IJD INTERDISCIPLINARY JOURNAL OF DIDACTICS
ISSN 3045-5413

Núm. 1, 2024, pp. 33-54
https://doi.org/10.14198/ijd.28095 Artículo

Influence of Critical Thinking on LLM Usage Among Universitat d'Andorra Students

Influencia del pensamiento crítico en el uso de los LLM entre los estudiantes de la Universitat d'Andorra

Marc Bleda Bejar, Aleix Dorca Josa, Begonya Oliveras Prat

Marc Bleda Bejar
Grup de Recerca en Tecnologia, Universitat d'Andorra, Andorra
mbleda@uda.ad
https://orcid.org/0009-0003-1703-2367

Aleix Dorca Josa
Grup de Recerca en Tecnologia, Universitat d'Andorra, Andorra
adorca@uda.ad
https://orcid.org/0000-0002-5749-8319

Begonya Oliveras Prat
Universitat Autònoma de Barcelona, España
Begona.Oliveras@uab.cat
https://orcid.org/0000-0003-1315-1126

Correo de correspondencia:
mbleda@uda.ad

Fecha recepción: 28/08/2024
Fecha aceptación: 20/09/2024

Financiación: este trabajo no ha recibido financiación.

Conflicto de intereses: los autores declaran que no hay conflicto de intereses.

Abstract:

As large language models continue to reshape educational practices, a comprehensive evaluation of critical thinking's influence on large language models' usage becomes essential. This study examines how students in the fields of education and computer science at the Universitat d'Andorra interact with large language models, with a particular focus on understanding their learning experiences, decision-making strategies, and problem-solving approaches. Using qualitative and quantitative methods, the research analyzes the frequency and purposes of using these technologies, as well as the critical thinking processes students employ to assess the reliability and relevance of content generated by artificial intelligence. Findings reveal a spectrum of attitudes towards large language models, ranging from enthusiastic adoption to skepticism. While many students appreciate the immediate and personalized academic support, content generation assistance, and writing skill improvement offered by these tools, concerns about the accuracy and potential biases of the outputs are prevalent. Notably, students demonstrate varying levels of the activation of their critical thinking skills when engaging with large language models, with some actively investigate the reliability of artificial intelligence generated information, while others exhibit a more passive reliance on these technologies.
The study also highlights distinct usage patterns between computer science and education students. The results contribute to a deeper understanding of student behavior in the context of artificial intelligence enhanced education, providing valuable insights for educational institutions aiming to integrate these tools into their curricula effectively. Furthermore, this research emphasizes the need to enhance critical thinking skills within educational programs to empower students to navigate the complexities of large language models capabilities and limitations.

Citación: Bleda Bejar, M., Dorca Josa, A., Oliveras Prat, B. (2024). Influence of Critical Thinking on LLM Usage Among Universitat d'Andorra Students. *Interdisciplinary Journal of Didactics,* (1), 33-54. https://doi.org/10.14198/ijd.28095

Keywords: Artificial intelligence; large language models; critical thinking; higher education.

Resumen:

A medida que los modelos extensos de lenguaje continúan transformando las prácticas educativas, se vuelve esencial una evaluación exhaustiva de la influencia del pensamiento crítico en su uso. Este estudio examina cómo los estudiantes en los campos de la educación y la informática de la Universitat d'Andorra interactúan con los modelos extensos de lenguaje, centrándose especialmente en comprender sus experiencias de aprendizaje, estrategias de toma de decisiones y enfoques para resolver problemas. Utilizando métodos cualitativos y cuantitativos, la investigación analiza la frecuencia y los propósitos del uso de estas tecnologías, así como los procesos de pensamiento crítico que emplean los estudiantes para evaluar la fiabilidad y la relevancia del contenido generado por la inteligencia artificial.

Los hallazgos revelan una gama de actitudes hacia los modelos extensos de lenguaje, que van desde la adopción entusiasta hasta el escepticismo. Si bien muchos estudiantes aprecian el apoyo académico inmediato y personalizado, la ayuda en la generación de contenido y la mejora de las habilidades de escritura que ofrecen estas herramientas, preocupan la precisión y los posibles sesgos de las salidas. En particular, los estudiantes demuestran niveles variables de activación de sus habilidades de pensamiento crítico cuando interactúan con los modelos extensos de lenguaje, algunos investigan activamente la confiabilidad de la información generada por la inteligencia artificial, mientras que otros exhiben una dependencia más pasiva de estas tecnologías.

El estudio también destaca patrones de uso distintivos entre estudiantes de informática y educación. Los resultados contribuyen a una comprensión más profunda del comportamiento de los estudiantes en el contexto de la educación mejorada por la inteligencia artificial, proporcionando información valiosa para las instituciones educativas que buscan integrar estas herramientas en sus planes de estudios de manera efectiva. Además, esta investigación enfatiza la necesidad de mejorar las habilidades de pensamiento crítico dentro de los programas educativos para empoderar a los estudiantes a navegar por las complejidades de las capacidades y limitaciones de los modelos de lenguaje masivo.

Palabras clave: inteligencia artificial; modelos extensos de lenguaje; pensamiento crítico; educación superior.

1. INTRODUCTION

Considering that Large Language Models (LLMs) are designed to generate coherent content based on probabilistic patterns rather than factual accuracy, the activation of Critical Thinking (CT) is crucial when using these tools. This article seeks to examine how higher education students engage with LLMs, their perceptions of the reliability of the information produced, and the strategies they employ to validate this information.

1.1. CRITICAL THINKING

A key focus of this research is CT, and to understand its essence, it is crucial to examine various definitions. Bloom et al. (1956) characterized CT as the mastery of skills including knowledge, comprehension, application, analysis, synthesis, and evaluation. Among these, the higher-order skills of analysis, synthesis, and evaluation are often considered central to CT. Paul and Elder (2005, p. 7) describe CT as "the process of analyzing and evaluating thinking with the purpose of improving it". Enhancing the quality of thinking, according to them, involves cultivating a set of essential conditions vital for developing CT. They emphasize the need to foster key thinking habits that involves all activities, such as reading, writing, speaking, and listening, applicable in both academic contexts and personal and professional life. In their point of view, CT skills are categorized into two types: general skills (relevant across all disciplines) and specific skills (pertinent to a particular discipline).

In 1990, the American philosophical association convened a panel of experts to establish a consensus definition of CT. The association's report identified two essential components of a critical thinker: cognitive skills and affective dispositions (Facione, 1990). For the cognitive aspect, they outlined six key skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation. However, there was some disagreement among experts regarding whether dispositions are integral to the definition of CT or simply qualities that enhance the practice of CT. Facione (1990) distinguished between cognitive skills, which are necessary for acquiring knowledge, and dispositions, which he described as virtues that guide the application of this knowledge in daily life. Regardless of this debate, the report emphasizes that dispositions add a personal and civic dimension to the critical thinker. This distinction between cognitive skills and dispositions has been adopted by other scholars (Paul & Elder, 2005; Ennis, 1987; Halpern, 1998; Dunn et al., 2009).

CT is closely tied to each person's value system and is influenced by the context (Izquierdo & Aliberas, 2021). Values and emotions are crucial in shaping judgments and making decisions (Tura et al., 2023). The integration of skills, dispositions, and values fosters the metacognitive dimension of a critical thinker, which encompasses self-awareness of one's knowledge, regulation, control, and organization of strategies and metacognitive abilities (Allueva, 2002). In a critical thinker, skills, dispositions, values, emotions, and metacognitive abilities are interconnected. Thus, this article aligns with the definition proposed by the Language and Science Teaching research group (LIEC) at the Universitat Autònoma de Barcelona (UAB), a definition crafted specifically for the educational field. CT is defined as "that set of cognitive, metacognitive, attitudinal and emotional processes that, while being based on knowledge of science, about science and values, allow to participate successfully in the evaluation of knowledge and ways of knowing school science through the application of criteria" (Couso & Márquez, 2023).

1.2. ARTIFICIAL INTELLIGENCE

1.2.1. Evolution of Artificial Intelligence

The concept of Artificial Intelligence (AI) is not a recent development; it has been a topic of discussion for many years, with debates about its definition emerging over four decades ago. AI refers to the simulation of human intelligence in machines, enabling them to think and learn in ways similar to humans. This involves creating algorithms and computational models that allow computers to perform tasks typically requiring human intelligence (Kurzweil et al., 1990). Such tasks include problem-solving, speech recognition, learning, planning, and perception (Bellman, 1978).

Today, AI spans various fields, including Natural Language Processing (NLP), machine learning, and speech recognition. NLP, a branch of AI, focuses on the interaction between computers and human language, aiming to enable machines to comprehend, interpret, and generate human language in a meaningful and contextually relevant way. Key elements of NLP include text processing, language understanding, and language generation, with applications in sentiment analysis, speech recognition, question answering, and dialogue systems. Despite advancements, NLP still faces challenges such as handling ambiguity, understanding context, and addressing language diversity. Ethical concerns, particularly regarding bias in language models, remain significant in NLP research.

The primary goal of a language model is to capture the inherent structure and patterns of natural language, enabling it to predict the likelihood of a word or sequence of words based on its context. The term "probabilistic" indicates that these models are grounded in probability theory (Bengio et al., 2000). In 2017, a group of researchers introduced a new neural network architecture called the Transformer, designed for sequence transduction tasks. This architecture relies exclusively on attention mechanisms, eliminating the need for recurrent or convolutional layers (Vaswani et al., 2017). The effectiveness of the transformer is evident in automatic translation tasks, where it has achieved outstanding results on various benchmark datasets, reduced training time, and improved parallelization. The authors also conducted experiments to explore the significance of different components of the Transformer and offered insights into its internal workings. For example, ChatGPT is largely based on this architecture (OpenAI, 2022). Recent progress, especially with models like GPT-3, has significantly enhanced NLP's capabilities, pushing the limits of what machines can achieve in language-related tasks.

1.2.2. Benefits of Artificial Intelligence

AI models offer significant benefits across various fields beyond text generation. In music composition, they generate accurate structures for melodies in genres with strict rules, though their accuracy drops for more complex genres like jazz (Alaeddine & Tannoury, 2021). In image editing, pre-trained GANs provide efficient restoration and customization, saving time and resources (Liu et al., 2023). Additionally, DATID-3D enhances 3D generative models' domain adaptation, improving image quality and text-image correspondence for virtual reality, gaming, and product design (Kim & Chun, 2023). ChatGPT performs well in translating high-resource European languages but struggles with low-resource or linguistically distant languages.

However, with the introduction of the GPT-4 engine, ChatGPT's translation capabilities have significantly improved, bringing its performance closer to that of commercial products, even for more distant languages (Jiao et al., 2023). ChatGPT also shows promise in supporting English as a Foreign Language (EFL) writing by offering personalized feedback, boosting engagement, and improving writing quality. While these benefits are notable, there are concerns that excessive reliance on AI in EFL writing could diminish human interaction and creativity. Therefore, a balanced approach is advised, using ChatGPT to complement and enhance, rather than replace, traditional teaching and learning methods (Ningrum et al., 2023). LLMs like ChatGPT have notable strengths, including their ability to generate high-quality text that closely mimics human writing, produce content in various styles and languages, and automate content creation across multiple industries. However, these models also face limitations, such as a tendency to generate biased or offensive material, challenges in producing coherent long-form texts, limited control over the output, and the high computational costs associated with large-scale Artificial Intelligence Generated-Content (AIGC) models. Although significant advancements have been made, continuous efforts are needed to further enhance the quality and diversity of AIGC (Zhang et al., 2023).

1.2.3. LLMs in education

AI is currently being used in education in various ways, including personalized learning, intelligent tutoring systems, learning analytics, assessment, and grading. Global organizations like UNESCO have advocated for incorporating AI and LLMs in education, though ethical concerns, such as student plagiarism, highlight the need for further research to ensure the ethical and effective use of tools like ChatGPT (Sabzalieva & Valentini, 2023). The potential benefits of AI in education include improved learning outcomes, increased efficiency and productivity, and greater access to education for marginalized or underserved populations. However, challenges remain, such as concerns about data privacy and security, the risk of bias or discrimination in AI algorithms, and the potential displacement of teachers and other education professionals. It is crucial to ensure that AI integration in education aligns with principles of human rights and social justice. To achieve this, involving educational authorities and coordinating collective efforts to promote AI usage with a focus on societal improvement is recommended (UNESCO, 2019). The growing use of AI in personalized learning, analytics, administration, and research support in the coming years is expected to be beneficial. However, a comprehensive examination of the ethical and societal implications is essential, requiring a collaborative, interdisciplinary approach involving educators, IT professionals, policymakers, and other stakeholders (Liu et al., 2023).

Generative Artificial Intelligence (GAI) has the potential to transform education by enhancing learning through personalized experiences, fostering collaboration, and improving assessment methods. It offers significant benefits for both instructors and students, providing capabilities such as generating course materials, offering suggestions, performing linguistic translations, creating assessment tasks, and evaluating student performance. However, it's important to recognize the associated risks, including privacy concerns and bias, which highlight the need for careful ethical consideration in the application of these tools (Solís et al., 2023; González et al., 2023). While LLMs show promise in increasing student engagement and creating interactive materials, their responsible use is crucial to prevent bias and ensure fairness.

The integration of CT and problem-solving skills in education is essential (George-Reyes et al., 2023). These models should be used to complement and enhance the learning experience rather than replace traditional teaching methods (Kasneci et al., 2023). ChatGPT, for example, can generate accurate and well-structured responses to university-level questions, raising concerns about potential academic misconduct. It also has the capability to produce complex CT questions and assess responses across various disciplines (Susnjack, 2022). Additionally, it can work as a virtual tutor, helping students with their doubts, facilitating collaboration, and generating dialogues to support language learning. Despite these advantages, there are concerns about the accuracy, reliability, and potential biases in ChatGPT's responses, as well as its potential to perpetuate inequalities. On the positive side, its performance across different domains has shown excellent results in CT, higher-order thinking, and economics, according to studies conducted in higher education settings across various countries (Lo, 2023). ChatGPT can also generate basic lesson plans, offering a flexible framework that teachers can adapt to meet their specific needs and context. Promoting CT and fostering openness in teacher education are essential for adapting to the evolving role of technology and its impact on pedagogy (Berg & Plessis, 2023). Furthermore, ChatGPT can enhance personalized and interactive learning experiences by encouraging collaboration among policymakers, researchers, educators, and technology experts to harness GAI tools for positive educational outcomes (Baidoo-Anu & Ansah, 2023).

Users generally perceive human-generated content and AIGC as equally credible, though they find AIGC to be clearer and more engaging than content produced by humans. Educating the public about LLMs is crucial for helping people understand and evaluate the potential risks associated with these tools. Promoting the responsible use of AIGC involves encouraging prudence, CT, and media literacy (Huschens et al., 2023). Users are advised to critically assess information sources and exercise caution, even when the content appears to come from a reliable source. Teachers, for their part, view these tools favorably, noting benefits such as well-structured information, personalized feedback, and enhanced CT (Baskara et al., 2023). From the students' perspective, the key strengths of AI tools in education include their ability to improve learning practices, personalize educational experiences, and provide immediate assistance, which in turn enhances their overall learning experience and engagement. However, students also identified areas where AI tools could be improved, reflecting a mix of optimism and concern regarding the role of AI in their education (Irfan et al., 2023). While many students are aware of these tools, not all use them regularly for academic purposes. Nevertheless, AI tools are valued for their assistance in writing, virtual tutoring, research support, and automated grading (Singh et al., 2023).

1.3. OBJECTIVES

These tools introduce new challenges in education, highlighting the need for higher education students to apply CT. This research aims to understand how students in higher education perceive and use LLMs, focusing on their attitudes toward these emerging technologies. Specifically, it is important to examine how students engage their CT skills when interacting with LLMs. The study will investigate the use of LLMs among students in Computer Science (CS) and Education, comparing their usage patterns. It will also assess how students perceive the reliability of LLM responses and their strategies for validating the content when uncertainties arise.

The specific objectives of this article are:

Objective 1: To examine and evaluate whether there are significant differences in LLMs usage between CS and Education students.

Objective 2: To describe how CS and Education students assess the reliability of LLMs responses.

Objective 3: To analyze the strategies that students use to validate the answers provided by LLMs.

2. METHODOLOGY

In November 2023, a survey combining quantitative and qualitative questions was conducted at the Universitat d'Andorra (UdA). Each question of the survey underwent a rigorous validation process to assess the degree of univocity, pertinence and importance, as detailed in Carrera et al. (2011), with input from five experts in various fields. Some questions were specifically aimed at gauging students' familiarity with these tools. The instrument included a mix of Likert scale questions, ranking questions, multiple-choice questions, and open-ended questions to capture a comprehensive view of students' experiences and opinions.

The survey was administered to all the students from the BSc programs in CS and Education at the UdA. Of the 129 students surveyed, 83 responded, resulting in a margin of error of 6% with a 95% confidence level. The distribution of the surveyed students is detailed in Table 1.

The survey was conducted through multiple in-person sessions across all three years of study in both fields. Participants completed the survey anonymously to ensure that their responses were honest and genuine. The collected data was compiled into a comma-separated values (CSV) file for analysis and interpretation. The questionnaire consisted of 15 questions designed to explore various aspects of students' perceptions and the usage of LLMs in an educational context. Three methods were employed to analyze the survey results.

Table 1

Distribution of the Users in the Dataset by Academic Year and Field

Academic year	Education	Computer Science	Total
1	26	13	39
2	14	9	23
3	16	5	21
Total	56	27	83

Source: own elaboration.

First, a descriptive analysis provided an overview of the responses from higher education students. Second, quantitative analyses with graphical representations were used for questions with closed options. Finally, responses to open-ended questions were analyzed using systemic networks, developed by Bliss et al. (1983). Systemic networks are tools that classify and categorize different meanings behind expressions or drawings. This approach, rooted in systemic linguistics, focuses on describing and representing the meanings conveyed by semantic resources in language. To address Objective 1, a quantitative analysis was performed on responses to a question that assessed the frequency of LLM usage in different contexts. This analysis aimed to identify significant differences between CS and Education students. Given that the question used a Likert scale with sufficiently large frequencies and independent data, a Pearson's χ^2 test of independence was applied to evaluate these differences. For Objective 2, a qualitative analysis was conducted using student responses to the Likert scale question "Responses provided by LLMs are reliable", along with their corresponding free-text explanations. These responses were compared with answers to a related sub-question, "I understand perfectly how LLMs work". A coding table developed from the data and a bottom-up coding approach were used to identify various codes and themes. To achieve Objective 3, a qualitative analysis of student responses to the open-ended question "How do you ensure the reliability of responses generated by LLMs when you have doubts about their content?" was conducted. Using the same coding methodology as in Objective 2, additional codes were identified to classify and categorize the students' strategies for validating LLM responses.

3. RESULTS

3.1. DESCRIPTIVE ANALYSIS

First, a descriptive analysis of the results has been conducted. The majority of the surveyed students are familiar with what a LLM is. Among the LLMs listed (BARD, Bing AI, ChatGPT, ChatPDF, LLaMa, and Perplexity), 98% of the students reported using ChatGPT. Bing AI was the second most popular choice, used by 48% of respondents. BARD followed at 20%, while ChatPDF and Perplexity each had 12% usage. LLaMa was the least used, with only 4% of the surveyed students reporting its use.

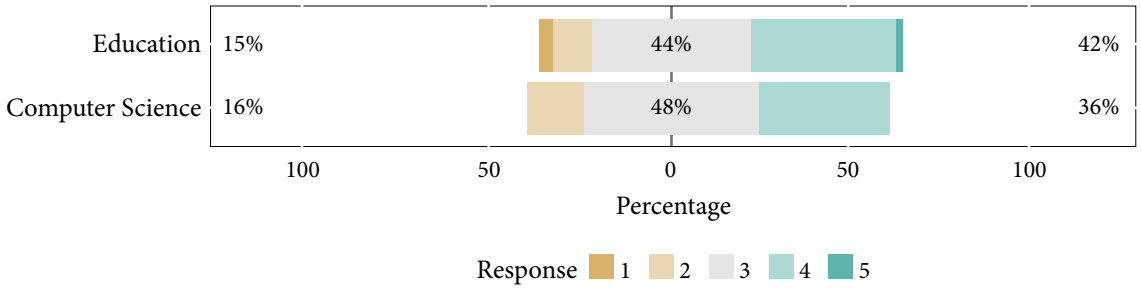
The frequency of students using LLMs over the past year was measured using a Likert scale. Among the respondents, 6% reported using LLMs "Never," while 22% indicated "Rarely". Some of these "Rarely" users find LLMs very useful for practical tasks like making shopping lists, planning menus, and creating gym routines, despite acknowledging limitations such as outdated databases. Although they have concerns about reliability, they appreciate LLMs' value in academic information research. Another 22% of students reported using LLMs "Occasionally" for tasks such as writing text and resolving programming issues. They often seek LLMs' suggestions to refine their work and gather general ideas after completing their tasks. A larger group, 29%, use LLMs "Often", valuing them for their time-saving benefits in information research and their role in providing inspiration, diverse perspectives, and project support in both academic and personal contexts. These users particularly rely on LLMs for navigating complex problems, correcting code errors, and conducting advanced or specialized research. Finally, 22% of the students use LLMs "Very often", integrating them extensively into

various aspects of their work. They find LLMs crucial for generating ideas and obtaining quick, reliable information for tasks such as research, writing, structuring, and creating multimedia content. Notably, CS students used these tools more frequently over the past year compared to their peers in Education.

Regarding the statement "Responses provided by LLMs are reliable", 3% of the surveyed students strongly disagreed and 13% disagreed. The most common response, selected by 45% of students, was "Neither agree nor disagree". This was followed by 39% who agreed with the statement, and just 1% who strongly agreed. Attitudes toward the reliability of LLM-generated responses were similar across both fields, as illustrated in Figure 1.

Figure 1

Reliability of LLM-Generated Responses by Fields of Study



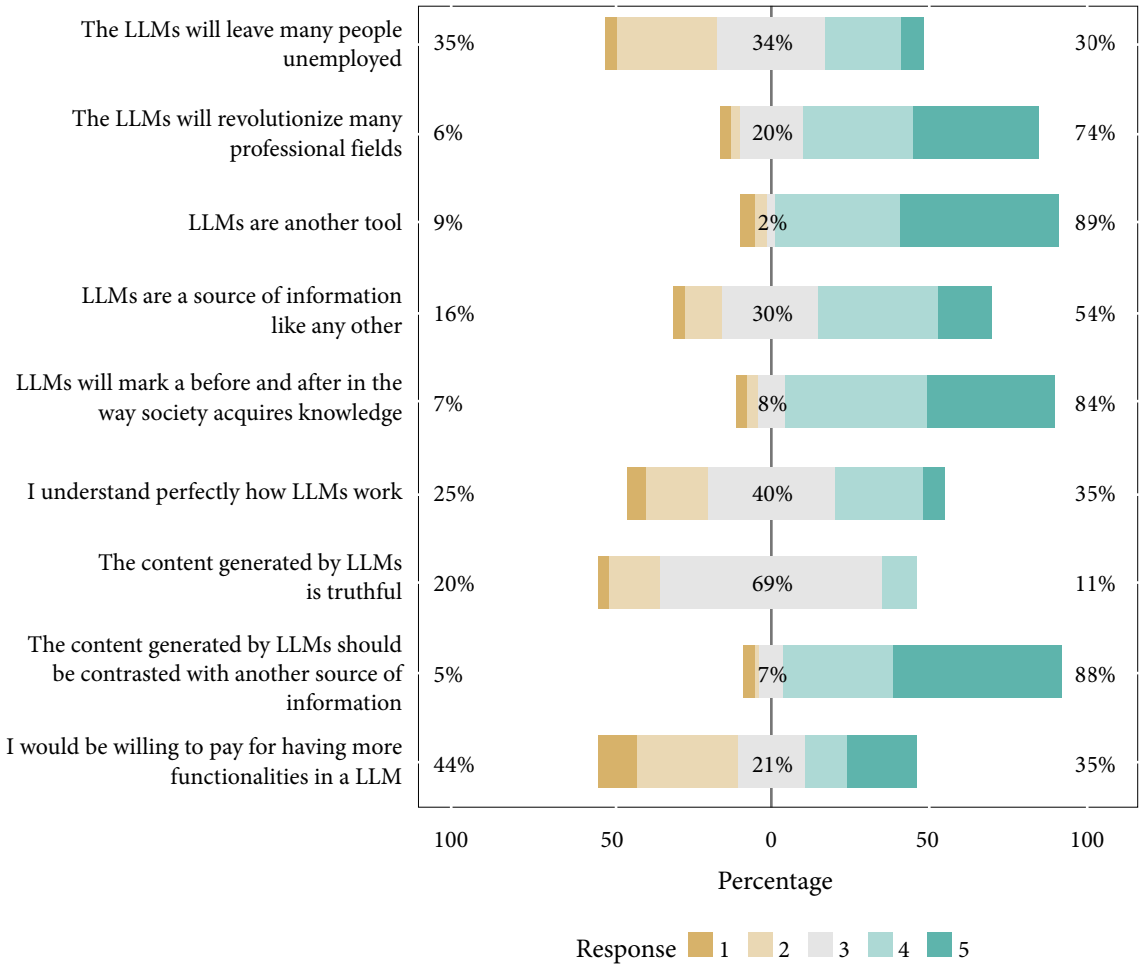
Source: own elaboration.

Students were asked to select from a list of sources they would use to verify the results generated by LLMs. Notably, 82% of students rely on search engines, while 58% use their own knowledge. Additionally, 53% refer to scientific articles, 31% consult Wikipedia, and 22% look at books or seek advice from experts in the field. Lastly, 17% use LLMs to compare results from other LLMs. Regarding their familiarity with using LLMs, 70% of students have engaged with these tools. Among these users, 42% have gained insights from online videos, 17% from reading scientific articles, 10% from attending courses, 7% from discussions with experts, and 6% from watching TV programs, with none citing specialized books as a source of information. In a subsequent question, students ranked five sources of information (in-person or virtual courses, LLMs, search engines, bibliography and scientific articles, and online encyclopedias) on a scale from 1 to 5. The rankings revealed that search engines are the most preferred source, with a total of 295 points. Bibliography and scientific articles follow closely with 274 points. In-person or virtual courses come in third with 251 points. LLMs are ranked fourth with 229 points, and online encyclopedias are the least favored, receiving 196 points.

To evaluate some other aspects of students' perceptions of LLMs, a Likert scale question was employed. The results are presented in Figure 2.

Figure 2

Students' Perceptions Regarding LLMs



Source: own elaboration.

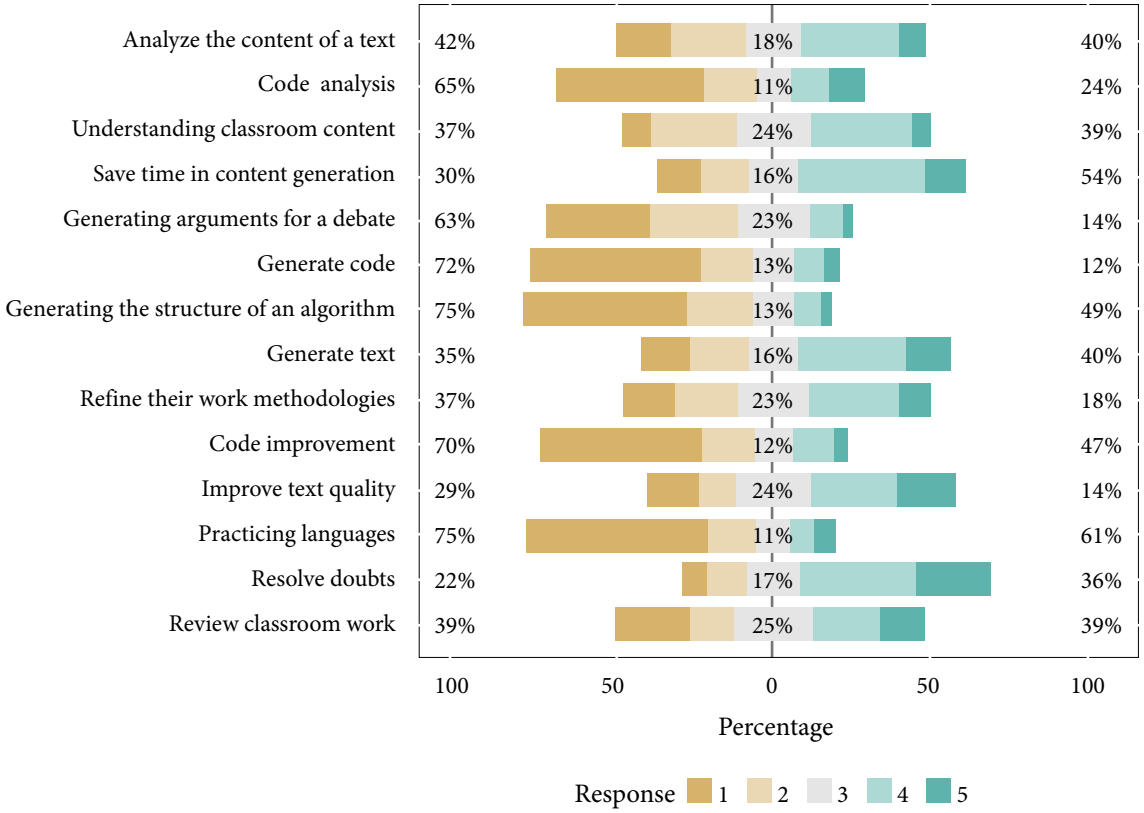
To assess how frequently the surveyed students use LLMs, a Likert scale from 1 to 5 was employed. Here, 1 represents "Never", 2 signifies "Rarely", 3 indicates "Occasionally", 4 stands for "Often", and 5 denotes "Very often". The distribution of responses is illustrated in Figure 3.

The question "Which of the previously mentioned aspects have LLMs helped you with the most?" seeks to determine the most significant impact of LLMs based on earlier evaluated factors. Among the respondents, 19% find LLMs most useful for resolving doubts, closely followed by 18% who appreciate improvements in text quality. Additionally, 12% value the time-saving benefits of LLMs in content generation. Another 6% of students find LLMs useful for generating text, understanding classroom content, and refining work methodologies. For 5%, LLMs are most impactful for reviewing classroom work or have no notable impact.

Analyzing text content and generating arguments for debates are chosen by the 4% of the surveyed students. In the CS field, 19% of students specifically highlight LLMs' assistance in code analysis, with 11% emphasizing their role in code generation. Structuring algorithms is noted by 7%, while 4% mention LLMs' support in code improvement. Notably, none of the CS students reported using LLMs for language practice.

Figure 3

Students' Frequency of LLMs Usage



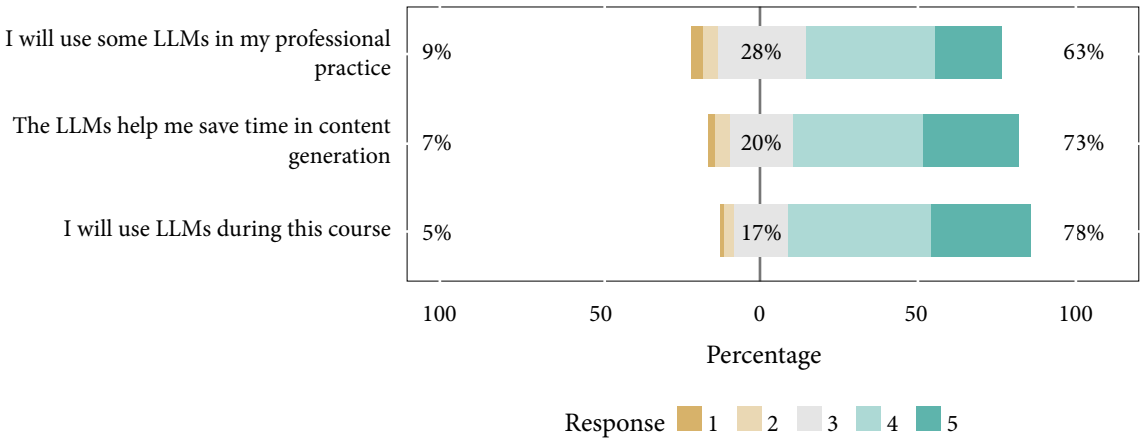
Source: own elaboration.

Looking ahead, 59% of the surveyed students plan to use LLM applications in their professional lives to resolve doubts. Additionally, 48% intend to use LLMs to save time in content generation, 46% for improving text quality, 35% for analyzing text content, 31% for refining work methodologies, and 25% for generating text. In the field of Education, 50% of students plan to use LLMs for generating academic content. In contrast, among CS students, 59% foresee using LLMs for code analysis, 52% for code improvement, 33% for algorithm structuring, and 22% for code generation. To assess three distinct aspects of LLM usage, a Likert scale was employed, with respondents rating from 1 to 5, where 1 means "Strongly disagree", 2

means "Disagree", 3 means "Neither agree nor disagree", 4 means "Agree", and 5 means "Strongly agree". The results are presented in Figure 4.

Figure 4

Students' Agreement Levels on LLMs Usage



Source: own elaboration.

3.2. DIFFERENCES IN THE USAGE OF LLMS BETWEEN THE STUDENTS OF BOTH FIELDS

To address this objective, the analysis of responses to the questions "How often have you used LLMs to: Analyze the content of a text; Code analysis; Understanding classroom content; Save time in content generation; Generating arguments for a debate; Generate code; Generating the structure of an algorithm; Generate text; Refine work methodologies; Code improvement; Improve text quality; Practicing languages; Resolve doubts; Review classroom work" was encompassed.

Using Pearson's χ^2 test of independence identified statistically significant differences in the frequency of LLM usage for four specific activities: code analysis, code generation, algorithm structuring, and code improvement. In the context of code analysis, a significant difference was observed between CS students and Education students, with a χ^2 value of 41.3 and a p-value of < 0.001. CS students demonstrated a higher frequency of using LLMs for code analysis, whereas Education students were more likely to report never using them for this purpose, as detailed in Table 2.

Table 2

Contingency Table for Code Analysis

Code analysis	Never (=1)	Rarely (=2)	Occasionally (=3)	Often (=4)	Very often (=5)	Total
CS	4.8%	4.8%	1.2%	10.8%	10.8%	32.5%
Education	43.5%	12.1%	9.6%	1.2%	1.2%	67.5%
Total	48.3%	16.9%	10.8%	12%	12%	100%

Source: own elaboration.

Regarding code generation, a statistically significant difference was found between CS students and Education students, with a χ^2 value of 36.7 and a p-value of < 0.001. CS students frequently reported using LLMs for generating code, while Education students were more likely to indicate that they never use LLMs for this purpose, as shown in Table 3.

Table 3

Contingency Table for Generate Code

Generate code	Never (=1)	Rarely (=2)	Occasionally (=3)	Often (=4)	Very often (=5)	Total
CS	4.8%	6%	8.5%	9.6%	3.6%	32.5%
Education	50.7%	10.8%	4.8%	0%	1.2%	67.5%
Total	55.5%	16.8%	13.3%	9.6%	4.8%	100%

Source: own elaboration.

Concerning the sub-question about generating the structure of an algorithm, a significant difference was observed between CS students and Education students, with a χ^2 value of 32.5 and a p-value of < 0.001. As shown in Table 4, CS students were more likely to report using LLMs "Occasionally" for generating algorithm structures, whereas Education students were more inclined to state that they never use LLMs for this task.

Table 4

Contingency Table for Generating the Structure of an Algorithm

Generating the structure of an algorithm	Never (=1)	Rarely (=2)	Occasionally (=3)	Often (=4)	Very often (=5)	Total
CS	6%	6%	9.6%	7.2%	3.6%	32.5%
Education	47%	15.8%	3.6%	1.2%	0%	67.5%
Total	53%	21.8%	13.2%	8.4%	3.6%	100%

Font: own elaboration.

When comparing the frequency of LLM usage for code improvement between CS students and Education students, a significant difference was found, with a χ^2 value of 35 and a p-value of < 0.001. CS students generally reported using LLMs frequently for enhancing code, while Education students were more likely to indicate that they never use LLMs for this purpose, as detailed in Table 5.

Table 5

Contingency Table for Code Improvement

Code improvement	Never (=1)	Rarely (=2)	Occasionally (=3)	Often (=4)	Very often (=5)	Total
CS	4.8%	4.8%	7.2%	10.8%	4.8%	32.5%
Education	48.3%	12.1%	4.8%	2.4%	0%	67.5%
Total	53.1%	16.9%	12%	13.2%	4.8%	100%

Source: own elaboration.

Although significant differences were found in four specific sub-questions, those related to code improvement, code generation, code analysis, and algorithm structure generation. These differences predominantly favor the CS field over Education. In contrast, the more generalized sub-questions, which apply to both domains, did not show statistically significant differences. This indicates that students from both fields use these tools with similar frequency in the broader context.

3.3. RELIABILITY THAT STUDENTS GIVE TO LLMS RESPONSES

Regarding Objective 2, students were asked to provide justifications for their Likert scale responses to the statement, "Responses provided by LLMs are reliable." However, 66.3% of respondents did not offer any justification for their answers. Given the limited number of justifications provided, it is not feasible to analyze whether significant differences exist between the responses from the CS and Education fields. The justifications that were submitted, which have been translated from Catalan, are summarized in Table 6.

Table 6

Code Table for Objective 2

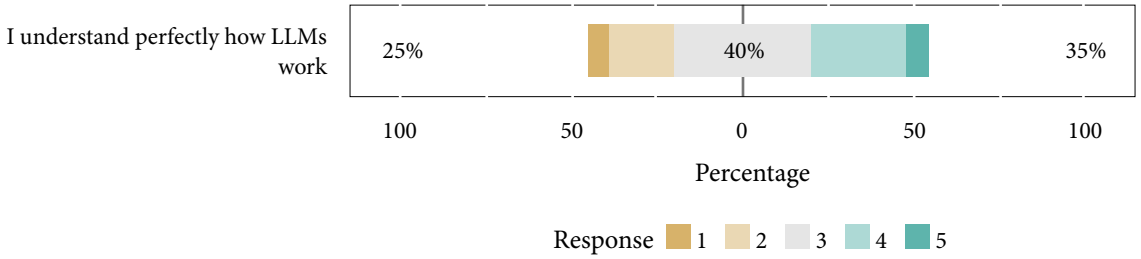
Answer	Code	Example	Frequency
Agree	Specific purpose	"I usually use AI to reformulate texts (explanations) that I've written but I don't like how they look."	1
	Trust in AI	"Since it's an AI, I guess I trust it."	5
	Cross-reference information	"It's usually reliable but you have to verify the information."	2
	Depends on the area	"Depending on the area I'm moving in, it's useful or not."	1
Neither agree nor disagree	Other points of view	"Most of the time the information is not reliable, but it can help you see other points of view. (Speaking of programming, since that's what I use AI with)"	1
	Tautological argument	"There are things that are true and others that are not so much (I guess)."	5
	Depends on the prompt	"Sometimes it is necessary to specify something specific to him so that he takes it into account and does not forget and even doing this he forgets to do it."	4
	Cross-reference information	"They can be reliable, but as long as you check the information on another reliable website."	3
	Need for prior knowledge	"It depends on the question you ask them, anyway you can't trust 100% of everything that comes out of information, since you have to have some prior knowledge."	1
	Uncertainty of the sources	"They are partially reliable, but we do not know exactly the sources from which the information has been taken, so it is not useful to quote or search if they are true."	2
Disagree	Depends on the area	"If it's a very specific question, etc., there's a higher chance that the whole text is reliable, but you have to check the text a lot."	2
	Depends on the source	"It really depends on what language and where it gets the information. The intelligence must be up to date and if connected to the internet know how to recognize if the sources are reliable."	1
	Need for prior knowledge	"You don't really know where they get them from so you can't really trust these things. Only if you know the answer."	1
NS/NC	Mistrust	"I don't know, but I would say that it is not reliable."	1

Source: own elaboration.

Among the respondents who provided justifications, there is a noticeable trust in AI, tempered by significant concerns about its reliability. The variation in AI performance, along with the importance of cross-checking information from multiple sources, underscores the need for a nuanced understanding of AI's strengths and limitations. Nevertheless, most respondents did not provide justifications for their views on the reliability of LLM-generated answers. This is particularly noteworthy given their apparent tendency to affirm their understanding of how these tools work, as shown in Figure 5.

Figure 5

Students' Perspectives on Understanding how LLMs Work



Source: own elaboration.

This paradox highlights a potential gap between respondents' perceived understanding of how LLMs works and their ability to provide well-reasoned justifications for their beliefs. Additionally, the lack of strong justifications suggests that respondents may not critically engage with the information provided by these technologies to the extent needed. A deeper understanding of LLM functionality could enable respondents to offer more nuanced and reasoned evaluations of the reliability of LLM-generated responses. Consequently, the gap between perceived understanding and the ability to justify beliefs points to a need for enhancing students' CT skills.

3.4. STRATEGIES USED BY THE STUDENTS TO VALIDATE THE ANSWERS GIVEN BY LLMS

This objective required analyzing responses to the question "How do you guarantee the reliability of the responses generated by LLMs when you have doubts about their content?". The responses were translated from Catalan. Out of the 83 respondents, 6 did not provide a valid answer. The valid responses are summarized in Table 7.

When asked about their strategies for verifying the reliability of LLM-generated responses, a significant majority (78%) stressed the importance of verification. This indicates a proactive stance among students, who understand the need to cross-check information from LLMs when in doubts through external validation. Additionally, some students emphasize using their

prior knowledge to assess LLM-generated content. By leveraging their own understanding of a subject, these students advocate for a contextual evaluation of LLM responses, suggesting that familiarity with a topic can be a useful benchmark for assessing the accuracy and credibility of AI-generated information. This approach reflects a critical mindset among students when evaluating the content produced by these tools.

Table 7

Code Table for Objective 3

Categories	Code	Example	Frequency
They justify	Cross-reference with other sources	"What I do is look for the same thing from different sources to make a comparison and determine how truthful or not an answer I doubt is."	22
	Cross-reference with internet information	"I look for it on the Internet, whether the information is correct or not, because many times these AI do not give you the exact information, for example in the case of ChatGPT."	17
	Cross-reference with prior knowledge	"Depending on what you know about the subject you can know if it's reliable or not but if you don't know anything about what you're asking for you have to be careful because you can't guarantee it's reliable."	11
	Cross-reference with references	"Ask for the bibliography of the text and read the pages."	10
	Cross-reference with other LLMs	"Comparing different LLMs to see if they all say the same or not."	6
They do not justify	Lack of verification	"I do NOT verify."	10
	Trust in AI	"I don't usually doubt the reliability of the answers."	8

Source: own elaboration.

4. CONCLUSIONS

In this article, a survey was conducted with 83 students from the UdA using a questionnaire with 15 questions aiming to explore their perceptions and usage of LLMs. The results indicate that most respondents are familiar with LLMs, with 98% having used ChatGPT or similar tools, primarily to resolve doubts. Interestingly, 40% of the students expressed confidence in the reliability of LLM responses. This is noteworthy in the context of higher education, where one might expect a lower level of trust. The range of perspectives included concerns about LLM sources, the need for validation of answers, and the importance of crafting precise prompts. Despite these insights, only a few students demonstrated a clear understanding of how LLMs operate. This highlights the need for a deeper comprehension of LLM functionality to enhance informed user engagement.

Analysis indicates no significant differences in LLM usage between CS and Education bachelor students, with notable exceptions in specific areas related to CS, such as generating, improving, and analyzing code, and structuring algorithms. The analysis also revealed that only a small portion of respondents provided thorough justifications for their trust in LLM responses. A key finding from the survey is that 66.3% of students did not provide substantial justification for their confidence in LLM reliability. This suggests a potential gap between perceived trust and a deeper understanding of these tools. Although both CS and Education students generally view LLMs positively regarding reliability, it is important to note that LLMs prioritize sentence coherence over reliability.

While students tend to verify answers when uncertain, their reliance on LLM-generated content may lead to insufficient scrutiny when they are confident in the answers provided. Therefore, enhancing students' understanding of how LLMs function is crucial for developing their CT skills, enabling them to more effectively evaluate the information provided by these technologies. Moreover, strengthening these skills is essential to address the gap between perceived and actual understanding of LLMs and to promote a more critical approach to LLM-generated content. Conversely, relying on the Internet as a primary source of scientific information can impact young people's understanding of credibility in various ways (Pimentel, 2022). The vast availability of information online often complicates the task of distinguishing between credible sources and misinformation or biased content. CT relies on deep knowledge and a thorough understanding of the subject matter (Sandoval & Millwood, 2005). To help students effectively evaluate the reliability of information obtained through LLMs, they need to cross-reference with their existing knowledge and other sources. This process enables them to differentiate between scientifically valid information and unsupported claims. Additionally, a solid grasp of how LLMs function is crucial for students to use these tools effectively and assess the reliability of their outputs. Thus, the rise of LLMs underscores the need to further enhance CT skills.

Universities should integrate LLMs into their teaching practices, acknowledging their growing importance in modern education. This integration will not only teach students to use these tools responsibly but also prepare them for a future where interaction with advanced LLMs is routine.

Moving forward, it is essential to expand this research to include other courses at the UdA, providing a comprehensive perspective on the only public university in the country.

ACKNOWLEDGEMENTS

We sincerely express our gratitude to the UdA for providing the opportunity to conduct the survey, offering us the necessary facilities and resources. Special appreciation is extended to the students from the fields of Education and CS who participated in the survey.

REFERENCES

Alaeddine, M. and Tannoury, A. (2021). Artificial Intelligence in Music Composition. in *Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings 17,* Springer, 387–397. https://doi.org/10.1007/978-3-030-79150-6_31

Allueva, P. (2002). Conceptos básicos sobre metacognición. *P. Allueva, Desarrollo de habilidades metacognitivas: programa de intervención,* 59–85.

Baidoo-Anu, D. and Ansah, L. O. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Baskara, F. R., Puri, A. D. and Wardhani, A. R. (2023). ChatGPT and the Pedagogical Challenge: Unveiling the Impact on Early-Career Academics in Higher Education. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *5*(3), 311–322. https://doi.org/10.23917/ijolae.v5i3.22966

Bellman, R. E. (1978). Artificial intelligence: Can computers think? *Boyd and Fraser Publishing Company*

Bengio, Y., Ducharme, R. and Vincent, P. (2000). A Neural Probabilistic Language Model. *Advances in neural information processing systems*, *13*. https://doi.org/10.1007/10985687_6

Berg, G. and Plessis, E. (2023). ChatGPT and Generative AI: Possibilities for Its Contribution to Lesson Planning, Critical Thinking and Openness in Teacher Education. *Education Sciences*, *13*(10), 998. https://doi.org/10.3390/educsci13100998

Bliss, J., Monk, M., Ogborn, J. and Black, P. J. (1983). Qualitative data analysis for educational research: A guide to uses of systemic networks. *London: Croom Helm*.

Bloom, B. S., Engelhart, M. B., Furst, E. J., Hill, W. H. and Krathwohl, D. R. (1956). TAXONOMY OF EDUCATIONAL OBJECTIVES The Classification of Educational Goals. *New York: Longmans Green*.

Carrera, F. X., Vaquero, E., Balsells, M. À. et al. (2011). Instrumento de evaluación de competencias digitales para adolescentes en riesgo social. *Edutec: revista electrónica de tecnología educativa*. https://doi.org/10.21556/edutec.2011.35.410

Couso, D. and Márquez, C. (2023). Pensar críticament a l'aula de ciències. Activitats competencials per a estudiants de secundaria. *Editorial Graó*.

Dunn, D. S., Halonen J. S. and Smith, R. A. (2009). Teaching Critical Thinking in Psychology: A Handbook of Best Practices. *John Wiley & Sons*. https://doi.org/10.1002/9781444305173.ch1

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. *WH Freeman/ Times Books/Henry Holt & Co*

Facione, P. (1990). Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction. *California State University.*

George-Reyes, C. E., López-Caudana, E. O. and Ramírez-Montoya, M. S. (2023). Research competencies in university students: Intertwining complex thinking and Education 4.0. *Contemporary Educational Technology, 15*(4), ep478. https://doi.org/10.30935/cedtech/13767

González, L. A. O., Baren, C. Y. O. and Zapata, E. J. P. (2023). El impacto de la inteligencia artificial en el ámbito educativo. *Revista Científica FIPCAEC (Fomento de la investigación y publicación científico-técnica multidisciplinaria). ISSN: 2588-090X. Polo de Capacitación, Investigación y Publicación (POCAIP), 8*(3), 342–354.

Halpern, D. F. (1998). Teaching Critical Thinking for Transfer Across Domains Dispositions, Skills, Structure Training, and Metacognitive Monitoring. *American psychologist 53*(4), 449-455. https://doi.org/10.1037/0003-066x.53.4.449

Huschens, M., Briesch, M., Sobania, D. and Rothlauf, F. (2023). Do You Trust ChatGPT? – Perceived Credibility of Human and AI-Generated Content. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2309.02524

Irfan, M., Murray, L. and Ali, S. (2023). Insights into Student Perceptions: Investigating Artificial Intelligence (AI) Tool Usability in Irish Higher Education at the University of Limerick. *Global Digital & Print Media Review, VI*(II), 48–63. https://doi.org/10.31703/gdpmr.2023(VI-II).05

Izquierdo, M. and Aliberas, J. (2021). Pensamiento crítico y valores en las distopías del no futuro [sesión de simposio]. *XI Congreso Internacional sobre Investigación en la Didáctica de las Ciencias: Aportaciones de la educación científica para un mundo sostenible,* 1923–1926.

Jiao, W., Wang, W., Huang, J.-t., Wang, X. and Tu, Z. (2023). Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2301.08745

Kasneci, E., Seßler, K., Kuchemann, S. et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and individual differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, G. and Chun, S. Y. (2023). DATID-3D: Diversity-Preserved Domain Adaptation Using Text-to-Image Diffusion for 3D Generative Model. in *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition*, 14203–14213. https://doi.org/10.1109/cvpr52729.2023.01365

Kurzweil, R., Richter, R., Kurzweil, R. and Schneider, M. L. (1990). The age of intelligent machines. *MIT press Cambridge*, *580*.

Liu, B. L., Morales, D., Roser-Chinchilla, J. et al. (2023). Harnessing the Era of Artificial Intelligence in Higher Education: A Primer for Higher Education Stakeholders.

Liu, M., Wei, Y., Wu, X., Zuo, W. and Zhang, L. (2023). A Survey on Leveraging Pre-trained Generative Adversarial Networks for Image Editing and Restoration. *Science China Information Sciences*, *66*(5), 1–28. https://doi.org/10.1007/s11432-022-3679-0

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Ningrum, S. et al. (2023). ChatGPT's Impact: The AI Revolution in EFL Writing. *Borneo Engineering & Advanced Multidisciplinary International Journal*, *2*, 32–37.

OpenAI, *Chatgpt*. (2022). [Online]. Available: https://chat.openai.com.

Paul, R. and Elder, L. (2005). Una Guía Para los Educadores en los Estándares de Competencia para el Pensamiento Crítico. *Estándares, Principios, Desempeño, Indicadores y Resultados. Con una Rubrica maestra en el pensamiento crítico*.

Pimentel, D. (2022). Learning to evaluate scientific evidence in the age of digital information. *eBook of Synopses*, 43.

Sabzalieva, E. and Valentini, A. (2023). ChatGPT and Artificial Intelligence in higher education: Quick start guide.

Sandoval, W. A. and Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction*, *23*(1), 23-55. https://doi.org/10.1207/s1532690xci2301_2

Singh, H., Tayarani-Najaran, M.-H. and Yaqoob, M. (2023). Exploring Computer Science Students' Perception of ChatGPT in Higher Education: A Descriptive and Correlation Study. *Education Sciences*, *13*(9), 924. https://doi.org/10.3390/educsci13090924

Solís, M. E. C., Martínez, E. L., Degante, E. C., Godoy, E. P. and Martínez, Y. A. (2023). Inteligencia artificial generativa para fortalecer la educación superior: Generative artificial intelligence to boost higher education. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, *4*(3), 767–784. https://doi.org/10.56712/latam.v4i3.1113

Susnjak, T., & McIntosh, T. R. (2024). ChatGPT: The end of online exam integrity? *Education Sciences, 14*(6), 656. https://doi.org/10.3390/educsci14060656

Tura, L. V., Bargalló, C. M., Prat, B. O. et al. (2023). Una propuesta para el diseño de actividades que desarrollen el pensamiento crítico en el aula de ciencias. *Revista Eureka Sobre Enseñanza Y Divulgación De Las Ciencias, 20*(1). https://doi.org/10.25267/rev_eureka_ensen_divulg_cienc.2023.v20.i1.1302

UNESCO. (2019). BEIJING CONSENSUS on artificial intelligence and education.

Vaswani, A., Shazeer, N., Parmar, N. et al. (2017). Attention Is All You Need. *Advances in neural information processing systems*, *30*. https://doi.org/10.48550/arXiv.1706.03762

Zhang, C., Zhang, C., Zheng, S. et al. (2023). A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2303.11717