Mathematics and Critical Thinking in the AI ERA: Rethinking Classroom Practices

Cristina P. S. Dias
cpsd@ipportalegre.pt
Portalegre Polytechnic University
Portalegre, Portugal
NOVAMATH – Center for Mathematics and Applications
New University of Lisbon, Portugal

Luísa M. S. Carvalho
luisacarvalho@ipportalegre.pt
CARE – Research Center on Health and Social Sciences
Portalegre Polytechnic University
Portugal
Center for Research in Education and Psychology (CIEP-UE)
University of Évora, Portugal

Sérgio D. Correia
scorreia@ipportalegre.pt
CARE – Research Center on Health and Social Sciences
Portalegre Polytechnic University
Portugal
Center of Technology and Systems (UNINOVA-CTS) and LASI
Caparica, Portugal

Abstract

In an educational context, increasingly shaped by Artificial Intelligence (AI), mathematics plays a strategic role in fostering critical thinking and problem-solving within an inclusive framework. Moving beyond traditional approaches centered on formulas and mechanical procedures is, therefore, a pedagogical priority. This study implemented and analyzed two didactic proposals integrating AI tools into 120-minute problem-solving sessions with higher education students. The first session emphasized guided exploration of an AI tool, focusing on question formulation, answer analysis, and strategic use of keywords. The second involved group work with mental calculation supported by AI, where the assessment considered strategy, interpretation, and collaboration. Findings highlight increased student autonomy, improved problem-solving skills, and deeper critical engagement with AI in mathematical reasoning.

Keywords

Artificial Intelligence, Critical Thinking, Inclusive Education

1 Introduction

The COVID pandemic accelerated the digital transformation of higher education, highlighting the role of emerging technologies in teaching and learning. Among these, artificial intelligence (AI) has gained relevance for its potential to personalize learning, support assessment, and foster student autonomy, while also contributing to inclusive practices [17, 5]. However, many teachers face difficulties in adopting such tools due to limited training and the absence of clear competency frameworks [11].

In mathematics education, AI represents a strategic opportunity to enhance learning experiences [4]. Tools, such as ChatGPT and Microsoft Copilot, can operate as virtual tutors, offering explanations, feedback, and adaptive guidance that reinforce understanding of abstract concepts. More broadly, AI can support the development of higher-order skills, critical thinking, argumentation, and decision-making, considered central to 21st-century mathematics education [9].

This study examines the integration of ChatGPT and Copilot into statistical problem-solving tasks, with a focus on their impact on student autonomy, hypothesis formulation, and critical thinking in higher education contexts, aiming for total inclusiveness.

Information Society 2025, Ljubljana, Slovenia

http://doi.org/10.70314/is.2025.digin.19

2 Artificial Intelligence and Critical Thinking: Risks, Challenges, and Educational Implications

AI has emerged as a transformative tool with the potential to amplify cognition, personalize learning, and support decision-making [13]. Its integration in education benefits students and institutions, while fostering inclusion [18]. Yet its ubiquity raises concerns for critical thinking, a skill requiring interpretation and reflection [6]. Nevertheless, over-reliance on AI risks leads to the blind acceptance of answers, weakening autonomy and reasoning [1]. Trust in virtual assistants or predictive systems may foster passivity, while deepfakes further blur reality, spreading disinformation and eroding trust [7].

Educating people to use AI critically and ethically is urgent. AI literacy must promote questioning, awareness of bias, and reliance on diverse sources [8]. In education, AI should stimulate reasoning, not replace it. Tools such as ChatGPT and Copilot can act as cognitive mediators, enabling hypothesis testing, comparison of strategies, and feedback [15]. Combining human and AI assessments strengthens evaluation of critical thinking [19], aligned with critical mathematical literacy [16].

AI is built by humans, with subjectivity influencing data, algorithms, and applications [12]. Thus, responsibility remains central. Preparing students requires cultivating ethical and critical interaction with intelligent systems, ensuring AI becomes a collaborator rather than a substitute in developing autonomous, creative citizens [13].

3 Methodology

This study adopts a qualitative, descriptive, and exploratory approach to examine teaching and learning in mathematics education supported by AI tools. As Bogdan and Biklen [3] note, qualitative research focuses on meanings participants attribute to their experiences in natural contexts.

3.1 Description of Pedagogical Practice

This proposal was designed for two 120-minute sessions in a computer room with internet access. Its main objective was to foster understanding and application of hypothesis testing through an interactive, AI-mediated approach. The first session introduced the fundamentals of statistical inference. Students worked in small groups, beginning with an activity to explore their preconceptions, followed by a teacher-led discussion of core concepts:

null and alternative hypotheses (H_0, H_1), type I and II errors, significance level, test statistic, and p-value. ChatGPT and Microsoft Copilot supported this stage by reformulating textbook concepts, generating examples (including intentionally incorrect ones), and illustrating AI's role in interpreting statistical decisions. Students were encouraged to ask questions such as: (1) How do I know if I reject H_0? (2) What does a p-value of 0.03 mean? (3) Can the sample mean be used to reject H_0?

The first session introduced the statistical language of hypothesis testing and encouraged critical reflection on AI-generated answers. The second focused on applying this knowledge collaboratively, with groups analyzing data sets and addressing inferential questions such as:

- *Do school students sleep less than 7 hours a night on average?*
- *Is there evidence that the proportion of users satisfied with AI exceeds 60%?*

Based on the data provided, the groups had to: (i) formulate the null hypothesis (H_0) and the alternative hypothesis (H_1), (ii) define the appropriate significance level (e.g. 0.05), (iii) calculate or interpret the test statistic and p-value, (iv) make a reasoned inferential decision, (v) use ChatGPT or Copilot to confirm the reasoning and explore alternative explanations or validate the conclusions generated with AI support.

Throughout the process, the teacher took on the role of facilitator, circulating among the groups, promoting debate, answering conceptual questions, and encouraging comparison between their own resolution and the resolution proposed by the AI.

3.2 Practice Evaluation

The assessment was formative, guided by explicit criteria, which valued: (a) the correct formulation of hypotheses, (b) the appropriate interpretation of the p-value, (c) the apparent justification of the statistical decision made, (d) the critical and conscious use of AI tools (avoiding automatic or uncritical responses), (e) the clarity and rigor in communicating the results.

To organize the teaching practice, the teacher distributes three problem situations to the groups to work on the topic of Hypothesis Testing. The three problem situations include simulated data and clear questions, ready to be used either on paper or using ChatGPT and Copilot. The first problem situation deals with students' hours of sleep:

- *The aim is to find out whether students at a school sleep, on average, less than 7 hours a night. A random sample of 20 students was taken.*
- *Based on this data supplied, is there statistical evidence that students sleep less than 7 hours a night?*

Instructions given by the teacher: (1) formulate the hypotheses H_0 and H_1; (2) consider the significance level $\alpha = 0.05$; (3) use AI to calculate the p-value or compare it with the critical value; (4) decide whether or not to reject H_0; (5) justify your decision with the support of AI (you can ask ChatGPT or Copilot to run the t-test for this sample and interpret the result).

For the second Problem-situation (satisfaction with an AI app), the following statement was given:

- *"A company wants to know if more than 60% of users are satisfied with its new AI application."*

Instructions given by the teacher were the same as the previous ones.

For the third problem situation, the aim is for students to recognize that this is a hypothesis test for the difference between two means, for which the following statement is provided:

- *"Two groups of students used different methods to study statistics. Group A used only textbooks; Group B used AI as support (ChatGPT/Copilot). After a test, the results (out of 20 points)."*

The data to consider when solving the problem was given to the students, and the students were asked to search for:

- *Is there a significant difference between the means of the two groups?*

Instructions given by the teacher were the same as the previous ones.

3.3 Pedagogical Approach to the First Problem Situation - Students' Sleeping Hours

In this phase, the teacher contextualizes the first Problem Situation and addresses the students by stating:

- *Let's explore whether the students at our school sleep less than 7 hours a night.*

To do this, they should review the following key concepts: the difference between H_0 and H_1, the meaning of the p-value, the significance level (α), and whether the test is one-sided ($\mu < 7$). During the resolution, the teacher moves around the classroom and interacts with the students, actively mediating and making some observations:

- *Have you formulated the hypotheses?*
- *Why is this a one-sided test and not a two-sided test?*
- *Does the p-value you considered make sense in light of the sample mean?*

The teacher began by encouraging students to share strategies freely, fostering collaboration. Discussion then shifted to traditional resources for solving statistical problems. Though initially shy, students soon interacted, and one group presented its solution on the board. The teacher promoted debate by questioning other groups, leading to disagreements about the meaning of the p-value. Inviting an alternative answer, the teacher compared strategies until the class identified the most statistically sound result, highlighting the value of critical analysis. Finally, the teacher gave a guided talk on digital resources, introducing AI and its applications in research and problem-solving.

3.3.1 A practical introduction to the use of AI. The teacher presented two AI tools, ChatGPT and Microsoft Copilot, and explained their potential applications in pedagogy. The teacher demonstrated how to formulate transparent and objective questions, how to interpret and evaluate the answers provided by AI, and how they can support mathematical and statistical reasoning. The teacher conducted a practical example that helped students better understand how the technology works, promoting a critical, ethical, and responsible attitude in its use. The importance of using appropriate keywords when formulating AI questions was also stressed. The teacher also pointed out that vague or poorly structured questions can generate inaccurate or decontextualized answers. For example, instead of asking "lower mean" or "statistical test", it would be more effective and specific to ask:

- *Can you do a t-test to see if the sample mean is less than 7?*

The teacher also warned of the risks of excessively long and confusing questions, which make it difficult for the AI to understand. The teacher used the following inappropriate wording as an example:

- *"We have a set of data and we want to know if the mean can be considered statistically different from the mean of another school because the students sleep little, and we want to know if this is relevant and what to do with the data..."*

This intervention aimed to help students reflect on clarity and precision in mathematical communication, as well as to utilize AI tools as a support for critical thinking, rather than as a substitute for autonomous reasoning. The students are given some guiding questions to think about, about what to ask the AI:

- *Does AI understand everything at once, or should we divide our question into clear and objective parts?*
- *In statistical terms, how can we make our question clearer?*
- *How does AI identify keywords such as "p-value", "mean", "significance" or H_0?*

It is explained to the students that the keywords serve as clues for the AI, allowing the tool to select the appropriate statistical method (t-test, z-test, p-value, etc.) and correctly interpret the desired outcome. The lesson continues with some questions for the students, which serve to direct what they want to get from the AI:

- *Do we want to know if there are significant differences?*
- *Is the mean higher or lower than a certain specified value?*
- *Is the proportion different?*

After the theoretical explanation and the initial example, the teacher returned to the first Problem Situation, illustrating it with the statement:

- *I want to know if the students' mean number of hours of sleep is less than 7.*

In this context, he reminded students of the importance of using relevant keywords, such as "t-test", 'mean', "less than 7", and "p-value". The teacher then challenged the students to consider the most effective way to ask AI questions, supporting them in formulating more transparent and more precise questions. Examples of guiding questions included, *"what is the parameter we are testing?"*, *"what is the null hypothesis (H_0) and the alternative hypothesis (H_1)?"*, *"what is the mean value taking into account H_0?"*, *"What is the significance level (α)?"*, and *"what type of test is most appropriate (t, z, one-sided, two-sided)?"*. To support the organization of thought, the teacher also explained how to divide the questions progressively and gave examples of how to do implement it: *"What is the mean of the sample?"*, *"What is the value of the mean that we are going to test?"*, *"What is the sample size?"*, *"What is the alternative hypothesis?"*. Finally, the teacher presented a well-structured instruction:

- *Run a one-sided t-test on this data to see if the mean is less than 7. Consider $\alpha = 0.05$.*

3.3.2 Group Work and Interactions with AI. Each group selected a problem and, based on prior examples, formulated questions for the AI to obtain rigorous statistical answers. With the teacher's support and real-time projection, they saw how small changes in wording—such as shifting from one- to two-sided tests—could alter conclusions. While groups worked, the teacher circulated, reviewing hypotheses (H_0, H_1), checking test choices (t, z, one-/two-sided), guiding clearer questions for the AI, and ensuring correct interpretation of p-values.

3.3.3 Mediation of Difficulties and Collective Discussion. One of the groups showed additional difficulties, prompting the teacher to intervene, asking:

- *Is your alternative hypothesis consistent with the problem question?*
- *Does the p-value you obtained indicate evidence against H_0? Why?*

After completing the tasks, the class held a collective discussion based on the projected answers. To encourage reflection, the teacher asked questions such as:

- *What was the result of your sample? Was the mean less than 7?*
- *What was the value of the t-statistic and the p-value?*
- *What decision did you make? Did you reject H_0 or not?*
- *Do you think AI helped to better interpret the problem? Why?*

The teacher then projected an answer generated by the AI, previously selected as clear (or confusing), asking the students to assess its validity. One of the groups compared the AI answer with their own, concluding that they preferred their resolution because it was simpler and they understood the reasoning better. The teacher took the opportunity to emphasize that AI does not replace human statistical reasoning, but only supports it. To deepen the assessment of statistical understanding, the teacher issued a provocative challenge:

- *If the AI told you not to reject H_0 with a p-value of 0.02, what would you say?*

The group answered correctly, "if the p-value is 0.02 and the significance level is 0.05, then as $0.02 < 0.05$, we must reject H_0. There is sufficient evidence against H_0." The teacher continued to stimulate critical thinking with new questions:

- *What if the significance level was 1%?*
- *What if the sample had 50 students?*

The answers given by the group revealed a solid understanding: "If $\alpha = 0.01$, then $0.02 > 0.01$, so we don't reject H_0"; "with 50 students, the test would be more accurate. With more data, it becomes easier to determine if there is a real difference in sleeping hours."

3.3.4 Discussion on statistical errors and AI limitations. To assess understanding of type I and II errors, the teacher made the following comment:

- *In the problem situation of hours of sleep, if H_0 is true but the p-value is 0.03 and we reject H_0 with $\alpha = 0.05$, what kind of mistake have we made?*

Responses generated by AI for each group:

- Group I: Type II error, because we rejected H_0 even though it was true (they used ChatGPT).
- Group II: It could be a type I or II error, depending on the interpretation (they used ChatGPT).
- Group III: Type I error, because we rejected H_0 when it is true (they used Microsoft Copilot).

The teacher projected the three answers and asked the groups to evaluate them. Group III acknowledged that only their answer was correct. Group II insisted that theirs also made sense, but eventually recognized that the AI shouldn't give contradictory answers. Group I remained undecided. The teacher took the opportunity to explain that sometimes AI can present "hallucinations" or incorrect answers. The teacher again reminded the students that a Type I error consists of rejecting H_0 when

it is true and that a Type II error corresponds to not rejecting H_0 when it is false. Finally, it reinforced the importance of critical thinking, emphasizing that students with essential thinking skills can explain their reasoning, explore multiple resolution strategies, and critique fallacious arguments [10].

3.4 Assessment and Reflection on Learning with AI Support

The aim of learning assessment in this pedagogical practice was not only to verify the acquisition of statistical content, but above all to gauge the development of critical thinking and intellectual autonomy among students when interacting with artificial intelligence tools. To this end, a descriptive evaluation summary was drawn up, focusing on four key dimensions of the work carried out by the students in the context of solving statistical inference problems:

A - *Formulating statistical hypotheses - Aims to assess the ability to distinguish and correctly state the null hypothesis (H_0) and alternative hypothesis (H_1), as well as the choice of the appropriate statistical test.*

B - *Interpretation of Results - Observes the understanding of the p-value, significance level, and type I and II errors, as well as the coherent justification of the decisions made.*

C - *Use of Artificial Intelligence - Analyzes how students used ChatGPT and Copilot as tools for reasoning and supporting statistical thinking, distinguishing between passive and critical use.*

D - *Collaboration - Considers students' involvement in group discussions, their ability to explain mathematical strategies, and debate the answers generated using AI.*

This synthesis has been structured into three performance levels: beginner, intermediate, and advanced, allowing for continuous and reflective formative assessment. This approach aims to promote an assessment culture that is more in line with the cognitive demands of the digital age, where in-depth understanding and the ability to question are more valued than simply reproducing procedures.

4 Final Considerations

Students developed greater autonomy in problem solving, improved mathematical reasoning, and deepened critical awareness of AI. They enhanced question formulation and strategy selection, promoting more meaningful learning approaches aligned with AI use in class. These findings agree with Zhou et al. [20], who showed that generative AI fosters self-regulation and strengthens critical thinking. They also support Trikoili et al. [19], who advocate combining human and AI assessment. The practice highlights AI's transformative role in promoting critical thinking, as argued in [14], though requiring careful attention to ethical and technical challenges. Properly integrated, AI enables more personalized learning, benefiting teachers and students alike [2].

Using tools like ChatGPT and Microsoft Copilot in statistical inference, students rigorously formulated hypotheses, interpreted evidence, and critically reflected on AI's role. Autonomy, clarity of statements, and detection of AI errors improved—key indicators of critical thinking. This underlines the need for teacher training in pedagogical AI use, not as substitutes for thought but as mediators of mathematical dialogue and statistical literacy. Teachers should integrate self-regulation strategies to maximize AI's impact (Satone et al., 2025). Further research and curricular references are recommended to embed critical thinking as a transversal competence in mathematics teaching.

Acknowledgements

This work is funded under the AI-Enable project (2022-1-SI01-KA220-HED000088368), and the Fundação para a Ciência e a Tecnologia, under project UIDB/00297/2020 and UID/05064/2023.

References

[1] Salim B Al Maqbali and Nooritawati Md Tahir. 2024. Ai opportunities and risks for students' decision-making skill. In *2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE)*. IEEE, 321–326.

[2] Ahmad Al Yakin, Ahmed J Obaid, Eka Apriani, Souvic Ganguli, Abdul Latief, et al. 2024. The efficiency of blending ai technology to enhance behavior intention and critical thinking in higher education. In *Embedded Devices and Internet of Things*. CRC Press, 242–266. doi:10.1201/9781003510420-14.

[3] Robert C. Bogdan and Sari Knopp Biklen. 2003. *Investigação Qualitativa Em Educação: Uma Introdução À Teoria e aos métodos*. Porto.

[4] Ann Marcus-Quinn and Tríona Hourigan, editors. 2017. *Transforming mathematics teaching with digital technologies: a community of practice perspective. Handbook on Digital Learning for K-12 Schools*. Springer International Publishing, Cham, 45–57. ISBN: 978-3-319-33808-8. doi:10.1007/978-3-319-33808-8_4.

[5] Sérgio D Correia, Ana Cunha, Maja Pusnik, and Bostjan Sumak. 2024. A micro-learning units package for improving inclusive digital education in hei. In *2024 Digital Inclusion in the Information Society (DIGIN)*. Information Society.

[6] PA Facione. 2020. *Critical thinking: What it is and why it counts*. The California Academic Press.

[7] Luciano Floridi. 2023. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press, (Aug. 2023). ISBN: 978-01988830.98. doi:10.1093/oso/9780198883098.001.0001.

[8] Sandra Fabijanić Gagro. 2024. Artificial intelligence in education–current challenges. *Annals of the Faculty of Law in Belgrade*, 72, 4, 725–747. doi:10.51204/Anali_PFBU_24405A.

[9] Wayne Holmes, Maya Bialik, and Charles Fadel. 2019. *Artificial Intelligence in Education. Promise and Implications for Teaching and Learning*. (Mar. 2019). ISBN: 978-1794293700.

[10] Jane Swafford Jeremy Kilpatrick and Bradford Findell. 2001. Adding it up: helping children learn mathematics. *National Academies Press*, (Nov. 2001). doi:10.17226/9822.

[11] Rose Luckin, Wayne Holmes, Mark Griffiths, and Laurie B. Corcier. 2016. *Intelligence Unleashed: An argument for AI in Education*. Pearson.

[12] Melanie Mitchell. 2019. *Artificial Intelligence: A guide for thinking humans*. Farrar, Straus and Giroux.

[13] Jorge Ortega-Moody, Kouroush Jenab, Saeid Moslehpour, Lizeth Del Carmen Molina Acosta, Edward Jhohan Marin Garcia, and José Neftalí Torres Marin. 2025. Swot analysis of artificial intelligence in education. In *2025 IEEE Engineering Education World Conference (EDUNINE)*. IEEE, 1–6. doi:10.1109/EDUNINE62377.2025.10981366.

[14] Teresa Chara-De lo Rios, Beymar Solis-Trujillo, Jhon Perez-Ruiz, and María Aquije-Mansilla. 2025. Systematic review of critical thinking using artificial intelligence. *Edelweiss Applied Science and Technology*, 9, 3, (Mar. 2025), 990–1001. doi:10.55214/25768484.v9i3.5405.

[15] Neil Selwyn. 2019. *Should robots replace teachers?: AI and the Future of Education*. English. (1st ed.). Polity Press, United Kingdom, (Nov. 2019). ISBN: 978-150952896-7.

[16] Ole Skovsmose. 2011. *An invitation to critical mathematics education*. Ole Skovsmose, editor. SensePublishers. doi:10.1007/978-94-6091-442-3.

[17] Boštjan Šumak, Sergio Duarte Correia, Ana Cunha, Tuncer Can, Irfan Simsek, Katja Kous, and Maja Pušnik. 2024. Identification of factors that impact e-inclusion in hei. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE, 478–483. doi:10.1109/MIPRO60963.2024.10569746.

[18] Boštjan Šumak et al. 2024. Ai-based education tools for enabling inclusive education: challenges and benefits. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*. IEEE, 472–477. doi:10.1109/MIPRO60963.2024.10569714.

[19] Anna Trikoili, Despoina Georgiou, Christina Ioanna Pappa, and Daniel Pittich. 2025. Critical thinking assessment in higher education: a mixed-methods comparative analysis of ai and human evaluator. *International Journal of Human–Computer Interaction*, 1–14. doi:10.1080/10447318.2025.2499164.

[20] Xue Zhou, Da Teng, and Hosam Al-Samarraie. 2024. The mediating role of generative ai self-regulation on students' critical thinking and problem-solving. *Education Sciences*. doi:10.3390/educsci14121302.