Universidad Pablo de Olavide (España)
International Journal of Educational Research and Innovation
número 22, 2024
ISSN: 2386-4303
DOI: 10.46661/ijeri.11036
Sección: Artículos
Recibido: 23-09-2024
Aceptado: 04-10-2024
Publicado: 03-12-2024
Páginas: 1-21

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

Inteligencia Artificial Segura, Transparente y Ética: Claves para una Educación Sostenible de calidad (ODS4)

Safe, Transparent, and Ethical Artificial Intelligence: Keys to Quality Sustainable Education (SDG4)

Francisco José Garcia-Peñalvo (*)
Computer Science Department
Research Institute for Educational Sciences (IUCE)
Universidad de Salamanca, Salamanca, Spain
(*) Corresponding author
fgarcia@usal.es

Marc Alier
UPC Universitat Politècnica de Catalunya - Barcelona Tech, Barcelona, Spain author
marc.alier@upc.edu

Juanan Pereira
Computer Science Faculty, UPV/EHU, San Sebastián, Spain
Juanan.pereira@ehu.eus

Maria Jose Casany
UPC Universitat Politècnica de Catalunya - Barcelona Tech, Barcelona, Spain
ma.jose.casan@upc.edu

RESUMEN

La creciente integración de la inteligencia artificial (IA) en los entornos educativos requiere un marco estructurado para garantizar su uso seguro y ético. Se ha propuesto un manifiesto que establece siete principios clave para una IA segura aplicada a la educación, destacando la protección de los datos del estudiantado, la alineación con las estrategias institucionales, la congruencia con las prácticas didácticas, la minimización de errores, interfaces de usuario comprensibles, supervisión humana y transparencia ética. Estos principios están diseñados para

García Peñalvo, F. J., Alier, M., Pereira, J., & Casany, M. J. (2024). Inteligencia Artificial Segura, Transparente y Ética: Claves para una Educación Sostenible de calidad (ODS4). IJERI: International Journal of Educational Research and Innovation, (22). https://doi.org/10.46661/ijeri.11036

guiar la implementación de tecnologías de IA en entornos educativos, abordando riesgos potenciales como violaciones de privacidad, uso indebido y dependencia excesiva de la tecnología. También se introducen las Aplicaciones Inteligentes para el Aprendizaje (SLApps), que integran la IA en el ecosistema tecnológico institucional existente, con especial atención a las plataformas de aprendizaje que son la base de los campus virtuales, permitiendo experiencias de aprendizaje seguras, adaptativas según el rol y específicas para cada asignatura. Si bien los grandes modelos de lenguaje como GPT ofrecen un potencial transformador en la educación, también presentan desafíos relacionados con la precisión, el uso ético y la congruencia pedagógica. Para enfrentar estas complejidades, se recomienda una lista de verificación basada en los principios de IA segura en la educación, que proporciona al profesorado e instituciones un marco para evaluar las herramientas de IA, asegurando que apoyen la integridad académica, mejoren las experiencias de aprendizaje y respeten los estándares éticos.

PALABRAS CLAVE

Inteligencia Artificial en Educación; AIED; Marco de referencia para una IA segura; Aplicaciones Inteligentes para el Aprendizaje (SLApps); Grandes modelos de lenguaje; Integridad académica e IA.

ABSTRACT

The increasing integration of artificial intelligence (AI) into educational environments necessitates a structured framework to ensure its safe and ethical use. A manifesto outlining seven core principles for safe AI in education has been proposed, emphasizing the protection of student data, alignment with institutional strategies, adherence to didactic practices, minimization of errors, comprehensive user interfaces, human oversight, and ethical transparency. These principles are designed to guide the deployment of AI technologies in educational settings, addressing potential risks such as privacy violations, misuse, and over-reliance on technology. Smart Learning Applications (SLApps) are also introduced, integrating AI into the existing institutional technological ecosystem, with special attention to the learning management systems, enabling secure, role-adaptive, and course-specific learning experiences. While large language models like GPT offer transformative potential in education, they also present challenges related to accuracy, ethical use, and pedagogical alignment. To navigate these complexities, a checklist based on the Safe AI in Education principles is recommended, providing educators and institutions with a framework to evaluate AI tools, ensuring they support academic integrity, enhance learning experiences, and uphold ethical standards.

KEYWORDS

Artificial Intelligence in Education; AIED; Safe AI Framework; Smart Learning Applications (SLApps); Large Language Models; Academic Integrity and AI.

1. INTRODUCTION

1.1. The ed-tech landscape until 2022

The worldwide education market was valued at approximately $6.6 billion in 2022 and is expected to expand at a compound annual growth rate of 6.34% during the forecast period, reaching $9.6 billion by 2028 (Dharmadhikari, 2024). 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, the field of 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. However, the early 2000s witnessed a surge in online learning platforms, 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 (García-Peñalvo & Seoane-Pardo, 2015). More recently, advancements in artificial intelligence (AI), virtual and augmented reality, and adaptive learning systems have further personalized the learning experience, catering to individual learning styles and needs (Pelletier et al., 2023; Segovia-García & Segovia-García 2024). 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 one hand, there's an undeniable commercialization, with education increasingly influenced by market-driven models and private enterprises. On the other hand, there's a growing movement towards open-source technologies and freely accessible content repositories. This contrast paints a complex picture of the current educational space, where the forces of the market coexist with a commitment to open access and knowledge sharing (Caulfield et al., 2012).

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:

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 (Castañeda & Selwyn, 2018). Tools like Learning Management Systems (LMS) tend to function more as management tools than learning aids, limiting the understanding of digital technology's role in education (Selwyn, 2016).

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, supporting and improving learning outcomes (Bartolomé et al., 2018). This concept is not new. Back in the 1980s, Seymour Papert (1987) observed similar issues within the LOGO community. He criticized the usual ways of evaluating educational technology, like 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 just looking at how technology fits into education, he suggested a more culturally aware evaluation of its role. This approach from Papert in 1987 remains relevant today as we continue to explore the best ways to integrate technology in learning environments.

Emotional and human impact. It is crucial to understand digital tools' emotional and human impact. These technologies influence students' and faculty'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 (Skinner, 1968). 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. While aimed at personalizing and enhancing learning, this constant surveillance raises privacy and psychological concerns. The educational journey can become heavily algorithm-driven, often without transparently acknowledging underlying decision-making processes.

1.2. Enter generative artificial intelligence

The use of AI is becoming increasingly widespread in today's society (Moral-Sánchez et al., 2023). Since late 2022, chatbots based on Large Language Models (LLMs) (Zhao et al., 2023), spearheaded by the unprecedented success of ChatGPT, have driven rapid and uncoordinated adoption of generative AI (GenAI) in educational settings.

This organic surge has emerged without a comprehensive strategy or regulatory framework, leading to diverse and often inconsistent applications. This highlights the urgent need to evaluate and guide the deployment of LLMs in education to harness their potential while mitigating associated risks (Lim et al., 2023), all of them related to the GenAI potential to generate new digital previously unseen synthetic content, in any form and to support any task, through generative modeling (García-Peñalvo & Vázquez-Ingelmo, 2023).

The dramatic shift in the perception of AI's role in education is illustrated by the overwhelming response to the different faculty development courses related to ChatGPT (or other similar software applications) applied in education from early 2023. These courses attracted many registrations, quickly filling up and necessitating additional groups, yet still failing to meet the demand. Teachers were drawn to these courses due to concerns about potential plagiarism, curiosity about automating academic tasks, and the potential to enhance student learning experiences.

Among the various educational advantages of ChatGPT and similar LLMs, we can mention the educational opportunities (García-Peñalvo et al., 2024) arising from these models, which can generate educational content, support discussions on diverse topics, create quizzes, evaluate assignments, and provide feedback; moreover, they can also assist in explaining complex concepts and offer coding examples in various programming languages; the teachers and students assistance (Dwivedi et al., 2023) of these LLMs suggesting topics, methodologies, and related studies, and finding connections between subjects, helping with statistical analysis, and proposing ideas for further studies; or the writing assistance (Reeves & Sylvia, 2024) through continuous feedback on writing, helping in organizing content, and strengthening the arguments.

However, the use of LLMs in education poses several challenges, too. For example, there are issues with the quality of the used prompts (Morales-Chan, 2023) because the effectiveness of an LLM heavily relies on the quality of the user prompts. High-quality prompting is not an easy task and is somewhat closer to an art than an engineering discipline (Henley et al., 2024); LLMs' variable quality of responses (Yang et al., 2024) due to the quality can fluctuate, especially in areas where the training data is limited or not comprehensive; the hallucinations (Huang et al., 2023) LLMs produce, it means content that seems credible but is actually false or irrelevant, which is particularly problematic in education where accuracy is crucial. However, this phenomenon can also be leveraged as a teaching opportunity. Educators can encourage critical thinking and media literacy skills by presenting students with examples of LLM-generated content that contains hallucinations. For instance, students can be asked to (a) detect and identify hallucinations in a given text, (b) explain why they think the LLM generated incorrect or irrelevant information, (c) discuss the potential consequences of relying on inaccurate information, or (d) develop strategies for verifying the accuracy of the information, especially when working with AI-generated content; the privacy, security, and legal concerns (Iskender, 2023), storing sensitive data on AI applications poses risks when most of the GenAI companies do not guarantee that conversations with their chatbots will not be used for other purposes, like training new models, except in paid enterprise plans.

AI's emotional impact on teachers and students must also be considered. According to Cambridge University's "AI and Scholarship: A Manifesto," AI can diminish key moments of satisfaction, such as "eureka moments" during research and learning. Over-reliance on AI can reduce the pleasure and fulfillment teachers and students derive from performing academic tasks manually, which is essential for skill development and personal growth. Teachers may feel anxious about losing control over academic integrity or even their relevance, while students may develop a false sense of competence by over-depending on AI tools (McPherson & Candea, 2024). Additionally, AI challenges include its potential to exacerbate inequalities, particularly when students have unequal access to premium versions of AI tools (Center for Teaching Innovation, 2024). Moreover, the over-reliance on technology (Duong et al., 2024) might cause dependency on the GenAI tools and anxiety that could be related to diminishing creativity and critical thinking skills (Choi et al., 2023), although it can also be a great tool to develop these skills if used properly (Vartiainen & Tedre, 2023).

LLMs' responses might reproduce hidden biases (Kamath et al., 2024) in their training data; there is a potential lack of human interaction (Choi et al., 2023). While AI chatbots aid learning, they cannot substitute essential human interaction in students' development; or the ethical concerns, which can be related to lack of content authorship (Johinke et al., 2023), plagiarism, and other dishonest uses (Gašević et al., 2023), or differential access and usage of these tools due to pricing and cost of premium paid versions (Cotton et al., 2024).

At the regulatory level, frameworks have emerged to guide the ethical use of AI in education. The European Commission, through its Digital Education Action Plan, emphasizes the need for fairness and transparency in AI applications. Projects like AgileEDU and AI4T advocate for developing explainable and transparent AI tools that promote inclusivity and equal access (European School Education Platform, 2024). The European AI Alliance also brings together stakeholders to develop ethical guidelines to prevent bias and ensure fair access to AI-driven education. These regulatory efforts aim to address the growing concern that AI tools, if not properly regulated, could perpetuate existing inequalities and create new forms of exclusion.

LLMs lack integration into the logical framework of educational activities, whether individual or group dynamics, teacher supervision, or learning analytics (García-Peñalvo, 2024). They are not fine-tuned (Christiano et al., 2023) to align with course-level content, educational activities, or specific pedagogical models. Teachers often see LLMs as threatening their roles and disrupting traditional functions. While some educators attempt to integrate these models into their activities, they face significant challenges such as content quality, lack of expertise in creating effective prompts, and absence of supervision mechanisms to monitor LLM student interactions. This can lead to issues like the aforementioned hallucinations, where the LLM generates incorrect or misleading information (Alier-Forment & Llorens-Largo, 2023).

The most notable effort to use a fine-tuned LLM for all-around educational purposes is Khan Academy's Khanmigo, which has not lived up to its hype and marketing efforts. Sethye (2024) analyzed the performance of Khanmigo as a language learning tool, specifically for learning French, through about 17.5 hours of interaction using Chapelle's (2001) evaluation framework for discerning the task appropriateness of a given Computer-Assisted Language Learning (CALL) tool. Results suggest that while holding some promise, Khanmigo does not show robust performance in all six criteria suggested for evaluation.

The state of the art suggests that there is no structured educational context for using LLMs, especially if they are intended to be an all-purpose tool for every scenario without an educational strategy or defined purpose. However, the emerging capabilities of LLMs and other GenAI technologies constitute a groundbreaking set of new technologies that may enable a new generation of new learning tools. In other words, new learning tools can be built on top of LLMs and tailored to the educational ecosystem, processes, and strategies.

1.3. Research goals

We have stated that a chatbot alone based on an LLM might not be the right tool for educational contexts. However, LLMs and related technologies can be used as platforms to develop learning tools that fit educational contexts precisely.

This paper aims to identify the principles that software tools based on GenAI tools must comply with to be able to be deployed in educational contexts. This implies addressing integration issues within the learning institution strategies, control, legal compliance, and ethical issues.

2. PRINCIPLES FOR A SAFE AND ETHICAL USE OF ARTIFICAL INTELLIGENCE IN EDUCATION

Before delving into the specific principles that define the safe and ethical use of AI in education, it is essential to provide a structured framework to guide the integration of these technologies. The following sections outline a set of fundamental guidelines that ensure AI applications align with educational strategies and maintain the necessary levels of security, accuracy, and ethical integrity. By adhering to these principles, educational institutions can harness the potential of AI while mitigating risks related to privacy, misuse, and the accuracy of information, all of which are crucial for safeguarding the quality and fairness of the learning process.

2.1. Safe artificial intelligence in education

Alier, García-Peñalvo, and Camba (2024) make a straightforward definition of "Safe AI in Education" that proposes five principles (extended with other two principles SAIE6 and SAIE7) that a safe AI application to be used in an educational system should fulfill:

(SAIE1) **It guarantees confidentiality**. The system must ensure the security and confidentiality of all student data, including identities, roles, academic records, and interactions.

(SAIE2) **It is aligned with educational strategies**. AI tools must align with the institution's strategy and IT governance policies to ensure they support educational goals and comply with operational standards. For example, AI tools should support learning and content creation yet be designed to prevent misuse, such as cheating or circumventing academic integrity measures. The system should not readily offer solutions to assignments or assist in paraphrasing to bypass plagiarism checks.

(SAIE3) **It is aligned with didactic practices**. AI applications must conform to predefined educational parameters when deployed in educational settings.

(SAIE4) **Accuracy and minimization of errors**. Despite their training in extensive data repositories, leading models like GPT-4 still risk delivering incorrect information or generating hallucinations. A safe AI system must prioritize the accuracy and relevance of its outputs, a task that becomes more feasible within narrowly defined application contexts.

(SAIE5) **Comprehensive interface and behavior**. The AI system should be presented in a manner that is understandable to students and teachers, clarifying its intended uses and limitations.

(SAIE6) **Human oversight and accountability**. AI tools in education must always complement, not replace, human educators. While AI can assist with administrative tasks like grading or providing feedback, all decision-making processes must remain under human supervision. AI-driven decisions should be explainable, and students must have the right to appeal these decisions through human-led processes. This ensures fairness, maintains the role of teachers as mentors, and protects the integrity of the educational process.

(SAIE7) **Ethical training and transparency**. AI models used in education must be trained in an ethical manner, with a clear commitment to transparency regarding the sources of training data and the methodologies used. It is essential that these models actively work to minimize biases and provide transparency about their training processes, allowing educators and students to understand the limitations and considerations involved in the AI's outputs.

These seven principles have been compiled on the Safe AI in Education Manifesto (https://manifesto.safeaieducation.org/) (Alier-Forment et al., 2024), which has been subscribed to by many academics and practitioners. The principles are constantly being revised and updated by the community of signatories.

These principles have strong implications for the kind of AI products to be used in education.

The SAIE1 principle (confidentiality guarantee) requires the educational institution to have a level of control over the AI tool so students' privacy and confidentiality are safeguarded. This can be achieved by owning and operating the whole technology stack or by requiring privacy in the service agreements with AI tools vendors. Needless to say, using free tools that require students to register on a website - like https://chatgpt.com - should be out of the question. Students can do out of their own will, but they should not be required to do so by faculty or to complete learning assignments. So far, the literature review regarding this topic shows that primary research rarely addressed privacy problems, such as participant data protection during educational data collection, and that there is a need to create or improve ethical frameworks (Alam & Mohanty, 2022; Fichten et al., 2021; Li et al., 2021; Manhiça et al., 2022; Otoo-Arthur & Van Zyl, 2020; Salas-Pilco et al., 2022; Salas-Pilco & Yang, 2022; Zawacki-Richter et al., 2019; Zhai & Wibowo, 2023).

The SAIE2 principle (alignment with educational strategies) creates tension with using general-purpose tools, like ChatGPT, which are intended to fit multiple use cases. Because of this, general-purpose LLMs might not be a good fit at the institutional level. This is a problem because:

- The complexity of using an LLM chatbot is deceptive; prompt engineering is proving to be a very complex discipline (Willison, 2023). This complexity should not be added to the complexity of a learning process. Adding complexity to a learning process is bad pedagogical practice because it increases the students' cognitive load (Chen et al., 2023).
- LLM chatbots are fine-tuned to follow the user's instructions, so avoiding their use for cheating, plagiarism, or other misuse is nearly impossible (González-Geraldo & Ortega-López, 2024).
- LLM chatbots always provide an answer. However, the quality of the answer varies from good to appalling. While a user with a certain expertise in a given domain will have enough knowledge and strategies to discriminate between good and bad answers, students can be deceived by hallucinations, incorrect answers, or answers not aligned with the educational institution.

The SAIE2 principle also requires that educational software has to be integrated with the educational institution's technological ecosystem strategy. This has other ramifications, such as accessing the tools using the institution credentials, the user interface complies with the institution's branding, and the tool complies with the institution's ecosystem governance policies (Bond et al., 2024).

The SAIE3 principle (alignment with the didactic practices) introduces the same problems as the SIAE2 principle but at a more specific level. If AI tools are used within a course, faculty, lecturers, and teachers need to clearly understand how the AI tool will fit into their instructional design. Examples of AI aligned with instructional design can be found in engineering or medical education (Hwang et al., 2024; Rabelo et al., 2024). This principle has implications for the tools:

- While remaining a viable option for an AI tool, the chat interface might not be the only approach. Specific didactic usages will require specific interfaces. For example, the VS Code programming environment uses LLMs as coding assistants and plugins that make suggestions in the code editor.
- There are challenges and considerations because faculty need to clearly understand how AI tools will fit within their instructional design. This includes selecting appropriate interfaces, configuring the AI's behavior to align with learning objectives, and ensuring that the tool supports rather than complicates the learning process (Bond et al., 2024).
- Just like with the online learning tools implemented within an LMS, the teacher might need to set up and configure the behavior of the AI tool. For example, in the domain of classic online learning tools, a teacher provides the quiz tool with a set of questions, correct answers, grading instructions, and time constraints.

The SAIE4 principle (accuracy and minimization of errors) is critical in educational settings. Since hallucinations are inherent to the current state of the art of GenAI technologies, this is a complex task, but there are ways to minimize erroneous or misleading answers. First, the task becomes more feasible within narrowly defined application contexts. Second, making AI tools to reference sources used to craft answers helps minimize errors and provides a method of answer validation. After all, the references and sources can be hallucinated, but a reference can be checked.

There is a need for rigorous quality assessments in AI tools to ensure they provide accurate information. A meta-systematic review of IA in higher education indicates that a significant proportion of the studies reviewed did not undertake comprehensive quality assessments, raising concerns about the reliability of the AI tools being used (Bond et al., 2024).

The SAIE5 principle (comprehensive interface and behavior) calls for experimentation with the interfaces and kinds of responses that AI educational tools provide. The tools need to make their usefulness and limitations explicit. As opposed to the behavior of ChatGPT, it mimics an omniscient all-purpose agent that delivers plain wrong or hallucinated answers with a pose of authority and confidence. Some AI systems deliver responses with unwarranted confidence. This behavior can mislead users, particularly in educational contexts where students may lack the expertise to critically evaluate the AI's outputs.

The SAIE6 (human oversight and accountability) calls for prioritizing human oversight of IA systems to ensure the educational process's fairness, transparency, and integrity. While AI can assist in tasks like grading or providing feedback, decisions must ultimately remain under human supervision to prevent the risks associated with over-reliance on automation. Studies have shown that AI-driven systems, such as those used for grading, are prone to errors and biases, which require human intervention to detect and correct (Fügener et al., 2022). Mouta et al. (2023) emphasize the ethical necessity of involving human educators in AI-driven processes, particularly when students' academic futures are at stake. Furthermore, explainability in AI decision-making is critical for maintaining accountability. Selbst and Barocas (2018) argue that decisions must be understandable to educators and students, allowing for the right to appeal and challenge AI-driven conclusions. Without such oversight, AI tools risk diminishing the human connection essential to the educational experience, a concern echoed by the AI Now Report (Whittaker et al., 2018), which calls for AI applications to support, rather than replace, human educators.

The SAIE7 (ethical training and transparency) is related to SAIE 4 and SAIE6. Many researchers in the field call for transparency and addressing biases in AI-driven tools used in education. This emphasizes the need for ethical considerations in AI model training processes and understanding IA systems' limitations (Guan et al., 2020; Mouta et al., 2024). Due to the current state of the art of LLMs, which requires huge amounts of training data, and the overwhelming training costs (Xia et al., 2024) are only available to well-funded organizations, the SAIE7 principle is likely to be compromised by any usage of LLMs. However, closed and open-sourced models now allow for fine-tuning of the models, which requires smaller datasets (Irugalbandara et al., 2024) that can be under the control or supervision of the learning institution or ethics evaluator.

This definition of Safe AI in Education emphasizes the integration of AI into educational settings in a manner that supports and enhances the teaching and learning experience while safeguarding against potential misuse and ethical concerns.

2.2. Smart learning applications

This "Safe AI in Education" definition has design implications for AI-based technologies to be used in education. Alier, Casañ, and Amo (2024) introduced the concept of Smart Learning Application, an advanced AI educational tool that goes beyond traditional learning applications because they take advantage of AI technologies - such as LLMs or diffusion models - for content generation. A learning application is part of or can be integrated within an LMS like Moodle, Sakai, or Blackboard, where they have appropriately termed "activities," such as a Forum, a Wiki, a Task, or a quiz. Smart Learning Applications are also crafted to function within the specific boundaries of a course and fit into its instructional design.

To ensure this, a Smart Learning Application (SLApp) must:

(SLApp1) Ensure secure access, which the educational institution manages. Utilizing the LMS for authentication and authorization.

(SLApp2) Adapt to learning roles. The LMS customizes the application's features to match the user's role: teacher, student, guest, or administrator.

(SLApp3) Provide course-specific settings. Each application instance is directly associated with a course and a learning activity, enabling a customized educational journey. For instance, these settings are represented explicitly in the Moodle user interface in the "breadcrumbs" links that display the course and activity where the user stands.

(SLApp4) Allow leveraging LLMs via APIs. To facilitate features such as on-the-fly content creation and personalized learning trajectories. This means that it can be used with all the LLMs and AI models that the state-of-the-art can offer, either provisioned by vendors as a service or open-source models running in the institution'' infrastructure.

This strategy boosts interactivity and customization, tackling challenges like guaranteeing content accuracy and adhering to data privacy laws. The goal is to present educational technology that is more closely aligned with educational objectives and capable of upholding academic integrity and offering.

The Safe AI in Education principles (SAIE 1-7) and the Smart Learning Apps characteristics (SLApp 1-4) provide a set of requirements for an AI-powered educational tool. The next section will discuss how Learning Assistants can be designed and built to satisfy these requirements.

3. SAFE AI IN EDUCATIONAL TECHNOLOGIES

Before evaluating the practical application of existing AI tools in educational settings, it is necessary to examine whether these tools comply with the established principles of Safe AI in Education. The following section aims to assess widely used AI technologies, such as ChatGPT, in terms of their alignment with key ethical and educational standards. This analysis will provide insight into how these tools perform concerning confidentiality, accuracy, didactic integration, and human oversight and whether they can be effectively integrated into educational contexts without compromising academic integrity or institutional strategies.

3.1. Is ChatGPT a safe AI in education?

Let us consider whether the popular ChatGPT complies with Safe AI in Education principles. For this, we will consider students and faculty using the free version of ChatGPT, which was opened to the public on Nov 30, 2022.

(SAIE1) ChatGPT does not guarantee the confidentiality of information. OpenAI may use the users' conversations for future model training, fine-tuning, or evaluation. For all we know, OpenAI personnel or consultants may have access to each free tier conversation (Salas-Pilco & Yang, 2022).

(SAIE2) ChatGPT is not aligned with educational strategies. ChatGPT is an obscure piece of technology. It is currently based on a closed LLM (GPT-3.5 and GPT-4). We do not have information about its architecture, training dataset, or fine-tuning process; its system prompt is secret, and we can only assess its future performance on previous observations (Shin, 2021).

(SAIE3) It is aligned with didactic practices. ChatGPT is a general-purpose system that will try to please the user and do its bidding. So, it will help the students cheat, copy, and paraphrase.

(SAIE4) At the bottom of its ChatGPT page, OpenAI states that, regarding the accuracy and minimization of errors, "ChatGPT can make mistakes. Check important info." The Chatbot is prone to hallucinations and changes in behavior as OpenAI changes the service and model.

(SAIE5) Related to the interface and behavior comprehension, the authoritative way ChatGPT delivers its correct or incorrect responses is confusing and counterproductive for students (Bender et al., 2021; Shin, 2021).

(SAIE6) Conversations between the students and ChatGPT are only available to OpenAI. While this availability to OpenAI potentially conflicts with privacy and confidentiality regulations like GDPR in Europe and FERPA in the USA, it does not provide oversight or accountability.

(SAIE7) Regarding ethical training and transparency, the dataset and algorithms used by OpenAI are a trade secret. There is no lack of controversy about the sources of the training dataset.

This quick analysis suggests that while ChatGPT is an interesting tool, and we can learn a lot by experimenting with it, ChatGPT does not comply with any of the stated principles of Safe AI in Education.

Let us consider a special application of ChatGPT, GTPs. OpenAI offers a product called GPTs, which are customized versions of ChatGPT designed for specific tasks or purposes. These GPTs allow users to tailor the system more effectively in various activities, such as learning, teaching, or assisting with specific work-related tasks. Users can create these customized models without requiring coding skills by simply providing instructions and additional knowledge and specifying the functions the GPT should perform, such as web searches, image generation, or data analysis. GPTs can be used individually, for internal company use, or shared with others (https://openai.com/index/introducing-gpts/). In this case, we can provide a little didactic alignment (in compliance with SAIE3) and make the GPT use specific, accurate, and relevant information (improving the standing relative to SAIE4.

3.2. Learning assistants as smart learning applications

Since late 2022, interest in searching for “AI assistant” has surged nearly tenfold, according to Google Search Trends. This increase reflects the rise of a new type of AI tool. A notable example is the search engine perplexity.ai (https://perplexity.ai), which responds to user queries by providing answers based on the search content and includes references to sources. It effectively uses LLMs for natural language processing, content analysis, summarization, and content generation, resulting in comprehensive responses with verifiable links and citations from authoritative sources.

In 2024, OpenAI introduced a new product called “Assistants,” built on its LLMs, shifting away from earlier enhancements to ChatGPT like Plugins and GPTs. The OpenAI Assistants have a dedicated API and a user interface, offering a distinct approach to creating these assistants.

The OpenAI assistants’ documentation clearly defines an AI assistant (OpenAI, 2024): “The Assistants API allows you to build AI Assistants within your own applications. It provides instructions and can leverage models, tools, and files to respond to user queries effectively.”

Unlike LLM-based chatbots, AI assistants have a different focus. They are designed to provide accurate, source-based information, with the ability to cite and link to those sources while maintaining LLMs’ natural language processing and reasoning abilities.

It is important to note that the quality of an assistant is not merely based on the volume of information baked into the LLM training or the cut-off date of its knowledge updates. Instead, it depends on the assistant’s ability to manage reasonably large contexts, retrieve relevant information and make sense of it, follow directions, and structure outputs. Therefore, the ideal LLM for building an AI assistant might not necessarily be the highest-performing model in all areas.

3.2.1. How to create a learning assistant

While LLMs play a crucial role in AI assistant development, they are not the only emerging technology involved. As outlined in Table 1, a range of other technologies and disciplines contribute to creating a comprehensive AI assistant. These complementary components augment the LLM’s capabilities, adding new features, predictability, and scope.

Table 1. Technologies involved in the creation of an AI assistant (in addition to LLMs)

Technology	Description
Retrieval-Augmented Generation (RAG)	RAG combines LLM's ability to generate responses and pull information from external databases or documents, improving the accuracy and relevance of responses. The retrieved data is inserted in the conversation with the LLM, usually called Context, so the LLM can use it to generate an accurate response
Semantic search in embeddings databases	Utilizes embeddings to organize and retrieve information semantically, improving the assistant's ability to understand and respond to queries accurately across different modalities. An embedding is a numerical representation of data that captures its meaning and relationships in a high-dimensional space. An embedding database can perform a similarity search and retrieve semantic-related objects to a given query. With RAG embeddings, databases confer an AI assistant capable of effectively using large amounts of unstructured information
Very large contexts	Modern LLMs have started to allow for extensive contexts. The context is the window of attention of LLM to a conversation. LLMs have displayed the emerging ability to learn new skills from the information and examples provided in a conversation. An LLM that can reliably attend to a very large context - see the needle in a haystack test (https://d66z.short.gy/ZumHZu) - will be most suitable for RAG strategies, requiring less precision in the retrieval strategy. However, the use of a lot of tokens in a context will come at higher computational costs
Code interpreters	The LLMs are not designed to perform calculations or complex tasks. While they can fake it with reasonable accuracy, especially the larger models, they are prone to error and hallucination. However, LLMs are increasingly proficient at generating code that can be passed to an interpreter, and they then use the execution output to complete their response
Function calling	Function calling is a feature introduced by OpenAI in June 2023. It provides the LLM with the option to respond with an invocation to the function of an API defined in the context. Function calling enables the LLM to interact with external information systems based on user commands
Prompt engineering	Prompt engineering (Sahoo et al., 2024) refers to crafting and optimizing prompts to guide the LLM responses, improving the quality and relevance of generated content
Evaluation	Involves assessing the assistant's performance using various monitoring tools, metrics, and benchmarks to ensure accuracy, reliability, and overall effectiveness. The results can be used to improve the prompt engineering and RAG processes and create datasets to fine-tune the underlying LLM further

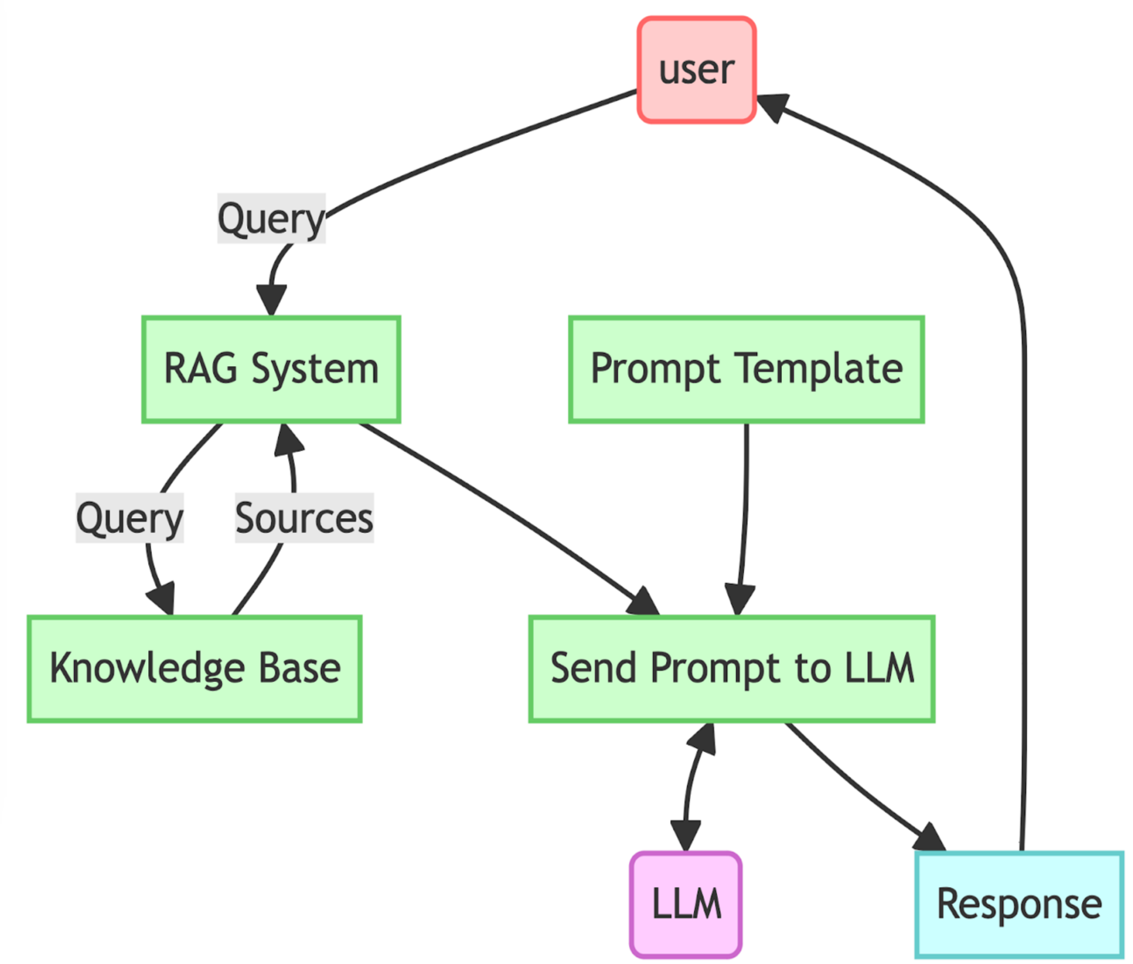
Fine tuning	Refers to retraining the underlying LLM on specific datasets to enhance its performance in particular domains or tasks

Beyond the technologies listed in Table 1, assistants demand strong software engineering skills and practices, focusing on deployment, scalability, and security. Ensuring the security of an AI assistant requires a deep understanding of information security principles and addressing unique concerns that arise from incorporating a LLM into the technology stack. Particular attention must be paid to mitigating risks such as prompt injection and LLM jailbreaking, among other potential threats (Yao et al., 2024).

3.2.2. The design of a basic safe learning assistant

Let us examine how a basic AI assistant using RAG techniques and prompt engineering aligns with the SAIE principles. We will use the simple AI assistant schema from Figure 1 as our reference point.

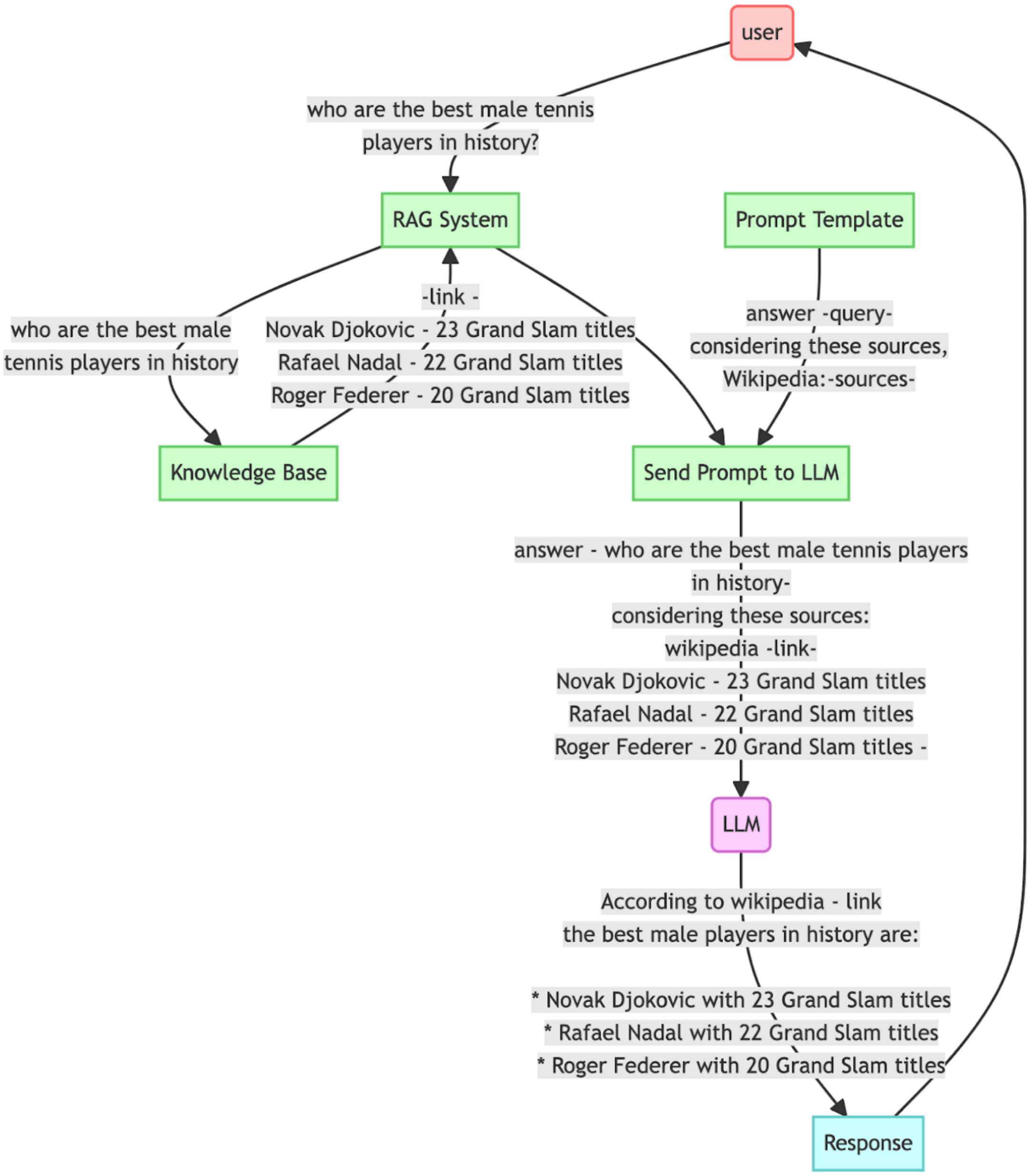
Figure 1. Schema of a simple AI Assistant



The assistant acts like a chatbot. When a user asks a question, the system finds relevant information from its database, selecting texts that are reliable sources. Combined with the user's question, these texts are sent to an LLM, which could be run online (like OpenAI's GPTs or Google's models) or on the organization's systems. The LLM processes this information and returns a response to the user.

Let us break it down into an example, as showcased in Figure 2.

Figure 2. Example of how a simple AI Assistant works


user
who are the best male tennis players in history?
RAG System
Prompt Template
who are the best male tennis players in history
-link -
Novak Djokovic - 23 Grand Slam titles
Rafael Nadal - 22 Grand Slam titles
Roger Federer - 20 Grand Slam titles
answer -query-
considering these sources,
Wikipedia:-sources-
Knowledge Base
Send Prompt to LLM
answer - who are the best male tennis players in history-
considering these sources:
wikipedia -link-
Novak Djokovic - 23 Grand Slam titles
Rafael Nadal - 22 Grand Slam titles
Roger Federer - 20 Grand Slam titles -
LLM
According to wikipedia - link
the best male players in history are:
* Novak Djokovic with 23 Grand Slam titles
* Rafael Nadal with 22 Grand Slam titles
* Roger Federer with 20 Grand Slam titles
Response

The assistant receives a "query" from the user:

"Who are the best male tennis players in history."

This query will be taken from the RAG subsystem of the agent, which will use the user query to select several texts from its knowledge base. In this case, let us say the knowledge base is

Wikipedia, and we get a link to ATP's Wikipedia page plus an excerpt listing the three best players and their grand slam wins. This text and the link will be considered valid sources of information. The sources of information are combined with the query of the user, according to a template, to create a message for the LLM:

"Answer the question: - who are the best male tennis players in history-

considering these sources:

Wikipedia -link-

Novak Djokovic - 23 Grand Slam titles

Rafael Nadal - 22 Grand Slam titles

Roger Federer - 20 Grand Slam titles -

"

The LLM will generate a response like the following one, which will be passed to the user. The response is an actual response from GPT-4o with the message proposed.

"The best male tennis players in history, based on Grand Slam titles and widely acknowledged by sources like Wikipedia, include:

1. **Novak Djokovic** - 23 Grand Slam titles
2. **Rafael Nadal** - 22 Grand Slam titles
3. **Roger Federer** - 20 Grand Slam titles

These three players have dominated the sport over the past two decades, setting numerous records and raising the level of competition in men's tennis."

3.2.3. AI assistant as safe AI in education

Now let us analyze this simple assistant according to the SAIE principles:

- The LLM is used as an API call using the assistant's code. This means that unless the user chooses to include personal information in the query, the confidentiality of the user is guaranteed. This satisfies (SAIE1); it guarantees confidentiality.

- The RAG system uses a knowledge base that the teacher/educational institution curates. This provides alignment with the educational institution's quality standards, views about the subject, and values in compliance with SAIE2.

- The prompt template will determine the LLM's behavior, not the user's query. This satisfies (SAIE3); it is aligned with didactic practices and helps with (SAIE5), presenting a comprehensive interface and behavior.

- According to the indications of the prompt template "Answer this question - <question> according to these sources - <sources>," the LLM is going to use the authoritative information provided by the knowledge base, Wikipedia, in the example. The LLM will base its answers on the sources provided, so the accuracy of the response will depend on the quality of the sources retrieved, not the training of the model or its cut-off knowledge date. This satisfies again (SAIE3); it is aligned with didactic practices and takes care of the principle (SAIE4), the accuracy and minimization of errors.

- The educational institution and the teacher design and control the assistant by selecting and curating the knowledge base and crafting the prompts to rule the assistant's behavior. The assistant expands and complements the teachers' and institution's capabilities and functions, and it is not designed to replace them in full compliance with the SAIE6 principle.

- While the usage of an assistant does not directly satisfy SAIE7 principle because it uses backend LLM-based technology via API, and we will not have control over the training of the model, the assistant being used provides itself a dataset of interactions - questions and answers - that can be later analyzed, verified and corrected to create a fine-tuning dataset that the educational institution can use to customize LLMs in the future. It is important here that the future dataset is in the hands of the educational institution.

So, the previous analysis suggests that an AI assistant tailored for specific didactic practices can satisfy SAIE1, 2, 3, 4, 5, 6, and, indirectly, 7. However, to satisfy SAIE2 fully (alignment with educational strategies), the AI assistant must behave as a smart learning application. This can be accomplished by using the IMS LTI interoperability protocol (IMS-GLC, 2014).

IMS LTI stands for Instructional Management System Learning Tools Interoperability. It is a standard developed by the IMS Global Learning Consortium that allows different LMS to integrate seamlessly with external educational tools and content. This means that tools like quizzes, assignments, and other learning resources from various platforms can work together within a single LMS, providing a smoother experience for both instructors and students.

4. ENSURING SAFE AI IN EDUCATION: A CHECKLIST

The Safe AI Manifesto outlines seven principles to ensure AI technologies' safe and ethical use in education. However, it is not intended to be a decision-making tool for determining whether to adopt a specific AI learning tool in an educational setting. For this reason, the authors have created a checklist to validate the safety of AI in education.

Although it may be challenging for an AI learning tool to meet all the criteria in the checklist, it offers decision-makers, educators, and developers a clear understanding of the potential benefits and risks. It also highlights areas needing improvement and aspects that should be approached cautiously.

The updated full version of the checklist can be accessed online at https://manifesto.safeaieducation.org/checklist. This initial version has been updated with the received comments of 20 international experts who have done a heuristic evaluation of the first checklist version.

The checklist is organized into three sections that specifically target different aspects of AI integration in education and might require different expert profiles and skill sets to provide accurate and comprehensive answers.

Here is a breakdown of the rationale for each section and the questions proposed.

1. About the AI learning tool, its data, and how it is processed

This section ensures the institution's control, compliance, and transparency regarding the AI educational tool's data handling and technology stack.

- **Ownership and control of the technology stack**. The questions examine whether the institution has full control over the AI tool and its infrastructure, which impacts data security, privacy, and compliance with laws like GDPR (Europe) and FERPA (USA). Whether the tool is on-premises, cloud-based, or SaaS (Software as a Service) dictates who is responsible for data protection, making these distinctions critical for ensuring compliance.
- **Ownership and control of data**. These questions explore the institution's ownership and control over sensitive personal data of students and teachers. They emphasize encryption, third-party sharing, and whether student interactions are handled securely. This is crucial for protecting privacy, particularly when dealing with third-party AI vendors. Understanding if and how data is shared and the rights institutions have over their own data is essential for protecting user privacy.
- **Research use of log data**. If the tool uses log data for research, transparency about what

CC BY NC ND

data is collected, why it is being used, and how it benefits education ensures ethical research practices. Obtaining permission from students and the institution ensures consent and ethical data usage.

- **Data minimization and deletion**. Collecting only necessary data and providing a way to delete data upon request securely adheres to privacy best practices. This ensures the institution is not unnecessarily collecting personal information and respects user rights over their own data.

2. About ethics and alignment with educational goals and practices

This section focuses on the ethical use of the AI tool, ensuring it complements educational goals and does not undermine the role of educators.

- **How does the AI tool fit in the educational context?** The questions ensure that the AI tool supports, rather than replaces, educators. By focusing on human supervision, explainability of AI decisions, and the ability for students to appeal, the aim is to prevent over-reliance on AI and maintain human oversight. This is vital for promoting fairness and ensuring that AI enhances, rather than disrupts, traditional educational practices.
- **Generated content disclosure and explainability**. These questions aim to ensure transparency about what content is generated by AI. This is important because students and teachers can trust the AI tool and understand the source of the information. Explainability is crucial for accountability, especially in educational environments where the accuracy and quality of content are critical.
- **Control and alignment of AI models**. Here, the focus is on ensuring AI models align with the institution's values and goals. Transparency over the AI models' training data and controlling output biases is critical to ethical AI use. Institutions should have some level of control or at least awareness of the training data used to ensure that AI models do not perpetuate biases or misinformation.

3. About the integration of the AI educational tool in the institution

This section discusses how the AI tool fits into the institution's overall technology and learning strategies.

- **Integration in the institution's technology strategy**. The focus is whether the AI tool can be easily integrated with existing systems, such as the LMS, and whether the institution can control user authentication, access, and scalability. This ensures that the tool is practical, secure, and can adapt to different institutional needs while being cost-effective.
- **Integration in the institution's learning and teaching strategies**. These questions assess how the AI tool fits the institution's educational goals. Identifying the risks and benefits ensures a careful approach to tool adoption. Training programs for teachers and strategies to support them are crucial to successfully incorporating the tool into teaching practices. The focus is on enhancing learning, identifying potential adverse impacts, like plagiarism or cheating, and addressing these proactively.
- **Integration in the institution's culture**. This section ensures that the institution communicates that the AI tool is there to enhance the work of human educators, not replace them. Training students and teachers on ethical AI use is essential to foster responsible use and a critical understanding of AI's capabilities and limitations.

In summary, the rationale behind the questionnaire is to ensure the ethical, compliant, and effective use of AI in education in compliance with the United Nations' Sustainable Development Goal 4 (Flores-Vivar & García-Peñalvo, 2023). The focus is on data protection, transparency, ethical decision-making, alignment with educational goals, and ensuring the tool supports—not replaces—human educators.

5. CONCLUSIONS

The seven principles of Safe AI in Education (SAIE) focus on ensuring that AI tools used in educational settings protect student data confidentiality (SAIE1), align with the institution’s educational strategies and information technology governance (SAIE2), conform to didactic practices and instructional design (SAIE3), prioritize accuracy and minimize errors, particularly avoiding misleading information (SAIE4), present a comprehensive, user-friendly interface that communicates the tool’s purpose and limitations to both students and teachers (SAIE5), do not try to replace humans in the teaching and learning processes (SAIE6). Their training processes have been defined under common ethical principles and in a transparent way. These principles aim to balance the benefits of AI in education with the need for security, accuracy, and pedagogical alignment.

The SLApp principles emphasize how AI applications should be integrated and managed within the institution’s technological ecosystem, basically with its LMS. SLApp1 ensures secure access by requiring the educational institution to handle authentication and authorization through the LMS. SLApp2 highlights the need for applications to adapt their features based on the user’s role, such as teacher or student. SLApp3 focuses on providing course-specific settings, allowing each application instance to be tailored to the specific course and learning activity, reflected in the LMS interface. SLApp4 supports using APIs to leverage advanced AI models, enabling features like personalized learning and on-the-fly content creation and ensuring that the latest AI capabilities can be utilized within the educational infrastructure.

Our analysis concludes that an LLM alone is not a good candidate to be a Safe AI in Education. However, using the emerging techniques of prompt engineering, RAG, function calling, structured outputs, etc., designing and building applications aligned with SAIE is possible. However, SAIE principles must be considered in the very design of such applications, and transparency should be required to guarantee SAIE. This transparency can only be provided by an open-source strategy or code that the learning institution can audit.

IMS LTI is the ideal candidate to transform SAIE applications into smart learning applications, allowing for the easy integration of the new crop of SAIE applications into most learning institutions' existing learning information systems strategies worldwide.

When selecting AI applications for education, managers and educators must be aware of the SAIE principles and include them in their selection criteria. To support the education community openly in the introduction of AI tools in education, the Safe AI in Education Manifesto (https://manifesto.safeaieducation.org/) and the Safe AI in Education checklist (https://manifesto.safeaieducation.org/checklist) have been defined and open to the community to achieve a natural evolution of these principles in compliance with the unavoidable advance of the AI technologies.

As the regulatory landscape surrounding AI in education continues to evolve, institutions need to engage actively in shaping these frameworks. Ensuring that AI adoption complies with local, national, and international regulations on data protection and ethical use is crucial to maintaining trust among students, educators, and stakeholders. Institutions should also consider the long-term sustainability of AI tools, promoting fairness and inclusivity to prevent deepening existing educational inequalities. By aligning the SAIE principles with these regulatory frameworks, academic institutions can create a safer and more effective environment for integrating AI technologies into education.

Declaration of competing interest

The authors have no competing interests to disclose.

Authors and Affiliations

Francisco José Garcia-Peñalvo Research Institute for Educational Sciences (IUCE), Universidad de Salamanca, Salamanca, Spain.
Marc Alier and Maria Jose Casany UPC Universitat Politècnica de Catalunya - Barcelona Tech, Barcelona, Spain.
Juanan Pereira Computer Science Faculty, UPV/EHU, San Sebastián, Spain.

Authors' contribution

Francisco José García Peñalvo: supervision; conceptualization; writing -original draft preparation; reviewing & editing. Marc Alier: conceptualization; data collection; writing -original draft preparation; visualization. Juanan Pereira: conceptualization; data collection; reviewing & editing; Maria Jose Casany: supervision; conceptualization; reviewing & editing.

Corresponding author

Correspondence to Francisco José Garcia-Peñalvo.

Funding

The Ministry of Science and Innovation **partially funded this research** through the AvisSA project grant number (PID2020-118345RB-I00).

REFERENCES

Alkuş, S., & Olgan, R. (2014). Pre-service and in-service preschool teachers' views regarding creativity in early childhood education. *Early Child Development and Care, 184*(12), 1902–1919. https://doi.org/10.1080/03004430.2014.893236

Altay, B. (2014). User-centred design through learner-centred instruction.*Teaching in Higher Education, 19*(2), 138–155. https://doi.org/10.1080/13562517.2013.827646

Beery, T. A., Shell, D., Gillespie, G., & Werdman, E. (2013). The impact of learning space on teaching behaviors. Nurse Education in Practice, 13(5), 382-387. https://doi.org/10.1016/j.nepr.2012.11.001

Bezanilla, M. J., Fernández-Nogueira, D., Poblete, M., & Galindo-Domínguez, H. (2019). Methodologies for teaching-learning critical thinking in higher education: The teacher's view. *Thinking skills and creativity, 33,* 100584. https://doi.org/10.1016/j.tsc.2019.100584

Biggs, J., & Tang, C. (2011). *Teaching for Quality Learning at University: What the Student Does*. 4th ed. McGraw-Hill.

Blannin, J., Mahat, M., Cleveland, B., Morris, J. & Imms, W. (2020). Teachers as Embedded Practitioner-Researchers in Innovative Learning Environments. *Center for Educational Policy Studies Journal, 10*(3), 99-116. https://doi.org/10.26529/cepsj.887

Byers, T., Imms, W., & Hartnell-Young, E. (2014). Making the case for space: The effect of learning spaces on teaching and learning. *Curriculum and Teaching, 29*(1), 5–19. https://doi.org/10.7459/ct/29.1.02

Byers, T., Imms, W., & Hartnell-Young, E. (2018a). Comparative analysis of the impact of traditional versus innovative learning environment on student attitudes and learning outcomes. *Studies in Educational Evaluation, 58*, 167–177. https://doi.org/10.1016/j.stueduc.2018.07.003

Byers, T., Imms, W., & Hartnell-Young, E. (2018b). Evaluating teacher and student spatial transition from a traditional classroom to an innovative learning environment. *Studies in Educational Evaluation, 58*, 156–166. https://doi.org/10.1016/j.stueduc.2018.07.004

Cleveland, B. (2016). Addressing the spatial to catalyse socio-pedagogical reform in middle years education. In K. Fisher (Ed.), *The translational design of schools* (pp. 27–49). Sense Publishers. https://doi.org/10.1007/978-94-6300-364-3_2

Consejo Europeo (2006). Recomendación del parlamento europeo y del consejo de 18 de diciembre de 2006 sobre las competencias clave para el aprendizaje permanente. Diario Oficial de Europa, L394/10, de 30 de diciembre de 2016. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:394:0010:0018:es:PDF

Cornelius-White, J. (2007). Learner-Centered Teacher-Student Relationships Are Effective: A Meta-Analysis. *Review of Educational Research, 77*(1), 113–143. https://doi.org/10.3102/003465430298563

Da Silva Santos, L. R., Souza-Pinho, M. J., Santos de Jesus, M., & Kalil, A. (2022). Educational games in the high school: implicate future teachers in the pursuit for new teaching strategies. *IJERI: International Journal of Educational Research and Innovation*, (17), 27–44. https://doi.org/10.46661/ijeri.4574

European Schoolnet (27 april 2024). *Future Classroom Lab*. https://fcl.eun.org/

European Schoolnet (29 may 2024). *iTEC*. http://itec.eun.org/web/guest

European Union. (2013). *Improving the quality of in-service teacher training system:analysis of the existing ETTA INSETT system and assessment of the needs for inservice training of teachers* (No. EuropeAid/130730/D/SER/HR). European Union.

Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., & Wenderoth, M.P. (2014). Active Learning Increases StudentPerformance in Science, Engineering, and Mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

French, R., Mahat, M., Kvan, T. & Imms, W. (2022). Viewing the transition to innovative learning environments through the lens of the burke-litwin model for organizational performance and change. *Journal of Educational Change, 23*, 115-130. https://doi.org/10.1007/s10833-021-09431-5

Generalitat Valenciana, (10 march 2024). Aules Transformadores D´espais i Metodologies Educatives. https://portal.edu.gva.es/aulestransformadores/es/inicio/

Gonzalez-Mohino, M., Rodriguez-Domenech, M., Callejas-Albiñana, A., & Castillo-Canalejo, A. (2023). Empowering Critical Thinking: The Role of Digital Tools in Citizen Participation. *Journal of New Approaches in Educational Research, 12*(2), 258-275. https://doi.org/10.7821/naer.2023.7.1385

Granda-Pinan, A.R., Lozano, O. R. & Alameda Villarrubia, A. (2024). Training on Innovative Learning Environments: Identifying Teachers' Interests. *Educational Sciences*, in press.

Granda-Pinan, A.R., Roda-Segarra, J. & Mengual-Andrés, S. (2024). Espacios digitales de aprendizaje como prolongación del espacio físico en el contexto de los Entornos Innovadores de Aprendizaje. En Morales Cevallos, M. B., Marín-Marín, J. A., Berbel Oller, P. y Villegas Castro, A. S. (Coords.), *Desafíos de la educación contemporánea: perspectivas formativas para una sociedad digital* (pp. 86-103). Dykinson.

Granda-Piñán, A.R. & Rojo-Bofill, L.M. (2024). Los Entornos Innovadores de Aprendizaje como respuesta a los retos educativos del siglo XXI. *Research in Education and Learning Innovation Archives*, 32,22-35. https://doi.org/10.7203/realia.32.27803

Hong, W. P. (2012). An international study of the changing nature and role of school curricula: from transmitting content knowledge to developing students' key competencies. *Asia Pacific Education Review, 13*(1), 27–37. http://dx.doi.org/10.1007/s12564-011-9171-z

Hoskins, B., & Deakin, R. (2010). Competences for Learning to Learn and Active Citizenship: different currencies or two sides of the same coin? *European Journal of Education, 45*(1), 121-137. http://dx.doi.org/10.1111/j.1465- 3435.2009.01419.x

Imms, W., Morris, J., Bradbeer, C., & Mahat, M. (2023). *What should be the focus of next generation learning spaces research? An international cross-sector response White Paper from the Innovative Learning Environments and Student Experience Scoping Study.* https://ilesescopingstudy.com.au/

INTEF. (4 march 2024). *Aula del Futuro*. https://auladelfuturo.intef.es/

Jiménez-Hernández, D., Sancho-Requena, P., & Sánchez-Fuentes, S. (2021). Estudio acerca de las opiniones del profesorado universitario en la Región de Murcia sobre la formación de métodos activos. *Revista Electrónica Interuniversitaria de Formación del Profesorado, 24(2)*. https://doi.org/10.6018/reifop.444381

Jorion, N., Taeyaerts, D., & Jeanes, W. (2016). Promoting collaboration using team based classroom design. *Creative Education, 07*(05), 724–729. https://doi.org/10.4236/ce.2016.75076

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Losada-López, J.L. y López-Feal Ramil, R. (2003). *Métodos de investigación en ciencias humanas y sociales.* Thomson-Paraninfo.

Martín-Gutiérrez, Á., Said-Hung, E. & Conde-Jiménez, J. (2024). Social media and non-university teachers from a gender perspective in Spain. *Journal of New Approaches in Educational Research,* 13 (10). https://doi.org/10.1007/s44322-024-00010-z

Mcmillan, J.H. & Schumacher, S. (2005). *Investigación educativa. Una introducción conceptual*. Pearson Educación.

Michael, J. (2006). Where's the Evidence That Active Learning Works?. *Advances in Physiology Education, 30*(4), 159–167. https://doi.org/10.1152/advan.00053.2006

Nzarirwehi, J. & Atuhumuze, F. (2019). In-Service Teacher Training and Professional Development of Primary School Teachers in Uganda. *IAFOR Journal of Education, 7*(1), 19-36. https://doi.org/10.22492/ije.7.1.02

Osamwonyi, E. F. (2016). In-service education of teachers: Overview, problems and the way forward. *Journal of Education and Practice, 7*(26), 83–87.

Pelletier, K., Robert, J., Muscanell, N., McCormack, M., Reeves, J., Arbino, N., Grajek, S., Birdwell, T., Liu, D., Mandernach, J., Moore, A., Porcaro, A., Rutledge, R., & Zimmern, J. (2023). 2023 EDUCAUSE Horizon Report, Teaching and Learning Edition. https://library.educause.edu/-/media/files/library/2023/4/2023hrteachinglearning.pdf?#page=36

Prince, M. (2004). Does Active Learning Work? A Review of the Research. *Journal of Engineering Education, 93*(3), 223–231. https://doi.org/10.1037/xap0000470

Rivera-Vargas, P., Calderón-Garrido, D., Jacovkis, J. & Parcerisa, L. (2024). Exploring student and family concerns and confidence in BigTech digital platforms in public schools. *Journal of New Approaches in Education Research 13* (5). https://doi.org/10.1007/s44322-023-00003-4

Ruiz-Primo, M.A., Briggs, D., Iverson, H., Talbot, R. & Shepard, L.A. (2011). Impact of Undergraduate Science Course Innovations on Learning. *Science, 331*(6022), 1269–1270. https://doi.org/10.1126/science.1198976

Sheth, S. (2004). Knowledge for teacher development in India: The importance of local knowledge for in-service education. *International Journal of Education Development, 24*, 39–52. https://doi.org/10.1016/j.ijedudev.2003.09.003

Stes, A., Min-Leliveld, M., Gijbels, D. & Van Petegem, P. (2010). The Impact of Instructional Development in Higher Education: The State-of-the-Art of the Research. *Educational Research Review, 5*(1), 25–49. https://doi.org/10.1016/j.edurev.2009.07.001

Taveras-Pichardo, L. (2022). Cognitive and affective variables predictive of the academic performance of university students. *IJERI: International Journal of Educational Research and Innovation,* (18), 118–131. https://doi.org/10.46661/ijeri.6189

Tristán-López, A. (2008). Modificación al modelo de lawshe para el dictamen cuantitativo de la validez de contenido de un instrumento objetivo. *Avances en medición, 6*(1), 37-48.

Wilson, M. E. (2004). Teaching, learning, and millennial students. *New Directions for Student Services,* (106), 59–71. https://doi.org/10.1002/ss.12