department at UNCPBA, member of NIEM Research Institute and Associate researcher at CONICET. Her research interests include mathematics teaching and use of resources for teaching.



Patricia Sureda

PhD in Mathematics Education from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), 2012. Associate professor in the Teacher Training Department at UNCPBA, member of NIEM Research Institute and Assistant researcher at CONICET. Her research interests include mathematics teaching and use of resources for teaching.



Ana Corica

PhD in Education Science from the Universidad Nacional de Córdoba (UNC), 2010. Associate professor in the Teacher Training Department at UNCPBA, director of NIEM Research Institute and Associate researcher at CONICET. Her research interests include mathematics teaching and use of resources for teaching.



Silvia Schiaffino

PhD in Computer Science from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), 2004. Full-time associate professor in the Computer Science Department at UNCPBA, member of ISISTAN Research Institute and Principal researcher at CONICET. Her research interests include recommender systems, user profiling and personalization.



Daniela Godoy

PhD in Computer Science from the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), 2005. Full-time associate professor in the Computer Science Department at UNCPBA, member of ISISTAN Research Institute and Principal researcher at CONICET. Her research interests include recommender systems, social networks and text mining.



# Evaluating ChatGPT-Generated Linear Algebra Formative Assessments

Nelly Rigaud Téllez<sup>1\*</sup>, Patricia Rayón Villela<sup>2</sup>, Roberto Blanco Bautista<sup>3</sup>

<sup>1</sup> Department of Industrial Engineering, FES Aragón, National Autonomous University of Mexico (Mexico)

<sup>2</sup> Universidad Internacional de La Rioja (Mexico)

<sup>3</sup> Department of Computer Engineering, FES Aragón, National Autonomous University of Mexico (Mexico)

Received 26 October 2023 | Accepted 23 January 2024 | Published 13 February 2024



## ABSTRACT

This research explored Large Language Models potential uses on formative assessment for mathematical problem-solving process. The study provides a conceptual analysis of feedback and how the use of these models is related in the context of formative assessment for Linear Algebra problems. Particularly, the performance of a popular model known as ChatGPT in mathematical problems fails on reasoning, proofs, model construction, among others. Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve student's achievement of intended instructional outcomes. The study analyzed and evaluated feedback provided to engineering students in their solutions, from both, instructors and ChatGPT, against fine-grained criteria of a formative feedback model that includes affective aspects. Considering preliminary outputs, and to improve performance of feedback from both agents' instructors and ChatGPT, we developed a framework for formative assessment in mathematical problem-solving using a Large Language Model (LLM). We designed a framework to generate prompts, supported by common Linear Algebra mistakes within the context of concept development and problem-solving strategies. In this framework, the instructor acts as an agent to verify tasks in a math problem assigned to students, establishing a virtuous cycle of learning of queries supported by ChatGPT. Results revealed potentialities and challenges on how to improve feedback on graduate-level math problems, by which both educators and students adapt teaching and learning strategies.

## KEYWORDS

Formative Assessment, ChatGPT, Linear Algebra, Math Word Problems, Polya's Strategy, Prompt Generator.

DOI: 10.9781/ijimai.2024.02.004

## I. INTRODUCTION

**L**ARGE Language Models (LLM) and the emergence of the popular ChatGPT, GPT-3.5 and GPT-4 by OpenAI [1] have spread significant developments in the context of Natural Language Processing. The underlying technology is becoming a meaningful turning point in the field of education [2].

Users enter clear commands or prompts to receive a wide range of natural-language tasks extending from text, image, videos, or code [3]. Such AI-driven educational dialogues have the potential to be a tool in education, as shown by the growing body of research, where attention focus in the improvement of active and personalized learning experience, reinforcement of learning, and assistance of the teaching processes [4] [5].

For instance, the rapid success of ChatGPT in a noticeably brief time seems to be an extremely useful tool to provide simple explanations of complex concepts [6], generate interactive educational materials like quiz questions and draft scripts for classes [7] [8]. Also, this technology can summarize longer texts [9], emphasize relevant content in a subject [10], provide learning through examples and generate formative

assessment [11]. It can also improve meaningful learning by assigning writing tasks [12], generate code explanations [13], or build up critical thinking by asking students to analyze responses of ChatGPT [14].

Moreover, the use of this technology could support the generation of statistical reports with measurements of skills and knowledge [15].

Nevertheless, implementing AI-based initiatives in education requires meticulous modeling and evaluation to ensure their effectiveness in supporting academic improvement [16]. While LLM has shown its accuracy as above mentioned, when reasoning tasks engage in the realm of solving math word problems, ChatGPT may provide erroneous outputs, presentation of false information as truth in cognitive tasks [17] or causing variations in motivational or metacognitive effects [18], elicited by feedback. Consequently, the accuracy of feedback to help students could be compromised.

### A. Math Word Problems

Verbal narratives, often expressed through less accurate descriptions, refer to math word problems presented in educational settings. These sorts of problems offer a comprehensive indicator of mathematical skills [19], exemplified in admissions exams designed to assess mathematical literacy.

Word problems present a realistic context described in a few sentences, where questions or dilemmas are sometimes accompanied by symbols, graphics, and pictures. Solving them requires applying mathematics [20].

\* Corresponding author.

E-mail address: nerigaud@unam.mx

The relevance of math word problems has increased because they support learning over math areas, for instance, algebra, linear algebra, counting and probability, geometry, number theory or intermediate algebra.

Also, math word problems can strengthen the potential of math learning over different subjects and aim to gain experience in accordance with their organization by complex levels of thinking and reasoning through solving problems strategies [21][22].

As math word problems usually present a textual format enriched by models and formulas, textbooks constitute a fundamental part of the teaching-learning process in the classroom, likewise, they serve as a basis for generating more balanced recommendations on the type of skills that one wishes to develop in the engineering student [23].

Given that books are a dominant educational resource that instructors review and use in teaching mathematics, these sources should facilitate opportunities for students to gain experience in problem-solving or developing new learning strategies or methods, for instance, based on common math mistakes [24][25].

In particular, and aligning with the purposes of this paper, one can use books of math word problems as a benchmark to evaluate performance of various methods. This includes examining responses when solving math word problems, considering not only accuracy, but also within the context of formative assessment [26].

### *B. Polya's Strategy*

For the provision of thorough feedback and constructive improvement suggestions, we advocate the application of Polya's problem-solving strategy. Introduced by the distinguished mathematician George Polya, this approach comprises four key steps. These four fundamental steps can address intricate mathematical problems in a structured and systematic manner, and encompass:

- Understanding the Problem: Begin by thoroughly understanding the problem statement, identifying the knowns and unknowns, and clarifying any ambiguities.
- Find a strategy: Develop a clear and organized plan to solve the problem. This may involve drawing diagrams, breaking the problem into smaller subproblems, or considering similar problems you have encountered before.
- Execution: Implement your plan step by step, performing calculations and logical reasoning to work towards a solution.
- Looking back: Once you have a solution, review, and verify it for accuracy. Ask yourself if the answer makes sense, if it aligns with your initial understanding of the problem, and if there are alternative approaches or insights that could provide further understanding.

### *C. Formative Assessment*

Research on formative assessment has expanded in a continuum, since Black and William [27] emphasized the need to better understand assessment for learning, as a mean to facilitate interactions between teacher, technology, and students within a learning environment that provides information for the student and teacher about the learner's performance.

Through formative assessment, and in particular by means of feedback, one could raise standards and improve learning, based on the approach of evidence, as an important opportunity to close the gap between current and desired performance by generating valuable information to both, teachers and students, consequently, yielding meaningful activities [28][29]. Moreover, researchers have considered formative assessment as an influence on future performance [30][31].

To identify concepts involved in providing effective feedback, some authors [32] found models and characteristics of feedback, where some of the most cited authors are Hattie and Gan [33]. Additionally, Jonsson, Panadero and Lipnevich [34][35] proposed a model, also instructional recommendations linked to different types of feedback: tasks (refers to understanding and performance when doing a task), process (the strategy needed to understand or perform a task), self-regulation (regulation of actions), and self (personal and affective aspects) [32].

Normally, teachers typically provide feedback such as comments related to the task and the self-level (personal). It is not common for them to offer comments on a solution process needed to perform the task, or at the metacognitive level (self-regulation), oriented to regulate and actively engage students' own learning [34].

More recent definitions on feedback associate tasks with information, considering it as the essence of feedback: instructors communicate it to the student with the intention of modifying his/her behavior linked to the learning. Jonsson and Panadero [34] consider as relevant components: information, gap, involved agents, and students active processing. In the latest definitions and models, Carless and Boud [36], also include similar components and oriented on how to help students to use the feedback.

From that point, Lui and Andrade [30], Panadero and Lipnevich [28], and Boud [37] posit the interaction of additional factors involved in formative assessment, which include internal process of the learner, such as motivation, and emotions elicited by feedback. These factors are related directly to behavioral response and academic achievement.

In this sense, the general model of Hattie and Gan [33] might be useful for the specific area of math word problems [38]. Despite the model of Panadero [28] requiring more research, their integrative model of feedback includes affective, motivational, and self-regulated learning processes that represent an important aspect when learning mathematics.

### *D. Purpose of the Study*

Feedback is essential for formative assessment in the context of math word problems [38], and the intention goes toward identifying what constitutes valuable feedback, critical attributes for receptiveness and effective use of feedback supported by LLM.

It seems LLM can enhance formative assessment through machine capabilities [39][40], where some stages might occur; (a) students solve math word problems through prompts, (b) ChatGPT receives answers or queries from students (full or partial), (c) the analysis carried out by the LLM models that involves summarizing and interpretations to feedback, and adaptation, as the information oriented to adjust teaching and learning [41].

As noted, ChatGPT can provide general answers, however math problems require precision and attention, and even the most insignificant mistake can lead to incorrect answers and frustration.

Therefore, when experienced instructors identify common math mistakes, this could lead to valuable learning opportunities.

The objective is to develop a framework for formative assessment in mathematical problem-solving using LLM. This framework aims to generate prompts, supported by common Linear Algebra mistakes within the context of concept development and problem-solving strategies.

The objective of this research is to highlight the conjunction between teacher evaluations and their integration with ChatGPT during an evaluation process. We took this initiative driven by the observed underperformance of students in Linear Algebra. The study aims to leverage the combined strengths of both human teaching expertise and ChatGPT's language model capabilities, enriched by

the collective teaching experience. The underlying assumption is that through an adequate and comprehensive assessment involving both agents, teachers and ChatGPT, the student performance can be enhanced, potentially alleviating negative emotions associated with studying this subject.

We focus on examining feedback in math word problems and evaluate the potential of ChatGPT, when oriented with prompts in the process of solving mathematical problems. Two main questions are: What is the contribution of ChatGPT or the instructor in formative assessment considering its appropriate components? and is it possible to propose prompts based on a methodology that includes knowledge of common errors and formative components?

In the following sections, based on the theoretical and empirical background, also, from research questions, we present the research method and main results.

Finally, we discuss theoretical, methodological, and practical implications in the context of math learning and formative assessment supported by LLM.

## II. METHODS

### A. Materials

To conduct this experiment, we chose the subject of Linear Algebra due to its recognition as a relevant mathematics. However, students find its learning challenging. Also, teachers find it challenging to teach.

We used a popular book of Linear Algebra named “Linear Algebra and its applications” by Lay and other authors [42] which includes a special section “Practice Problems”. These problems serve to address potential challenges within the exercises or serve as a valuable prelude, and their solutions often include beneficial tips and cautions concerning homework.

We implemented a distance learning class, where students had to address, for this experiment, a set of five Linear Algebra practice problems from the specified textbook, aligning with the curriculum of a Linear Algebra course. Below, there are the five practice problems arranged from the easiest to the most difficult:

Problem one. “Construct one different augmented matrix for linear systems whose solution set is  $x_1=-2$ ,  $x_2=1$ ,  $x_3=0$ ”.

Problem two. “Suppose the solution set of a certain system of linear equations can be described as  $x_1=5+4x_3$ ,  $x_2=-2-7x_3$ , with  $x_3$  free. Use vectors to describe this set as a line in  $R^3$ ”.

Problem three. “Suppose a  $4 \times 7$  coefficient matrix for a system of equations has 4 pivots. Is the system consistent? If the system is consistent, how many solutions are there?”

Problem four. “Suppose an economy has three sectors: Agriculture, Mining, and Manufacturing. Agriculture sells 5% of its output to Mining and 30% to Manufacturing and retains the rest. Mining sells 20% of its output to Agriculture and 70% to Manufacturing and retains the rest. Manufacturing sells 20% of its output to Agriculture and 30% to Mining and retains the rest. Determine the exchange table for this economy, where the columns describe how the output of each sector is exchanged among the three sectors.”

Problem five. “Let  $A$  be a  $4 \times 4$  matrix and let  $x$  be a vector in  $R^4$ . What is the fastest way to compute  $A^2x$ ? Count the multiplications.”

These exercises included two at a basic level, two at an intermediate level, and one at an advanced level. Additionally, a concluding question addressed students’ emotional responses to the learning process, encompassing emotions such as boredom, anxiety, anger, indifference, and frustration [43], which have been identified as pertinent emotional reactions to feedback in mathematical learning [29].

The process and results of each exercise, along with the emotion expressed by the learner, when applicable, were used to formulate a series of prompts. These prompts were designed to elicit feedback from the student before the instructor’s review, considering both a problem-solving approach and the identification of compound emotions.

### B. Participants

Our experiment took place at the Faculty of Superior Studies Aragon from the National Autonomous University of Mexico. The online classes’ main goal is to improve knowledge, comprehension and problem solving of Linear Algebra.

We invited thirty-five low performance students from Industrial (60%), Mechanical (25%) and Electric-electronic (15%) careers to join the Linear Algebra course; therefore, the sample was non-probabilistic.

The total duration of the course was 32 h with four sessions per week. Three experienced teachers instructed students with explanations of Linear Algebra’s fundamental concepts and resolved problems to successfully tackle the set of five Linear Algebra practice problems. Also, as requested, each of the instructors provided help to participants during interventions with ChatGPT.

Furthermore, these three teachers contributed to review and generate manual feedback to students’ responses. Finally, three more teachers conducted a meta-evaluation of the feedback, as well as its comparison with ChatGPT’s feedback.

The main function of ChatGPT was to provide explicit feedback according to user’s prompts.

We informed all participants about the conducted experiment and obtained their consent for data collection during the process, including videotaping.

### C. Tasks and Methods

As a first step, the participating students enrolled in a course of two-hour. They also engaged in assessment exercises and responded to surveys in which they provided information about their self-perception of learning difficulties. As a result of this process, information about whether the student has learned difficulties is stored in the “Common Linear Algebra mistakes” (Table I).

Table I lists a sample of a few common Linear Algebra mistakes related to concept development and problem solving. Three instructors analyzed answers. We classified outputs in accordance with Polya’s strategy [22] and provided exemplifications of recommendations for students based on the prompts.

The diagram on Fig. 1, shows a general process to help the instructor to give better feedback to students based on the Polya’s method, the student’s emotion, and fundamental common Linear Algebra mistakes as an entrance to LLM.

An expert in the math field is necessary to obtain effective feedback, by identifying common errors of the math discipline, which are then stored in the feedback database. In this case, we focus on Linear Algebra problems and utilize the Polya’s method to identify whether the error generated belongs to the comprehension (understanding the problem), planning (find a strategy), doing (execute), or revision stage (looking back). The instructor uses this information from the problem selection and its solution, and adopts a multifaceted strategy encompassing problem-solving processes, self-regulation, self-reflection, and the acknowledgment of mistakes. Within this comprehensive strategy, the instructor leverages these elements to generate prompts, seeking enriched feedback from ChatGPT to provide more insightful and constructive learning experience.

A relevant tool of LLM is the employment of natural language processing to generate prompts. Particularly in the context of this paper, establishing effective communication using LLM like ChatGPT is of great relevance to obtain clear and concrete answers.



TABLE I. EXAMPLE OF FUNDAMENTALS COMMON LINEAR ALGEBRA MISTAKES AND PROMPTS RECOMMENDATIONS

	Understanding the Problem	Find a strategy	Execute	Looking Back
Doesn't identify what the problem is	Provide at least two different descriptions of the problem			
Erroneous selection of appropriate concepts and procedures		Can you explain me the concept of... Can you explain me the method... Why the method ... is not appropriate to solve the problem	Verify the outcome ... Test the solution through method ...	Why the method ... is appropriate to solve the problem
	Doesn't know how to communicate the solution	Express how the problem makes you feel		Why is the solution effective? o Why doesn't the proposed solution cover what was expected? How can I interpret the problem?
Do not identify the characteristics of a system		Can you help me to identify if the system is consistent, inconsistent, or dependent? How can you identify that a system is consistent, dependent or		

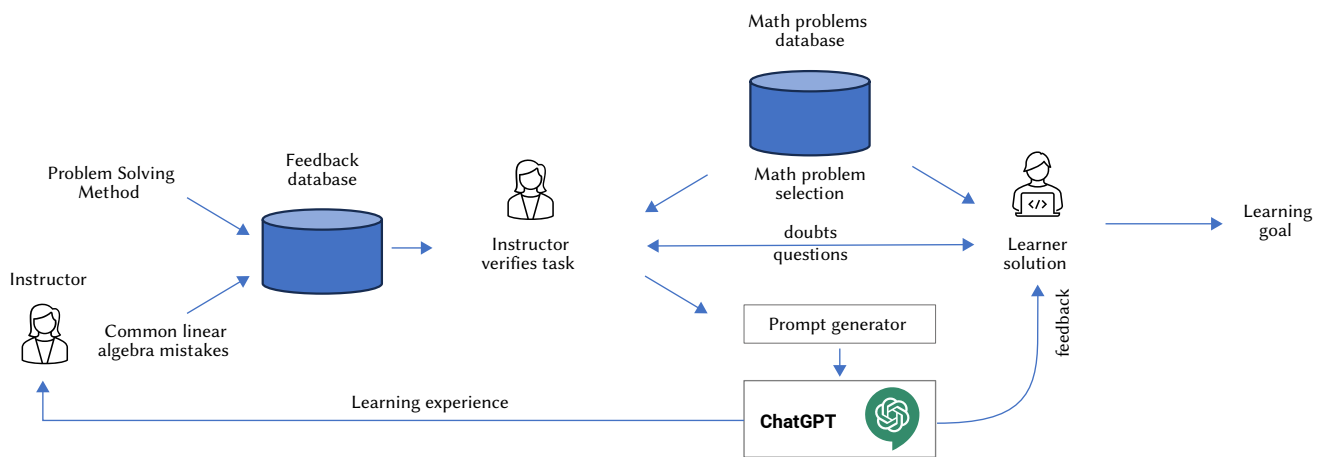


Fig. 1. Framework for formative assessment using LLM for mathematical problems.

As we have mentioned throughout the document, it is important to provide formative assessment to the student and provide some enriched prompts that the student can use with LLM.

To elicit an appropriate and constructive response from ChatGPT for students, one effective approach involves crafting specific questions. These questions serve to generate targeted feedback, incorporating motivating elements that enhance the overall quality of the student’s training.

According to the prompt generation stage, the required information is:

1. Teacher Role and Course Features
2. Criterion
  - i. Give the problem and the correct solution
  - ii. Solving process: Polya’s strategy
  - iii. Solving process stage: compression, planning, doing or revision
  - iv. Specify self-regulation: detail, precision, and tone
  - v. Specify Self: student emotion and recommendations
  - vi. Emphasize the mistake
3. Give the task (problem and solution) to ChatGPT
4. Request ChatGPT, with the information numbered as 1, 2 and 3, to generate a teaching strategy
5. Request ChatGPT to exemplify the strategy according to step 3.

For instance, generic prompts are:

- Prompt 1: *I am (1) the interest is in the following math problem (i).*  
 Prompt 2: *For the given problem consider the (ii) at the phase of (iii), use (iv) for (v).*  
 Prompt 3: *Identify the process stage to improve...*  
 Prompt 4: *Request some resources...*

Some examples for prompt generation are in Fig. 2.

Prompt 1	I am a teacher of Linear Algebra for engineering bachelor the interest is in the following math problem: “Construct an augmented matrix for linear systems whose solution set is $x_1=-2, x_2=1, x_3=0$ ”, test the following “ $4x_1+6x_2+3x_3=-2, -2x_1+5x_2+2x_3=9$ and $x_1-7x_2+4x_3=-9$ ”
Prompt 2	For the given problem consider the Polya’s solving process at the phase of “problem understanding”. Use adequate tone and accuracy for a frustrated student.
Prompt 3	For the given problem consider the Polya’s solving process at the phase of “search strategy”. Use adequate tone and accuracy for a frustrated student.
Prompt 4	Give me recommendations for public link resources for the students to improve “search strategy” for “matrices”

Fig. 2. Example of prompts generation.

To examine the performance of both agents (ChatGPT and teachers) on math word problems in the context of formative assessment we based on the models of Hattie and Timperley [32] to structure relevant components on feedback, in the sense to reduce gaps between current understanding or performance and the learning goal. Furthermore, this study delves into the association of emotions triggered by feedback and self-regulation [29][34], as outlined in a well-established model for mathematical word problems. It analyzes the intricate process of solving these problems, emphasizing a thoughtful and systematic approach for complete comprehension [20]. The results are presented in the following section.

### III. RESULTS

We present the results of the analysis conducted on the feedback from ChatGPT and the teachers in Table II. As observed, we transformed the model components of Hattie and Gan [32] (Task, Solving Process, Self-regulation, and Self) into a sequence of yes-no response questions, which were then used for the assessment.

As seen in Table II, in the ‘Task’ component, teachers outperform ChatGPT, with an 85% accuracy compared to ChatGPT’s 40%. Despite ChatGPT being capable of solving all five problems when requested individually, it makes errors when reviewing solutions generated by others. For instance, one of the most frequent errors was erroneously grading an incorrect student’s response as correct.

TABLE II. EVALUATION OF FORMATIVE ASSESSMENT ON MATH WORLD PROBLEMS

Component	Feedback	Frequency of Yes Answer	
		ChatGPT	Teacher
Task	Does the agent give a correct answer?	40%	85%
Solving process	Does the agent provide elements for understanding the problem? e.g., verbal, schematic, tabular, and so on.	90%	10%
	Does the agent model the problem?	90%	5%
	Does the agent provide calculations to resolve the model?	90%	5%
	Does the agent interpret output(s)?	90%	90%
	Does the agent evaluate the solution?	90%	90%
	Does the agent communicate the whole solution?	80%	5%
Self-regulation	Does the agent show any sort of self-management? a) Awareness of own errors	No	Yes
	b) Timing of feedback	No	Yes
	c) Level of detail	No	Yes
	d) Accuracy	No	Yes
	e) Tone	No	Yes
Self	Does the agent encourage engagement/ commitment through answers?	80%	60%
	Does the agent promote self-efficacy? (recommendations)	90%	90%

In the ‘Solution Process’ component, we observe that there are some aspects in which ChatGPT shows better results than a human. This is because, being an automated process, it can generate longer responses tailored to each situation, including verbal elements to understand the problem, model the problem, display the procedure’s calculations, and most of the time, communicate a final solution. On the other hand, human feedback was shorter (on average, three lines) and focused on determining whether the result was correct or incorrect. In the latter case, it briefly pointed out where in the procedure the student’s first error occurred but did not provide an explanation of what the correct solution and procedure should be.

It is important to note that the evaluators independently analyzed the ‘task’ component within the ‘solution process’ component. For instance, ‘Does the agent provide calculations to resolve the model?’ is assigned ‘yes’ when the agent tries to include such calculations in its feedback, regardless of whether they are correct or not.

In the ‘Self-regulation’ component, evaluators decided that it was challenging to assess these aspects individually in each of the samples and that a global conclusion had to be drawn for the complete set of results.

The conclusion was that although ChatGPT can regulate aspects such as tone, the level of detail in the response, etc., this is done as part of the prompt generation. However, this is externally imposed regulation by a human and not self-regulation. In the case of the teachers, there was no indication of a response that was out of context in terms of tone, level of detail, etc. According to the meta-evaluators of the experiment, all the responses provided by the humans would be the responses a teacher would typically give in a classroom.

Finally, in the ‘Self’ component, we can observe that ChatGPT always considered the result of emotion interpretation to craft feedback and included elements to encourage engagement and promote self-efficacy. In this regard, it is notably contrasting that the teachers’ responses did not exhibit elements indicating that they considered the emotional state of the student, and the feedback was focused on problem-solving.

In the same phase of the experiment, as noted when the difficulty of problems increased, frustration is the most common emotion as shown in Fig. 3.

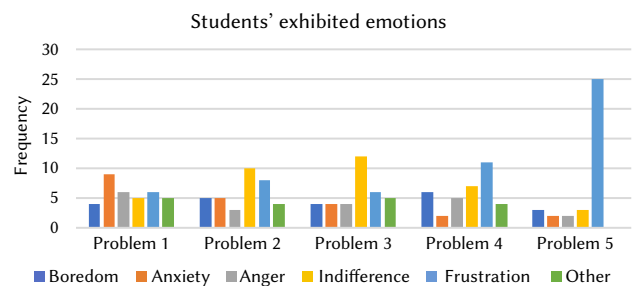


Fig. 3. Students' exhibited emotions.

Fig. 3 shows emotions exhibited by participants and provides frequencies of the experienced emotions by the students, during the solution of set of five Linear Algebra practice problems, from number one to five, as their complexity increase from less complex to more complex.

The suggested emotions include boredom, anxiety, anger, indifference, frustration, and others such as happiness or surprise. In Fig. 3, the frequency of these emotions experienced in each problem of increasing complexity is illustrated. As observed, anxiety decreased as the complexity of the problem increased. This suggests that, as students progressed in solving the problem, they were more focused on the task at hand.

The emotion of boredom remains constant throughout the problem-solving process, diminishing only in the most complex problem. The same pattern is observed for anger. Indifference increases from problem one to problem three and then decreases from problem three to problem five. In the case of frustration, it consistently increases with each new problem and experiences a significant spike in the final one. As observed, frustration appears to be the emotion that could have had the most pronounced negative impact on the group, theoretically suggesting that their performance did not improve.

Fig. 4 shows that four students did not answer any problem, and nineteen answered three problems.

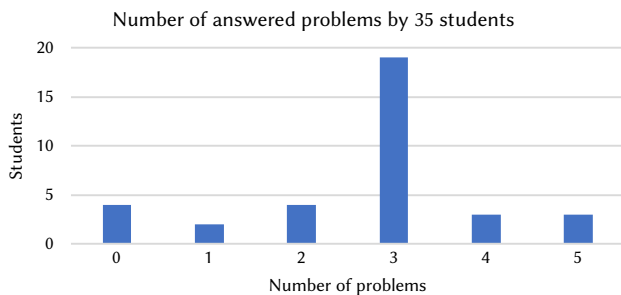


Fig. 4. Number of answered problems per participants.

Fig. 4 illustrates the distribution of students who answered a certain number of problems. The cumulative frequencies within the group of thirty-five students are as follows: four students did not solve any problem, two students answered one problem, four students answered two problems, nineteen students answered three problems, three students answered four problems, and three students successfully completed all five problems.

#### IV. DISCUSSION

ChatGPT is an appropriate tool to provide more effective formative feedback due to the inclusion of four main aspects: tasks, problem solving, self-management, and self.

Interpretation of Fig. 3 and Fig. 4 suggests that there is a need for reinforcement and improvement in students' performance concerning knowledge, attitude, and dominant emotions [11][13]. This interpretation is attributed to a low proficiency in Linear Algebra, lack of comprehension of the problems, and a prominent level of distraction hindering performance improvement. Based on these findings, the recommended approach for the instructor is to prioritize feedback, incorporating both quantitative and qualitative criteria of formative assessment. Implementing strategies like Polya's problem-solving method can aid in enhancing student understanding and regaining their self-confidence.

As seen, in mathematical problems, when ChatGPT is employed independently, its performance is low. Something similar happens to the instructor. However, through the employment of the framework that includes both agents, feedback could improve learning outcomes. Additionally, the support of LLM for the students benefit their motivation to continue their math studies and reinforce math learning [26].

Nowadays, the use of these technologies is particularly important, such as LLM and the adequate use of prompts generators. Moreover, when the teachers function as a guide to construct them, recommendations are strong [44][45].

The transformative impact of LLM on mathematics learning presents key challenges that are central to the scope of this research, as follows:

To effectively integrate tools like ChatGPT into educational settings, it is imperative to establish explicit guidelines encompassing teaching and learning assessment strategies.

Specifically, within the realm of evaluation, ChatGPT tools should play a role in fostering critical thinking and logical reasoning, particularly in STEM careers, where disruptive technologies, such as those facilitated by AI, contribute to innovative and creative environments.

Considering this, prompt engineering becomes essential for shaping the approach to queries directed at ChatGPT. Well-crafted prompts should provide resources, such as relevant books available on the web, and adhere to a clear structure akin to the one proposed by the authors of this paper. This approach ensures that the generated answers are not only accurate but also engaging. The teacher's role is pivotal in this phase, serving as a verifier to confirm the correctness of the responses, as verified by the three teachers during the experiment.

For students struggling in mathematics, experiencing emotions like frustration and indifference that negatively impact their performance, ChatGPT can serve as a valuable tool. Leveraging a more human-like interaction through conversational agents, it has the potential to promote motivation and reinforce positive emotions.

#### V. CONCLUSION

Undoubtedly, the use of AI technologies with LLM represents a tool for educative support, as shown with the proposed feedback framework to improve formative assessment.

As final recommendations at the level of tasks, the instructor could propose a math word problem and assign it to the student. After the student solves it, the teacher reviews and provides regular feedback. The student asks ChatGPT to become an immersive choose-your-own task. The purpose is to reinforce the prior knowledge of the student.

For self-regulation, and from obtained feedback, students reflect and communicate about the mathematical task. Students ask ChatGPT to generate structured activities to correct his/her performance, and to encourage them to think about their learning process and math progress. Therefore, the use of ChatGPT to generate feedback is tailored to each student's needs and goals.

Another conclusion is that the teacher should encourage students to self-assess, reflect, and monitor their math work. The teacher asks ChatGPT to generate self-assessment tools, such as rubrics or the entire process for solving a math word problem that helps students evaluate their own work.

Finally, at the personal level, from provided feedback, the teacher asks ChatGPT to generate follow-up activities that encourage students to apply the feedback they have received.

For self-regulation, students engage in reflective practices based on feedback received. They utilize ChatGPT to request structured activities aimed at correcting their performance and fostering thoughtful consideration of their learning process and mathematical progress. Consequently, the use of ChatGPT for feedback generation is tailored to each student's individual needs and goals.

Alternatively, teachers can empower students to self-assess, reflect, and monitor their mathematical work. In this scenario, the teacher prompts ChatGPT to generate self-assessment tools, based on Polya's problem-solving strategy, such as rubrics or comprehensive guides for solving math word problems, facilitating students in evaluating their own work.

On a personal level, leveraging the feedback provided, the teacher can instruct ChatGPT to generate follow-up activities. These activities are designed to encourage students to apply the received feedback,



promoting a more firsthand and practical application of their learning experience, and reducing negative emotions that hinder academic performance.

This personalized approach contributes to a more comprehensive assessment tailored to individual learning needs, supported by AI technologies.

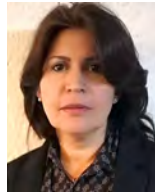
#### ACKNOWLEDGMENT

This paper has been possible thanks to the support received from The National Autonomous University of Mexico, DGAPA, PROJECT PAPIIME PE112723.

#### REFERENCES

- [1] OpenAI, "ChatGPT: Optimizing language models for dialogue," Open AI 2015-2024. Accessed: Aug 13, 2023. [Online]. Available at <https://openai.com/blog/chatgpt/>.
- [2] W.M. Lim, A. Gunasekara, J.L. Pallant, J.I. Pallant, and E. Pechenkina, "Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators," *The International Journal of Management Education*, vol. 21, no. 2, pp. 1-13, 2023, <https://doi.org/10.1016/j.ijme.2023.100790>.
- [3] J. Zhou, P. Ke, X. Qiu, M. Huang, J. Zhang, "ChatGPT: potential, prospects, and limitations," *Frontiers of Information Technology & Electronic Engineering*, pp. 1-6, 2023, <https://doi.org/10.1631/FITEE.2300089>.
- [4] C. K. Lo, "What is the Impact of ChatGPT on Education? A rapid review of the Literature," *Education Sciences* vol. 13, no. 4, pp. 410, 2023, <https://doi.org/10.3390/educsci13040410>.
- [5] R. Gruetzemacher and J. Whittlestone, (2022). "The transformative potential of artificial intelligence," *Futures*, vol. 135, pp. 1-11, 2022, <https://doi.org/10.1016/j.futures.2021.10288.4>.
- [6] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the evil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, no. 1, pp. 1-24, 2023, <https://doi.org/10.1186/s40561-023-00237-x>.
- [7] R. Dijkstra, Z. Genc, S. Kayal, and J. Kamps, "Reading Comprehension Quiz Generation Using Generative Pre-trained Transformers," in *4th International Workshop on Intelligent Textbooks, iTextbooks*, Durham, UK, 2022, pp. 1-14.
- [8] E. Gabajiwala, P. Mehta, R. Singh, and R. Koshy, "Quiz Maker: Automatic quiz generation from text using NLP," in *Futurist trends in networks and computing technologies*, vol. 936, P.K. Singh, S.T. Wierzchoń, J. K. Chhabra, and S. Tanwar, Eds. Springer Lecture Notes in Electrical Engineering, 2022, pp. 523-533.
- [9] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Ellermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, ..., and G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," *Center for Open Science*, vol. 103, 2023, <http://dx.doi.org/10.35542/osf.io/5er8f>.
- [10] X. Zhai, (2022), "ChatGPT user experience: Implications for education," *Social Science Research Network Electronic Journal*, vol. 18, <https://doi.org/10.2139/ssrn.4312418>.
- [11] A. Herft, "A Teacher's Prompt Guide to ChatGPT: Aligned with 'What Works Best,'" CESE NSW "What Works Best in Practice", 2023. Accessed: Aug. 15, 2023. [Online]. Available: <https://usergeneratededucation.files.wordpress.com/2023/01/a-teachers-prompt-guide-to-chatgpt-aligned-with-what-works-best.pdf>.
- [12] A. R. Mills, "Seeing Past the Dazzle of ChatGPT," *Inside Higher Education*, 2024. Accessed: Jan 19, 2023. [Online]. Available: <https://www.insidehighered.com/advice/2023/01/19/academics-must-collaborate-develop-guidelines-chatgpt-opinion>.
- [13] S. MacNeil, A. Tran, D. Mogil, S. Bernstein, E. Ross, and Z. Huang, "Generating diverse code explanations using ChatGPT-e large language model," in *Proceedings of the 2022 ACM Conference of International Computing Education Research*, New York, NY, USA, Association for Computing Machinery 2022, pp. 37-39.
- [14] E.R. Mollick and L. Mollick, "Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts," *The Wharton School Research Paper*, 2023. Accessed: Oct. 15, 2023. [Online]. Available: <https://ssrn.com/abstract=4391243> or <http://dx.doi.org/10.2139/ssrn.4391243>.
- [15] J. F. Wu, "Effective use of machine learning to empower your research," *The Campus Learn, Share, Connect*, 2022. Accessed: Aug 15, 2023. [Online]. Available: <https://www.timeshighereducation.com/campus/effective-use-machine-learning-empower-your-research>.
- [16] A. Tack and C. Piech, "The AI teacher test: Measuring the pedagogical ability of blender and GPT-e in educational dialogues," in *Proceedings of the 15th International Conference on Educational Data Mining*, Durham, UK, 2022, pp. 1-8, <https://doi.org/10.48550/arXiv.2205.07540>, to be published.
- [17] L. M. Sánchez-Ruiz, S. Moll-López, A. Nuñez-Pérez, JA. Moraño-Fernández, and E. Vega-Fleitas, "ChatGPT Challenges Blended Learning Methodologies in Engineering Education: A Case Study in Mathematics," *Applied Sciences*, vol. 13, no. 10, 2023, <https://doi.org/10.3390/app13106039>.
- [18] Shakarian P., Koyyalamudi A., Ngu N., and Mareedu L. (2023). "An independent evaluation of ChatGPT on Mathematical Word Problems,"
- [19] A. R. Strohmaier, F. Reinhold, S. Hofer, M. Berkowitz, B. Vogel-Heuser, and K. Reiss, "Different complex word problems require different combinations of cognitive skills," *Educational Studies in Mathematics*, vol. 109, pp. 89-114, 2022, <https://doi.org/10.1007/s10649-021-10079-4>.
- [20] L. Verschaffel, B. Greer, and E. De Corte, *Making sense of word problems*, Países Bajos: Swets & Zeitlinger, 2000.
- [21] T. S. Barcelos, R. Muñoz-Soto, R. Villarreal, E. Merino, and I. F. Silveira, "Mathematics Learning through Computational Thinking Activities: A Systematic Literature Review," *Journal of Universal Computer Science*, vol. 24, no. 7, pp. 815-845, 2018.
- [22] G. Polya, *Cómo plantear y resolver problemas*, Cd. México, Méx.: Editorial Trillas- Colección "Serie de Matemáticas", 1969.
- [23] S. Frieder, L. Pinchetti, R. R. Griffiths, T. Salvatori, T. Lukaszewicz P. C. Peterses, A. Chevalier, and J. Berne, "Mathematical Capabilities of ChatGPT," *Neural Information Processing Systems- Datasets and Benchmarks Track*, pp. 1-37, 2023, <https://doi.org/10.48550/arXiv.2301.13867>.
- [24] J. K. Kim, M. Chua, M. Rickard, and A. Lorenzo, "ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine," *Journal of Pediatric Urology*, vol. 19, no. 5, pp. 598-604., 2023, <https://doi.org/10.1016/j.jpuro.2023.05.018>.
- [25] A. Tack, E. Kochmar, Z. Yuan, S. Bibauw, and C. Piech, "The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues," in *Proceedings of the 18th Workshop on innovative Use of NLP for Building Educational Applications*, Toronto, Canadian Association for Computational Linguistics, 2023, pp. 785-795, <https://aclanthology.org/2023.bea-1.64.pdf>.
- [26] Y. Hicke, G. W. Masand, and T. Gangavarapu, "Assessing the efficacy of large language models in generating accurate teacher responses," in *Proceedings of the 18th Workshop on innovative Use of NLP for Building Educational Applications (BEA 2023)*, Toronto, Canada, 2023, pp. 745-755.
- [27] P. Black and D. Wiliam, "Developing the Theory of Formative Assessment," *Educational Assessment Evaluation and Accountability*, vol. 21, pp. 5-31, 2009, doi:10.1007/s11092-008-9068-5.
- [28] E. Panadero and A. A. Lipnevich, "A Review of Feedback Models and Typologies: Towards an Integrative Model of Feedback Elements," *Educational Research Review*, vol. 35, 2022, doi: 10.1016/j.edurev.2021.100416.
- [29] A. Ramaprasad, "On the Definition of Feedback," *Behavioral Science*, vol. 28, pp. 4-13, 1983 doi:10.1002/bs.3830280103.
- [30] A. M. Lui and H. L. Andrade, "Inside the Next Black Box: Examining Students' Responses to Teacher Feedback in a Formative Assessment Context," *Frontiers in Education*, vol. 7, pp. 1-14, 2022, <http://dx.doi.org/10.3389/fev.2022.751548>
- [31] L. Allal, "Assessment and the Co-regulation of Learning in the Classroom," *Assessment in Education: Principles, Policy & Practices*, vol. 27, no. 4, pp. 332-349, 2019 doi:10.1080/0969594X.2019.1609411.

- [32] J. Hattie and H. Timperley, "The Power of Feedback," *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, 2007, doi:10.3102/003465430298487.
- [33] J. A. C. Hattie and M. Gan, "Instruction Based on Feedback," *Handbook of Research on Learning and Instruction*, R. Mayer and P. Alexander Editors New York: Routledge, 2011.
- [34] A. Jonsson and E. Panadero, "Facilitating Students' Active Engagement with Feedback," in *The Cambridge Handbook of Instructional Feedback* Editors, London, England: Routledge, 2018, pp. 28.
- [35] A. A. Lipnevich and E. Panadero, "A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions." *Frontiers in Education*, vol. 6, 2021, doi: 10.3389/educ.2021.720195.
- [36] D. Carless and D. Boud, "The development of student feedback literacy: enabling uptake of feedback," *Assessment and Evaluation in Higher Education*, vol. 43, no. 8, pp.1315-1325, 2018.
- [37] D. Boud, "Sustainable Assessment: Rethinking Assessment for the Learning Society," *Studies in Continuing Education*, vol. 22, no. 2, pp. 151–167, 2000, doi:10.1080/713695728.
- [38] A. Lipnevich, F. Preckel, and S. Krumm, "Mathematics attitudes and their unique contribution to achievement: Going over and above cognitive ability and personality," *Learning and Individual Differences*, vol. 47, pp. 70–79, 2016, <https://doi.org/10.1016/j.lindif.2015.12.027>.
- [39] B. McMurtrie, "AI and the future of undergraduate writing," *The Chronicle of Higher Education*, 2022. Accessed: Sept. 12, 2023. [Online]. Available: <https://www.chronicle.com/article/ai-and-the-future-of-undergraduate-writing>.
- [40] A. R. Mills. "ChatGPT just got better: What does that mean for our writing assignments?," *The Chronicle of Higher Education*, 2023. Accessed: March 26, 2023. [Online]. Available: <https://www-chronicle-com.libproxy.library.unt.edu/article/chatgpt-just-got-better-what-does-that-mean-for-our-writing-assignments>.
- [41] J. Warner. "Freaking Out About ChatGPT–Part I", *Inside Higher Education*, 2022. Accessed: Aug. 13, 2023. [Online]. Available: <https://www.insidehighered.com/blogs/just-visiting/freaking-out-about-chatgpt%E2%80%94part-i>
- [42] D. C. Lay, S. R. Lay, and J. J. McDonald, *Linear Algebra and its applications*, Maryland, USA: Pearson (5th Ed.), 2016.
- [43] A. Behera, P. Matthew, A. Keidel, P. Vangorp, H. Fang, and C. Susan, "Associating Facial Expressions and Upper-Body Gestures with Learning Tasks for Enhancing Intelligent Tutoring Systems," *International Journal of Artificial Intelligence in Education*, vol. 30, pp. 236–270, 2020, <https://doi.org/10.1007/s40593-020-00195-2>.
- [44] F. J. García-Peñalvo and A. Vázquez-Ingelmo. "What do we mean by GenAI? A systematic literature mapping of AI-driven solutions for content generation". *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4. pp. 7-16, 2023, doi: <https://doi.org/10.9781/ijimai.2023.07.006>
- [45] S. S. Gill, M. Xu, P. Patros, H. Wu, R. Kaur, K. Kaur, S. Fuller, M. Singh, P. Arora, A. K. Parlikad, V. Stankovski, A. Abraham, S. K. Ghosh, H. Lutfiyya, S. S. Kanhere, R. Bahsoon, O. Rana, S. Dustdar, R. Sakellariou, S. Uhlig, and R. Buyya. "Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 19-23, 2024, <https://doi.org/10.1016/j.iotcps.2023.06.002>.



Patricia Rayón Villela

Patricia Rayón has more than 20 years of experience in research and teaching in Computer Science. Experience in management and participating in research projects related to data mining, artificial intelligence, and pattern recognition issues. She is coordinator of the Master in Artificial Intelligence at UNIR-México and full professor at this university.



Roberto Blanco Bautista

Roberto Blanco received his B. Eng. Degree from the Veracruzana University. He has studies of Systems Engineering from the National Polytechnic Institute. He has more than 50 years of experience in soft computing where he has been combining teaching and counselling for many public and private organizations in soft engineering projects. His research is concerned with knowledge representation, software, and algorithms optimization.



Nelly Rigaud Téllez

Nelly Rigaud is a Full Professor for the Industrial Engineering and Systems Department. Counselor and advisor of the Open and Distance Education System at the National Autonomous University of Mexico. She received her Engineering Doctorate from the Institute of Applied Sciences and Technology in Mexico. She holds a Master of Engineering (Planning and Projects Management) and a degree in Mechanical Engineering. Her research interests include math education and knowledge-based systems, systems modeling and simulation, and decision support systems.

