Guided Integration of Generative AI in Radiography Education: Effects on Presentation Quality and Emerging AI Literacy

Sami Suhonen[1]

[1] Tampere University of Applied Sciences

Abstract

The rapid emergence of generative AI tools, such as ChatGPT, Copilot, Gemini and NoteGPT, challenges higher education institutions to develop effective approaches for their integration into teaching and learning. Rather than restricting or prohibiting AI use, intentionally guiding students toward responsible and structured utilization of AI presents significant potential for enhancing learning outcomes, particularly in fields requiring mastery of complex scientific concepts, such as radiography.

The study addressed three research questions: (1) whether guided AI use improves the quality of students' presentations compared with previous non-AI cohorts, (2) how ChatGPT's rubric-based evaluations align with expert human grading, and (3) how students actually used and experienced AI during their work. All student groups were tasked with preparing presentations on radiography-specific topics: ionization chambers, Geiger-Müller tubes, thermoluminescence dosimeters, direct ion storage dosimeters, and scintillation detectors. The uniformity of these topics across multiple cohorts enables a comparative qualitative analysis, focusing on criteria such as scientific accuracy, clarity, depth of content, structure, and overall presentation coherence. Participating students documented their interactions with AI tools, recording prompts utilized, identifying aspects of their assignments where AI contributed most effectively, and addressing challenges encountered during the AI-assisted collaborative process. They were also instructed to reflect upon and articulate their key learning experiences related to AI use.

Across 28 presentations (12 AI-assisted), no statistically significant differences were found in scientific accuracy, clarity, or visual quality between AI and non-AI groups, indicating that human factors dominated quality variation. Correlations between AI (ChatGPT) and human scoring were negligible ($r \leq 0.22$), contrasting with earlier essay-based findings ($r \approx 0.6$–0.8). Qualitative reflections showed that students used AI primarily for conceptual understanding and explanation rather than content generation, displaying critical awareness of reliability and verification needs. Guided integration thus supported metacognitive learning and AI literacy even without measurable performance gains. Students reported that AI prompts helped in clarifying complex radiographic concepts, facilitating efficient literature searches, and assisting in the coherent organization of technical information. However, students also reported facing challenges, particularly related to critically assessing the reliability of AI-generated content, refining prompt engineering skills, and balancing dependence on AI tools with independent critical thinking and professional judgment.

Keywords

Guided AI Integration, AI Literacy, AI Assessment Reliability

1. Introduction

Generative artificial intelligence (GenAI) tools such as ChatGPT, Copilot and Gemini have moved rapidly from pilot curiosities to everyday study companions in higher education, prompting sector bodies to publish practical guidance for teaching and assessment (Alcock 2024). Policy advice has shifted away from blanket prohibitions toward human-centred, responsible integration, emphasising clear institutional principles, capacity-building for staff and students, and safeguards around ethics and data (Holmes & Miao 2023).

Radiography education requires students to master demanding radiation-physics constructs and detector operating principles that are safety-critical in practice, yet cohorts are heterogeneous in prior mathematics and physics preparation. This creates a need for instructional scaffolds that translate abstract models into actionable understanding without displacing students' own reasoning. Generative AI tools are a promising candidate for such scaffolds, but their pedagogical role is best framed through contemporary theories of extended cognition and the extended mind. According to the extended mind thesis (Clark & Chalmers 1998) cognition need not be confined to the brain: when external resources, such as notes, diagrams, instruments, or AI systems, are functionally integrated into a person's problem solving and memory, they become constitutive parts of the cognitive process. From the extended-mind perspective, human thinking routinely unfolds across hybrid systems that couple brains with bodies, tools, and external media; new technologies (including LLMs) can become part of these cognitive circuits. Rather than simple "offloading," productive human–AI loops can amplify creative problem-solving while demanding new meta-skills: assessing reliability, calibrating trust, and learning how to question AI-suggested ideas. This view positions GenAI neither as a threat nor a silver bullet, but as a resource whose value depends on how it is integrated into the wider cognitive ecosystem of learning (Clark 2025).

In science education, researchers distinguish between two main ways of using generative AI. Substitutive use means letting AI do the work for the learner, which can make students passive and reduce their own thinking. Complementary use, in contrast, keeps students active and engaged. Recent studies describe three useful complementary ways to use AI in teaching: (1) AI-generated feedback that helps students notice and correct their misunderstandings, (2) supportive functions that make learning more accessible without replacing the learner's effort, and (3) interactive or game-like dialogues that encourage explanation and critical thinking. The key pedagogical goal is to design learning activities where AI supports and extends human thinking, rather than acting as a replacement or main author (Rivera-Novoa & Duarte, 2025).

Our study applies this approach in a radiography course at Tampere University of Applied Sciences by giving students guided opportunities to use generative AI while preparing presentations on standardised detector-physics topics. Students were instructed to ask AI for step-by-step explanations, find sources and technical information, and organize that information. They also recorded the prompts, sources, and verification steps, so that evaluation and reflection became explicit parts of the learning process. The resulting presentations were compared with those of earlier cohorts that had no AI guidance to explore three questions: (1) Does guided use of generative AI improve presentation quality in terms of accuracy, clarity, depth, structure, and coherence? (2) How well do large language model (LLM) rubric-based ratings align with human expert evaluations? (3) How do students actually use and evaluate generative AI when working under guidance?

2. Methdos

The study was conducted in a radiography course in which student teams prepared presentations on detector physics (ionization chambers, Geiger–Müller tubes, direct-ion-storage dosimeters, thermoluminescence dosimeters, and scintillation detectors). Using identical topics across cohorts allows direct comparison of the

outputs. Prior cohorts prepared the assignment using mainly textbooks and web sources. The 2025 cohort received an intentionally guided AI intervention: in addition to textbooks/web, students were encouraged to use AI tools (e.g., ChatGPT, Copilot) in preparing their presentations. Short how-to videos of using Photomath and ChatGPT were provided, together with some example prompts. In preparing the presentations, the students were encouraged to use AI tools to explore the topic through summaries, definitions and concept checks, to outline and structure slides, and to clarify difficult concepts. They were further instructed to conclude with one to two slides that list the prompts used, explain when, where and how AI supported their work, and offer brief reflections on what worked, what proved challenging, and the key takeaways about AI use. The student presentations were examined qualitatively on accuracy, clarity, depth, visualizations, terminology and the usage of sources. Students' metacognitive reflections were investigated on benefits and challenges. This approach follows international recommendations that recognise GenAI literacy, including the critical evaluation of accuracy, credibility, and truthfulness, as an essential emerging competence.

The material consists of 28 presentations, of which 12 were AI-assisted. The AI-supported presentations covered the same technical topics as the non-AI ones but included an additional slide or section explicitly describing how artificial intelligence had been used during the work. Thematically, the AI-assisted groups relied on ChatGPT or Copilot mainly for information retrieval, conceptual clarification, and structuring their presentations, while a few used image-generation tools for visualization. The study addresses three questions:

1. Does guided AI use improve presentation quality relative to earlier cohorts? This was examined by evaluating the presentation within each topic area against criteria for scientific accuracy, conceptual depth, clarity of explanation, structure and coherence, quality of visualizations, correct terminology, use of sources and citations, and overall professionalism. Each aspect was rated from 0 to 5, and the presentation's final mark was computed as the simple average of all criterion scores.

2. How do large-language-model (ChatGPT) ratings align with expert human evaluation and grading of the presentations? To assess alignment, the same presentation PDFs were submitted to ChatGPT with the identical rubric (scientific accuracy, conceptual depth, clarity, structure and coherence, visual quality, terminology, use of sources, overall professionalism), instructing it to produce 0–5 scores per criterion and an overall mean with brief justifications.

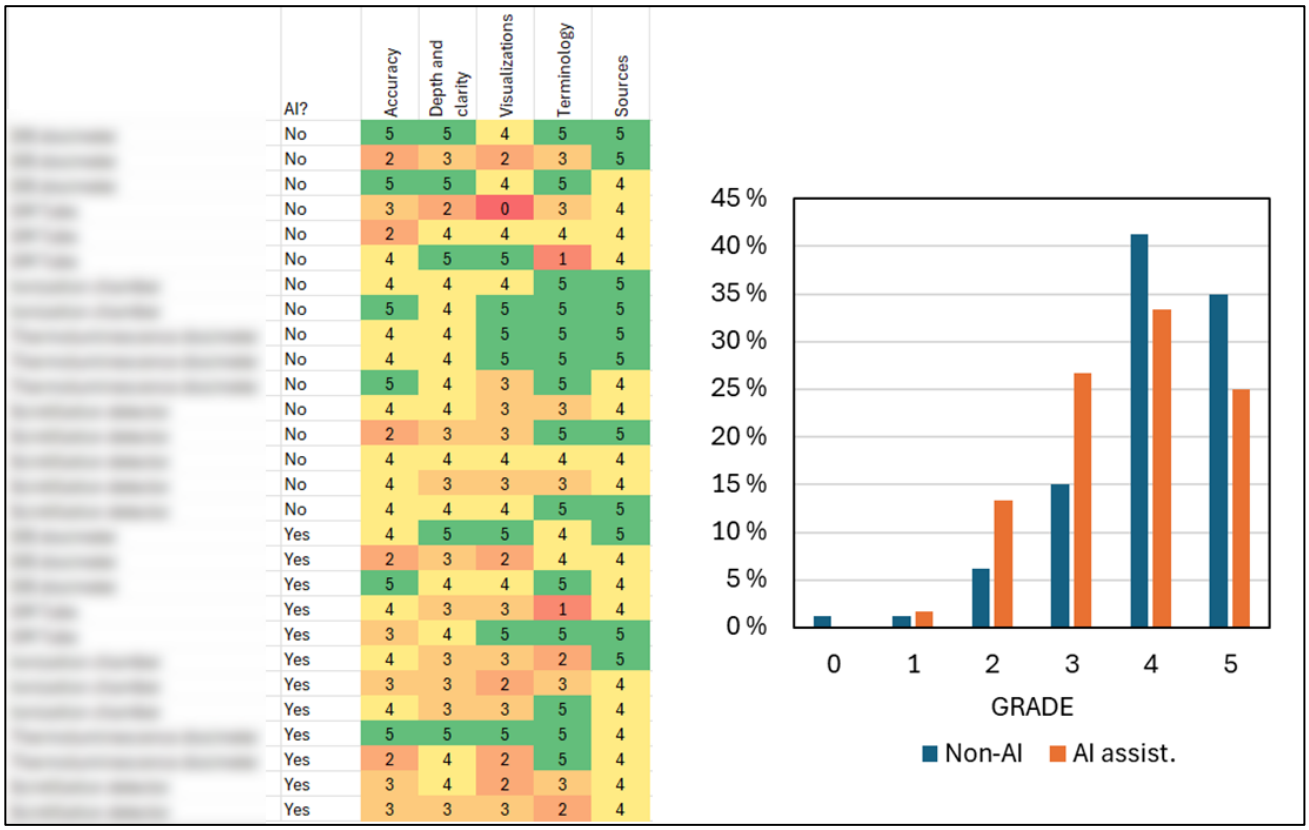
3. How do students actually use AI when preparing their presentations (prompts, verification behaviours, perceived benefits)?

3. Results

3.1 Quality of presentations in AI and non-AI presentations

Figure 1 presents the evaluation results of all student presentations across the five assessment categories: 1) scientific accuracy and depth, 2) clarity and structure, 3) visuals and diagrams, 4) sources, and 5) terminology (left) and the distribution of these grades between AI-assisted and non-AI groups (right). There were no statistically significant differences between AI-assisted and non-AI presentations in any of the rubric dimensions or in the overall scores. Variation was markedly larger at the topic/group level (e.g., between student groups) than between the AI versus non-AI condition, suggesting that quality differences were primarily attributable to the groups themselves rather than to the use of AI. Presentation-to-presentation variation within each cohort was substantial, with scores for individual presentations spanning the full 0–5

range and considerable overlap between cohorts; this within-cohort dispersion outweighed the average AI vs. non-AI differences on most criteria. These results indicate that human factors dominated quality outcomes in this implementation and guided AI did not systematically lift performance beyond that background variability.



AI?	Accuracy	Depth and clarity	Visualizations	Terminology	Sources
No	5	5	4	5	5
No	2	3	2	3	5
No	5	5	4	5	4
No	3	2	0	3	4
No	2	4	4	4	4
No	4	5	5	1	4
No	4	4	4	5	5
No	5	4	5	5	5
No	4	4	5	5	5
No	4	4	5	5	5
No	5	4	3	5	4
No	4	4	3	3	4
No	2	3	3	5	5
No	4	4	4	4	4
No	4	3	3	3	4
No	4	4	4	5	5
Yes	4	5	5	4	5
Yes	2	3	2	4	4
Yes	5	4	4	5	4
Yes	4	3	3	1	4
Yes	3	4	5	5	5
Yes	4	3	3	2	5
Yes	3	3	2	3	4
Yes	4	3	3	5	4
Yes	5	5	5	5	4
Yes	2	4	2	5	4
Yes	3	4	2	3	4
Yes	3	3	3	2	4

Figure 1. Evaluation of student presentations on radiation measurement topics. The heatmap (left) shows individual scores across five assessment categories classified by AI-assisted and non-AI groups. The bar chart (right) presents the overall grade distribution, indicating no significant difference between AI-assisted and non-AI presentations.

3.2 AI and human grading

The anonymized student presentations were given to ChatGPT for evaluation, and they were also independently evaluated by the course's instructor. The presentations were assessed using a five-criterion rubric addressing accuracy and coverage, clarity and structure, visual design, sources, and terminology. The rubric was used both by the human evaluator and ChatGPT, which was instructed to produce 0–5 scores per criterion and an overall mean with brief justifications. Table 1 presents the rubric in detail, describing the performance levels (1, 3, 5) and their corresponding qualitative characteristics for each assessment criterion. Presentations achieving higher scores demonstrated factual precision, coherent organisation, effective visual communication, appropriate referencing, and accurate use of disciplinary terminology. Lower scores indicated superficial understanding, limited structure, insufficient or unclear visuals, inadequate sourcing, or inconsistent terminology.

Table 1: Assessment rubric.

Criteria	Score 5 (Excellent)	Score 3 (Average)	Score 1 (Weak)
Accuracy & Coverage	All content accurate and comprehensive. Covers structure, principle, use cases. Level suitable to radiography students. No errors.	Mostly correct with minor inaccuracies or omissions. Some important aspects missing.	Several errors or misconceptions. Major gaps in coverage, very superficial.
Clarity & Structure	Logical flow, slides are well-structured and easy to follow. Key points highlighted clearly.	Generally understandable structure, but occasional inconsistency or weak transitions.	Confusing, poorly organized, difficult to follow. No clear progression.
Visuals	Images, diagrams and layout strongly support understanding. Text is readable, visually appealing. No AI-generated, background images.	Some visuals but uneven quality. Images partially support content only partially.	Weak visuals. Irrelevant images or unreadable layout.
Sources	Reliable, diverse sources (e.g. STUK, textbooks, scientific articles). Full reference list.	Some sources present.	Unreliable references, if no sources, then score = 0.
Terminlogy	Correct and consistent use of physics/radiation terminology (e.g., scattering, atomic number, µSv/h). Professional style.	Terminology mostly correct, but some mistakes or inconsistent usage.	Terminology inaccurate or wrong. Casual language instead of scientific.

Figure 2 illustrates the relationship between human and ChatGPT evaluations of student presentations across all assessment criteria. Each bubble represents one or more paired scores, with its size corresponding to the number of occurrences. While the general distribution suggests a weak positive trend between the two evaluators, the overall correlation coefficient was only $r = 0.22$, indicating that ChatGPT's ratings aligned with human judgement only to a limited extent. ChatGPT exhibited a narrower scoring range, avoiding the lowest scores and showing a tendency to concentrate around values 3–4, whereas human assessors used the full scale more distinctly. This compression effect implies that ChatGPT was less sensitive to qualitative differences between average and excellent presentations.

To investigate the differences between AI and human evaluations in more detail, grade distributions for each criterion were visualised as boxplots in Figure 3. It presents a criterion-wise comparison between AI (ChatGPT) and human evaluations across five rubric dimensions and the overall score. The boxplots show that the two evaluators produced relatively similar median grades, but with substantial variation within and between categories. The correlation coefficients (r) indicate only weak alignment for all criteria: from $r = 0.02$ in Visuals to $r = 0.45$ in Sources, with a correlation of $r = 0.15$ for the overall presentation grade. The strongest agreement appeared in the Sources category, where both evaluators tended to reward the presence of clear references

and identifiable source material. In contrast, Visuals, Terminology, and Accuracy & Coverage showed almost no correlation, suggesting that ChatGPT was incapable to assess the presentations deeply enough.

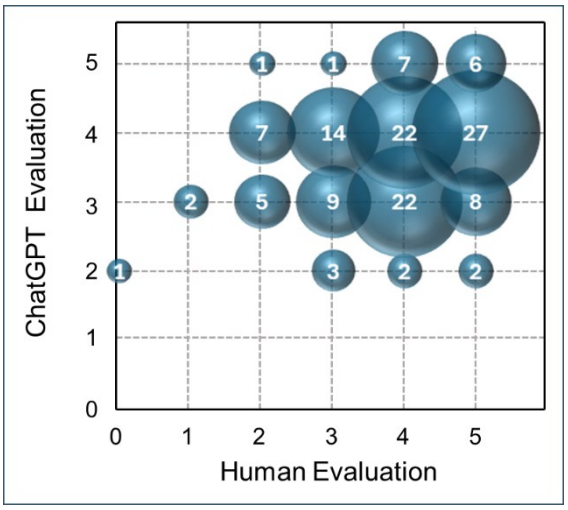


Figure 2. Relationship between human and ChatGPT evaluations of student presentations across all assessmen t criteria. Bubble size represents the number of occurrences.

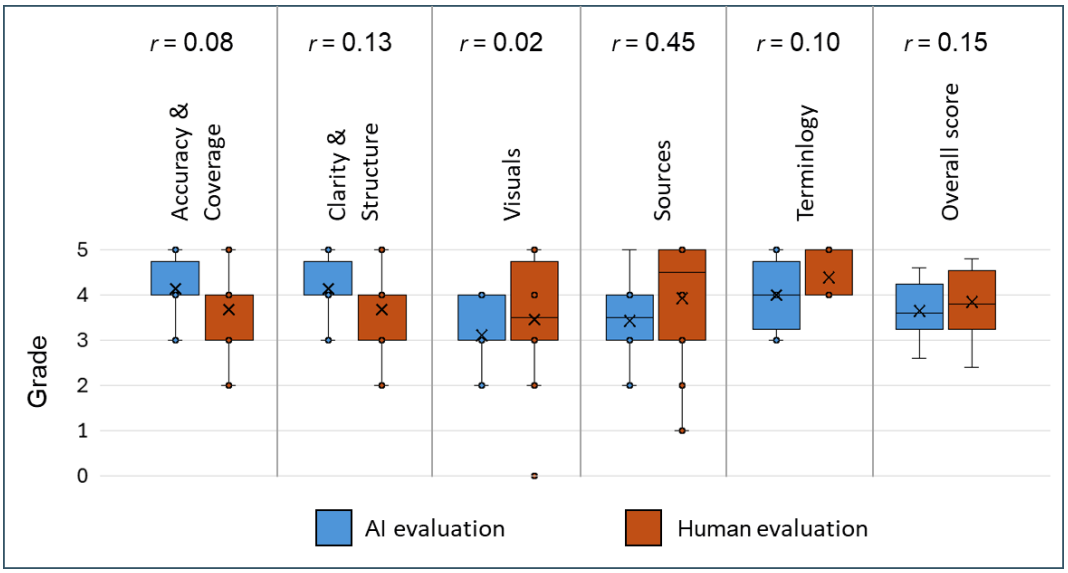


Figure 3. Criterion-wise comparison between AI (ChatGPT) and human evaluations across five rubric dimensions and the overall score with correlation coefficients (*r*).

Compared with the findings of Quah et al. (2024), where correlations between ChatGPT and human assessors in essay scoring reached $r = 0.83$ and $r = 0.60$ for two open-ended clinical questions, the correlations observed in this study were negligibly small ($r \approx 0.22$ or below). This difference is likely explained by the fundamentally different nature of the assessment tasks. While Quah et al. evaluated extended written responses that allowed the AI model to recognise linguistic patterns, argumentation structure, and generic essay features, the present study involved PowerPoint presentations focused on technically precise and physics-based content. Such material requires understanding of quantitative reasoning, conceptual accuracy, and disciplinary terminology, areas in which large language models currently seem to perform less reliably. Overall, the pattern supports the interpretation that ChatGPT captures surface-level features, such as text organisation and reference presence, better than deep disciplinary understanding or contextual accuracy. Evaluating the quality of diagrams and

schematic representations of radiation detectors poses particular challenges for AI, which remains insufficient as a stand-alone assessment tool for this type of task.

3.3 Student reflections on AI use.

Information on students' use and experiences of artificial intelligence was collected through a reflective self-report task that accompanied each presentation. After completing their group work, students were asked to describe how and for what purposes AI tools were used, what benefits or challenges they observed, and how AI influenced their learning or understanding of the topic. The reflections appended to presentation slides. The data were gathered from all AI-assisted groups (n = 12) across the five detector topics (DIS dosimeter, TLD, ionisation chamber, scintillation detector, and Geiger tube). Responses were analysed qualitatively to identify typical patterns of AI use.

Table 2 summarises the main purposes for which students reported using AI tools during the preparation of their group presentations. Six distinct categories of use were identified from the students' reflective reports: information retrieval and conceptual understanding, structuring and text generation, searching for sources, image creation and visualisation, language support and translation, and reflection and learning support.

Table 2: Purposes, descriptions, and frequency of reported AI use in the preparation of student presentations.

Purpose of AI use	**Description and typical examples**	**No of obs.**
Information retrieval and conceptual understanding	Students asked AI to explain technical principles, definitions, and applications (e.g., *"What is an ionization chamber and how does it work?"*, *"What does surface dose mean?*	12
Structuring and text generation	AI used to create outlines, slide headings, or summaries (*"Can you make a PowerPoint template for this topic?"*, *"Condensed version on one slide"*).	3
Searching for sources	Used as Google search. AI helped locate or list industrial actors and reference materials when search engines were insufficient.	3
Image creation and visualization	Students employed AI to generate or design illustrative figures (*"Create a picture showing scintillation detector applications"*, *"Make an image of Geiger counter use"*).	2
Language support and translation	AI assisted in rephrasing, summarising, or translating text between Finnish and English.	2
Reflection and learning support	AI used for clarification of misunderstood concepts and as a conversational tool for self-explanation (*"Good to use AI to explain things when something is difficult to understand"*).	2

The most frequent and clearly valued use concerned information retrieval and conceptual understanding, where students asked AI to explain technical principles, definitions, or applications in radiation physics. This indicates that AI primarily served as a conceptual tutor or explanatory aid, helping students make sense of complex physical phenomena. Other uses—such as structuring text, locating sources, or language translation—appeared occasionally but were less central. Image generation was used only in a couple of presentations. Based on course's instructor's comments, the AI-generated images and illustrations were mostly irrelevant, wrong or misleading for scientific topics. Several students also described using AI to clarify misunderstandings or to rehearse explanations in their own words, which suggests a potential role for AI in supporting metacognitive learning processes. However, this type of use was observed only in 2 presentations. Therefore,

reflective use of AI and iterative improvement of presentations based on AI feedback should be encouraged in the future.

Students' reflections also revealed a notably critical and self-aware approach to using AI tools. Several groups reported that the responses provided by ChatGPT or Copilot were sometimes inconsistent, incomplete, or even misleading—for example, offering contradictory information about manufacturers or confusing "ionization chambers" with ionizing air purifiers. Others noted that the system occasionally redirected them to irrelevant online sources. Such experiences led students to emphasize the need for fact-checking and source evaluation: they learned that question formulation affected the precision and depth of AI's responses and that outputs could not be accepted uncritically. Overall, the reflections demonstrate an emerging form of AI literacy characterized by cautious engagement, iterative questioning, and awareness of the reliability limits of generative systems. From a pedagogical standpoint, this aligns with recent extensions of the extended mind thesis in AI-enhanced learning, where generative AI is seen not merely as a tool but as a cognitive extension of the learner, yet one that requires metacognitive control to avoid passive offloading (Rivera-Novoa et al., 2025).

4. Conclusions

This study explored guided use of generative AI in radiography education through a comparative analysis of student presentations and reflections. While structured AI guidance did not measurably improve presentation quality relative to non-AI cohorts, it enhanced students' reflective awareness and conceptual engagement. The minimal correlation between ChatGPT and human grading ($r \approx 0.2$) contrasts with higher correspondence in essay-based assessments ($r \approx 0.6$–0.8), suggesting that AI evaluation is more effective for linguistic or narrative tasks than for technical and physics-oriented reasoning.

Across cohorts, a consistent pattern emerged: good students produced good presentations both with and without AI, and weaker groups tended to perform modestly regardless of AI support. This indicates that generative AI alone does not equalise learning outcomes; rather, its benefits depend on students' existing skills in critical thinking, conceptual understanding, and self-regulated learning.

Students' reflections showed that AI functioned mainly as an explanatory and organising aid, but in some cases also as a substitute for understanding by providing oversimplifications. Students recognised both the utility and the limits of AI tools, particularly regarding factual accuracy and conceptual depth demonstrating the emergence of AI literacy as a critical professional competence.

Pedagogically, these findings align with the extended mind perspective, where generative AI operates as a complementary cognitive extension that supports but does not replace human reasoning. Effective learning design should therefore foster iterative questioning, verification, and reflection when using AI, ensuring that human judgment and disciplinary expertise remain central. Future research should explore how guided AI practices can be scaled within STEM and health education, balancing automation with epistemic responsibility and maintaining a strong role for educator-led scaffolding.

5. References

Alcock, E. (2024). Assuring and enhancing the quality of AI-transformed higher education: Staying ahead of the curve. European Quality Assurance Forum (EQAF) 2024, European University Association. https://www.eua.eu/publications/conference-papers/assuring-and-enhancing-the-quality-of-ai-transformed-higher-education-staying-ahead-of-the-curve.html

Clark, A., & Chalmers, D. (1998). The extended mind. analysis, 58(1), 7-19.

Clark, A. (2025). Extending minds with generative AI. nature communications, 16(1), 4627.

Holmes, W., & Miao, F. (2023). Guidance for generative AI in education and research. Unesco Publishing.

Rivera-Novoa, A., & Duarte Arias, D. A. (2025). Generative Artificial Intelligence and Extended Cognition in Science Learning Contexts. Science & Education, 1-22.

Smuha, N. A. (2025). Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (EU Artificial Intelligence Act). International Legal Materials, 1-148.

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. BMC Medical Education, *24*(1), 962.