



OPEN ACCESS

EDITED BY

Sergio Ruiz-Viruel,
University of Malaga, Spain

REVIEWED BY

Vinhthuy Phan,
University of Memphis, United States
Kwan Yi,
Eastern Kentucky University,
United States

*CORRESPONDENCE

Daniele Agostini
✉ daniele.agostini@unitn.it

RECEIVED 21 November 2025

REVISED 06 February 2026

ACCEPTED 16 February 2026

PUBLISHED 25 March 2026

CITATION

Agostini D, Serbati A, Picasso F and
Lipnevich A (2026) When ChatGPT joins
the team: a mixed-methods study
of AI-mediated collaborative lesson
design.

Front. Educ. 11:1751618.

doi: 10.3389/feduc.2026.1751618

COPYRIGHT

© 2026 Agostini, Serbati, Picasso and
Lipnevich. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](#). The use, distribution or
reproduction in other forums is
permitted, provided the original author(s)
and the copyright owner(s) are credited
and that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

When ChatGPT joins the team: a mixed-methods study of AI-mediated collaborative lesson design

Daniele Agostini^{1*}, Anna Serbati¹, Federica Picasso² and Anastasiya Lipnevich³

¹Department of Psychology and Cognitive Sciences, University of Trento, Trento, Italy, ²Unit for Teaching Innovation, Skills Development and Certification, Advanced and Postgraduate Education Office, Area for Education, Student Services, Guidance and Internationalization, University of Genoa, Genoa, Italy, ³National Board of Medical Examiners, Philadelphia, PA, United States

Introduction: The application and influence of artificial intelligence (AI), and specifically Large Language Models (LLMs), in educational processes is widely discussed. However, there remains a gap in research on using LLMs as peer-like contributors in collaborative learning contexts.

Methods: This article reports a mixed-methods quasi-experimental study investigating how positioning ChatGPT as a peer-like feedback provider shapes student-teachers' learning and collaboration during group lesson-design activities. The study employed a counterbalanced crossover structure for knowledge assessment and a sequential two-task design for authentic artifact production. A total of 102 teachers in training ($M_{age} = 38.87$, $SD = 8.01$), organized into 21 groups, completed two authentic design tasks within a single session.

Results: Across the session, students progressively adapted to AI interaction, refining how they queried the model and how they evaluated and integrated its suggestions. Results indicate a Post-Withdrawal Sustained Performance (PWSP) effect: improvements observed during AI-available phases were not followed by a detectable decline in the immediately subsequent AI-withdrawn phase within the study timeframe. This pattern was clearest for technology-related knowledge and was consistent with stable artifact quality after AI removal. While ChatGPT support increased efficiency and contributed to technology-focused insights, qualitative evidence also pointed to tensions, including reduced peer-to-peer idea-building in some groups and concerns about creativity.

Discussion: Overall, the findings suggest that integrating LLMs as a feedback team-mate can support collaborative design work without immediate post-withdrawal performance costs, particularly when learners are scaffolded to engage critically with AI output rather than accept it unreflectively. These results carry implications for the design of AI-enhanced collaborative activities, highlighting the need to balance AI efficiency gains with sustained opportunities for authentic peer dialogue.

KEYWORDS

AI in education, AI-in-the-loop, artificial intelligence, computer-supported collaborative learning, educational feedback, educational technology, human-computer interaction, large language models

1 Introduction

OECD defines an AI system as a “machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment” (OECD, 2024, p. 4).

The field of Artificial Intelligence in Education (AIED) has made significant strides over the past 25 years, advancing both theoretical understanding and practical applications (VanLehn, 2006; Koedinger and Corbett, 2006; Heffernan and Heffernan, 2014; Roll and Wylie, 2016). Instead of just automating the instruction of students sitting in front of computers, AI could help open up teaching and learning opportunities that would otherwise be difficult to achieve: it questions conventional pedagogies and may assist instructors in becoming more successful.

Other AIED technologies are designed to monitor student progress and provide personalized feedback, enabling educators to assess whether a student has achieved mastery of a given topic (Corbett and Anderson, 1994; Meyer et al., 2024; VanLehn, 2011). Similarly, AIED tools developed to facilitate collaborative learning can collect comparable data (Chandler et al., 2025), while advanced intelligent essay assessment systems can analyze and infer a student’s level of understanding (Pack et al., 2024). These technologies have the potential to capture a wide range of data throughout a student’s formal education journey (Mohammadi et al., 2025). This aligns with the long-standing recognition in the learning sciences of the importance of constructive assessment activities in fostering deeper learning (Black and Wiliam, 1998; Nicol and Macfarlane-Dick, 2006).

Moreover, such systems could extend beyond formal education settings to track participation in non-formal learning experiences, such as acquiring skills in music, crafts, or other hands-on disciplines, as well as informal learning opportunities, including language acquisition or cultural immersion (Volta and Di Stefano, 2024; Zhou et al., 2025). By integrating insights from these diverse learning contexts, AIED technologies can offer a more holistic understanding of a student’s knowledge, skills, and growth (Holmes et al., 2023).

In the light of this evidence, the current research is based on a shared vision of AI in Education as a powerful medium to co-create knowledge and experiences in class, both in relation with teachers and students. The relationship between AI, students and university teachers is seen through the lens of Computer Assisted Collaborative Learning approach (CACL) which, within educational contexts, involves the use of computers to facilitate and enhance collaborative learning processes. The main aim of introducing technology—in our case for AI enhanced learning experiences—is to help students learn effectively together by enabling them to exchange ideas, access information, and exchange feedback during problem-solving or authentic activities (Stahl et al., 2006; Thorsteinsson and Page, 2012).

Within this framework, our aim is to use Artificial Intelligence as an active agent in the teaching, learning, and assessment processes, grounded in a constructivist perspective that emphasizes the co-construction of knowledge. Rather than merely functioning as a technological tool, AI operates as an adaptive and interactive facilitator that supports learners in building understanding and

developing competences in collaboration with the peers and the teacher. In this sense, the integration of AI acquires particular significance for its potential to scaffold students’ learning within their Zone of Proximal Development (Agostini and Picasso, 2024; Vygotsky, 1966; Yousif, 2025), thereby reinforcing the idea of AI as a mediating presence that dynamically supports learners’ cognitive growth and self-regulated learning.

Despite the rapid diffusion of LLMs in higher education, the empirical picture is still uneven. Much of the available evidence focuses on individual adoption, attitudes, and self-reported practices, rather than on learning-sensitive designs that observe what changes in students’ work and learning processes (Acosta-Enriquez et al., 2024; Arum et al., 2025; Bektik et al., 2025). At the same time, empirical work on LLMs in collaborative settings is emerging, but it remains fragmented across contexts (e.g., ideation and programming), often relies on between-group comparisons or context-specific implementations, and rarely triangulates outcomes with artifact quality and process evidence (Kovari, 2025; La Scala et al., 2025; Shaer et al., 2024; Wang et al., 2025; Yan et al., 2025). This makes it difficult to conclude not only whether AI “works,” but also how it changes collaboration and learning, and to what extent effects are confounded by practice, task familiarity, or task sequence (Bektik et al., 2025; Kovari, 2025). This study addresses these gaps by examining LLM-as-peer collaboration in authentic small-group lesson-design tasks through a counterbalanced crossover design and by triangulating knowledge gains, artifact quality, and process evidence.

2 Theoretical framework: AI in education

The potential of AI in educational settings, as well as the necessity for AI literacy, places educators at the forefront of these new and exciting breakthroughs that were previously relegated to obscure computer science laboratories. At the same time, teachers and administrators are required to have clear perspectives on the potential of AI in education and, eventually, to incorporate this ground-breaking technology into their practice (Holmes and Tuomi, 2022).

To focus on the characteristics of Artificial Intelligence in Education (AIED) concept, Holmes et al. (2019) created a taxonomy for AIED systems, which is helpful to categorize tools and applications into three different but intersecting categories: (1) student-focused, (2) teacher-focused, and (3) institution-focused AIED.

Student-focused tools enable learning support through individualized guidance and feedback systems, which include intelligent tutors, dialogue-based tutors, AI-assisted applications and simulations, automated essay writing, chatbots, formative assessment tools, learning-network orchestrators, exploratory environments and lifelong learning assistants. Teacher-focused systems, on the other hand, improve educational processes through plagiarism detection, smart material selection, classroom observation, automated summative assessment, and AI-based teaching support and classroom management tools. Finally, the institution-focused systems concentrate on governance and operational management through their implementation of student

selection systems, course scheduling, school security, student risk and dropout identification and online proctoring tools.

Looking at the current study through the lenses of this classification, the focus is on Student-focused AIED. We may categorize it as Automatic Formative Assessment with the interface of a Chatbot, that we see as a category about the means of interactions more than about the scope of the AI employed.

2.1 AI in the loop

The concept of “human-in-the-loop” is a well-established framework that spans the intersection of computer science, cognitive science, and psychology (Wu et al., 2022). It represents a semi-supervised learning paradigm in which human input and machine learning systems work together to achieve optimal outcomes. This method leverages the precision and scalability of machine learning while incorporating the nuanced understanding, creativity, and adaptability of human intelligence. By integrating human oversight and intervention, the human-in-the-loop approach aims to enhance the accuracy of machine learning models while simultaneously supporting and enriching human learning experiences (Maadi et al., 2021, Monarch, 2021).

Notwithstanding this well-established framework, there is a new, growing, one that is thought to better address the ideal situation in most of the contexts of AI-Human collaboration: the AI-in-the-loop approach, that moves from an automation paradigm, where the human have a general oversight and some space of intervention on the AI-lead process, to a paradigm of collaboration and structured interaction between AI and human(s) (Natarajan et al., 2025; Sharma et al., 2023). This approach moves its steps from documents like the European Commission (2019b) and European Commission (2019a) that focused on the idea of the Human-Centric AI and the Human-in-Command (HIC) concepts, and gained traction in events like 2022 Stanford HAI’s conference titled “AI in the Loop: Humans Must Remain in Charge” where the general topic of discussion was to rethink AI systems where humans remain at the center of the decision-making (Lynch, 2022).

Unlike the traditional “Human-in-the-loop” model, where humans merely oversee an automated system, the “AI-in-the-loop” framework posits the human as the primary driver of the workflow. In this configuration, AI is not a replacement but a specialized tool integrated into the human process (Shneiderman, 2020). It is utilized selectively for specific computational tasks—such as processing data or interactions at a scale unmanageable by humans (Ebel et al., 2021)—while the human retains full ownership of the context, the direction, and the final decision-making. This approach ensures that AI complements human expertise rather than replacing it (Amershi et al., 2019).

In educational contexts, this framework can be particularly powerful, creating a synergistic relationship that enhances both teaching and learning. For example, LLMs, teachers, and students might interact within the same educational process: AI can handle repetitive or data-intensive tasks, such as grading or progress tracking, while teachers focus on fostering critical thinking, creativity, and emotional engagement. Students, in turn, benefit from personalized learning pathways and immediate feedback, all while being guided by the expertise and mentorship of their teachers. This triadic interaction between AI, educators, and

learners creates a dynamic and adaptive educational ecosystem that maximizes the strengths of each participant (Luckin and Holmes, 2016).

One has to consider not only the development of AI models but also the design of the interactions and behaviors that compose the human experience around the AI models (van Allen, 2018).

This leads to the definition of two new terms related to the relationship between humans and AI models that go beyond cooperation in learning, called “Usable AI” and “Useful AI” (Xu, 2019), which are fundamental to ensure that an AI models’ implementations are successful. Usable AI focuses on designing AI systems that are intuitive, accessible, and easy to interact with for the people who use them. Useful AI, on the other hand, takes a broader perspective. It goes beyond usability to ensure that AI systems are meaningful and beneficial to society as a whole. This approach considers the human conditions and contexts in which AI operates, striving to address real-world challenges and improve quality of life (Mosqueira-Rey et al., 2023; Xu, 2019).

Importantly, when an LLM is introduced into a collaborative task, it is not merely an additional resource: it becomes a new interaction partner that can reshape participation, co-regulation, and the division of labor within the group (Kovari, 2025). Recent studies of LLM-supported teamwork illustrate both the potential to enrich collaborative activity (e.g., expanding idea spaces in group ideation) and the risk of shifting work from co-construction to selection-and-editing when AI output becomes the “fastest voice in the room” (La Scala et al., 2025; Shaer et al., 2024). Related classroom-oriented quasi-experimental work also shows how LLM-based or agent-based support can change outcomes and perceived workload in collaborative programming contexts, reinforcing the need to conceptualize AI as a contributor to group cognition rather than a passive tool (Wang et al., 2025; Yan et al., 2025). This is where the AI-in-the-loop stance matters: the pedagogical design must ensure that humans remain epistemically responsible (contributing, interpreting, judging, and deciding), while AI remains a contingent contributor whose output is treated as fallible input for collective sense-making (Bektik et al., 2025). Framing ChatGPT as a “peer-like feedback provider” therefore operationalizes AI-in-the-loop in a way that is directly testable: it specifies an intended role (team member), an intended process (discussion and validation), and a clear point of tension (efficiency vs. collaborative meaning-making).

2.2 Large language models

Lately, Large Language Models (LLMs) have grown increasingly common in society and educational contexts. These AI-powered models can generate, analyze, and summarize text, as well as communicate dialogically with humans (Kasneji et al., 2023). Most of them are based on the General Pretrained Transformer (GPT) architecture, a model that looks at every word in context at once (self-attention) to figure out how they relate, then writes by guessing the next word statistically, given the input, using the probabilities it learned during training (Vaswani et al., 2017). Recently, other techniques add up to the GPT and, for example, enable multi-step inference at generation time through

internal computation (called test-time compute), giving the new models “reasoning” capabilities (Chen et al., 2024).

OpenAI’s ChatGPT, which is built on several GPT models, is one of the most well-known instances of LLMs. Anthropic’s Claude, Google’s Gemini and X-AI’s Grok are also prominent LLMs, delivered by the biggest technology players in the World scene. While these models are powerful, there are concerns about data privacy and results consistency (Chen et al., 2023). However, there are other options available. There are open-weight foundation models (such as Meta’s LLAMA, DeepSeek’s V3 and R1, OpenAI’s GPT-OSS, AliBaba’s Qwen and many others) that share weights, tokenizer and inference/fine-tune code, but not the pretraining data, with licenses that range from permissive (MIT/Apache-2.0) to custom restricted. These models can be used, customized, fine-tuned, or even trained specifically for one’s use case, allowing for greater flexibility and control (Martin et al., 2023). Such open-weights are very capable thanks to the advancement on the algorithmic and computational optimizations, and can run on powerful personal computers and servers of several institutions, that may, in this way, retain total control on the data, the behavior of the model several other aspects of the interactions. For example, OpenAI reports that GPT-OSS-20B (a model fittable on a laptop) matches or outperforms o3-mini (a 7-months-older reasoning model selectable in ChatGPT) on the same benchmarks, achieving even better results in competitive mathematics and medical reasoning despite its smaller size (OpenAI, 2025).

Generally speaking, LLMs can analyze massive amounts of text (in the order of the hundreds of thousands words for the average LLM) aggregate it, and then offer feedback based on previously established standards, rules and preferences (Tamkin et al., 2021). From an educational standpoint, the outcomes of that analysis can be applied to provide feedback to the student as well as to assist the instructor in evaluating the text. LLMs can support teachers and instructors in some time-related sustainability difficulties, enabling them, with the appropriate pedagogical and technical framework, to scale meaningful and authentic assessment and feedback approaches, where in the past generic feedback and structured tests (such as multiple choice) were the only sustainable ways.

Kasneji et al. (2023), for example, see opportunities for higher education students, where large language models assist in research and writing tasks, and in developing critical thinking and problem-solving skills. For teachers they envision LLMs used to create personalized learning experiences. These models, in general, are capable of analyzing students’ writing and responses, offering individualized feedback, and suggesting instructional materials that match their particular learning requirements. This type of support has the potential to economize teachers’ time and effort in crafting personalized materials and feedback, allowing them to focus on other aspects of teaching, such as creating engaging and interactive lessons (Kasneji et al., 2023).

2.3 AI and feedback: promises and challenges

According to Lipnevich et al. (2016) and Lipnevich and Smith (2018), instructional feedback is any information related to a performance that students can use to improve their performance

or learning. The most effective feedback bridges the gap between where a student currently is, the state of the goal they are aiming for and the steps they need to take to get from the current state to the desired state (Hattie and Timperley, 2007). Shute (2008) also argues that feedback should help reduce uncertainty between performance and goal, as well as being supportive, timely, non-evaluative and specific. Feedback might come from any source; for a long time literature has focused on teachers’ feedback, which remains a valuable resource, but the recent studies on feedback have broadened this vision with a deeper focus on other sources such as peers, or the task itself.

Hattie and Timperley (2007) suggested that feedback, to support the student learning process, should explicitly clarify expected learning outcomes to students, in order for the feedback process to be effective and truly “for learning;” promote awareness, providing information on where the learner is on the path taken, what progress he/she has made and what is lacking to achieve the set goals and indicate strategies for bridging the gap. In other words, feedback must be provided through strategies that can support the learner in achieving the expected learning outcomes, providing useful information to bridge the gap in knowledge and competences detected (Hattie and Timperley, 2007; Wisniewski et al., 2020).

Feedback can differ regarding three main elements. The medium through which feedback is proposed changes according to the context and the task; it can be in the form of written comments, or through the aid of software and learning platforms, or feedback can be proposed orally (Hattie and Timperley, 2007). The feedback can be differentiated regarding the recipient: it can in fact be individual, thus provided directly to the author of the task/product, or group-based, thus given to the entire class, or to a group, with respect to the critical aspects detected by the instructor or by peers (Nicol and Macfarlane-Dick, 2006). Another fundamental aspect is the source of the feedback, which may be external or internal. External therefore provided by the teacher or by the student’s peers; while internal is understood as feedback self-generated by the subject him/herself through processes of reflection and self-assessment (Carless and Boud, 2018; Nicol and Macfarlane-Dick, 2006).

Nicol (2021) theorizes the idea of “inner feedback” as an internal generative process through which knowledge and understanding of the discipline are constructed through evaluative action and the formulation of judgment. Literature has shown that any feedback, if processed by students, becomes self-feedback (To et al., 2023; Panadero et al., 2019; Tomazin et al., 2023). Students produce self-feedback by evaluating their ongoing work against various reference points, such as teacher feedback, peer feedback, technology-generated feedback, rubrics, exemplars, or instructional materials. This allows them to assess how well they are meeting task goals and their performance level, decide whether to seek academic support, and refine their goals, strategies, and efforts for future tasks (Nicol, 2021).

By generating feedback on their own, students enhance their awareness of their strengths and areas needing improvement while also cultivating crucial metacognitive skills that foster lifelong learning and academic achievement (Lipnevich and Smith, 2022; Nicol and Kushwah, 2024).

Lee and Moore (2024) in their systematic review present findings which demonstrate that AI delivers varied feedback

across different contexts, serving a range of instructional purposes. By automating routine grading and feedback tasks, AI systems alleviate instructor workload, enabling educators to concentrate on more complex teaching responsibilities with enhanced efficiency (i.e., identifying gaps, reducing cognitive load, and correcting information). From this study it seems that integrating AI-powered feedback systems enhances educational outcomes while creating a more effective and inclusive experience for both students and instructors.

An interesting perspective is that one proposed in some studies, where the AI is integrated in students' work as a tutor/agent that can act as an intermediary among students, teaching assistants, and lecturer (Hobert and Berens, 2023). Moreover, recent studies show (Zheng et al., 2024) the efficacy of AI-driven feedback and feedforward approach to enhance collaborative knowledge construction, coregulated behaviors, and group performance. When it comes to students' performance, research findings are still contradictory, as it is a very new research field. Escalante et al. (2023) reported findings from a writing task for university-level English as a New Language (ENL) learners over a 6-week period, where the experimental group received writing feedback from ChatGPT, while the control group received feedback from a human tutor. In this study there was no significant difference in learning outcomes between the two groups and the authors claim for a blended approach that maximizes the strengths of both feedback types.

3 The current study

This quasi-experimental study contributes to research on pedagogical approaches and technologies that strengthen students' self-feedback generation and support authentic learning. Departing from models that position AI as a tutor or authority, we implemented ChatGPT as a collaborative group member within a constructivist task. In this role, the AI offered information and preliminary feedback for students to analyze and interpret together, encouraging critical engagement with AI-generated content and promoting collective sense-making. Framing AI as a peer was intended to enhance students' ability to evaluate diverse inputs, generate meaningful self-feedback, and build critical thinking skills—competencies essential in emerging human–AI collaborative contexts.

This study advances current research on GenAI in collaborative learning in three ways that directly respond to limitations highlighted in recent syntheses and to the methodological fragmentation of the field (Bektik et al., 2025; Kovari, 2025). First, we examine LLM use in an authentic, discipline-relevant design setting (collaborative lesson design by pre-service teachers), where the key outcomes include both individual knowledge development and the quality of group-produced artifacts. Second, we adopt a counterbalanced crossover structure (AI in Task 1 vs. AI in Task 2), which helps disentangle AI effects from task order, practice, and familiarity—issues that can otherwise complicate interpretation when studies rely on a single sequence or simple between-group comparisons (Wang et al., 2025; Yan et al., 2025). Third, we triangulate outcomes with mixed-methods evidence, combining repeated knowledge tests, expert rubric

ratings of artifacts, and qualitative reflection/process data, thereby connecting “whether it works” claims to plausible mechanisms of change that remain under-specified in much of the emerging collaborative GenAI literature (La Scala et al., 2025; Shaer et al., 2024).

The study's primary aim is to examine whether, and how, ChatGPT influences the design and implementation of learning activities for a specific student group. By analyzing its use in collaborative tasks, we explore how ChatGPT may enrich learning processes, support decision-making, and contribute to more effective educational experiences.

RQ1. What is the impact of using ChatGPT as a peer-like feedback provider on the acquisition of knowledge and skills during collaborative activities, particularly in the execution of specific tasks?

RQ2. How do the outcomes differ between groups that use ChatGPT at different stages?

4 Materials and methods

4.1 Participants

102 post-grad students, perspective special education teachers (82 women, 15 men, and 5 with another gender identity) enrolled in a course of instructional design with technology for inclusion, within a 1-year program of specialization and habilitation for special education teachers. The age groups distribution was: 5 students from 23 to 26 y.o. (4.9%), 66 from 27 to 42 y.o. (64.7%), and 31 from 43 to 58 y.o. (30.4%). The estimated mean age was 38.87 years (SD = 8.01).

They were divided into groups of 4–6 students based on spatial proximity, resulting in 21 groups, 8 of which were composed exclusively of women.

4.2 Study design and procedures

The study employed a quasi-experimental, counter-balanced crossover design, as illustrated in the study flow chart (Figure 1). The 102 participants, organized into 21 clustered groups, engaged in two sequential, authentic tasks over the course of the experiment.

A key procedural instruction for all groups, regardless of their condition, was to treat ChatGPT 3.5 as a collaborative “team member” and to keep brief notes for their own use about how they used it (e.g., the kinds of questions asked and how outputs informed discussion), to support the post-task written reflection. The experimental manipulation involved controlling which task was AI-assisted. The 21 groups were randomly assigned to one of two conditions for the first task:

- Experimental Group 1 (GT1): Used ChatGPT for Task 1 and worked without it for Task 2 (hence the name GT1: Group that used ChatGPT at Task 1).
- Experimental Group 2 (GT2): Worked without ChatGPT for Task 1 and used it for Task 2.
- The experimental procedure unfolded across four main phases that took place contiguously in a single session of 240 min:

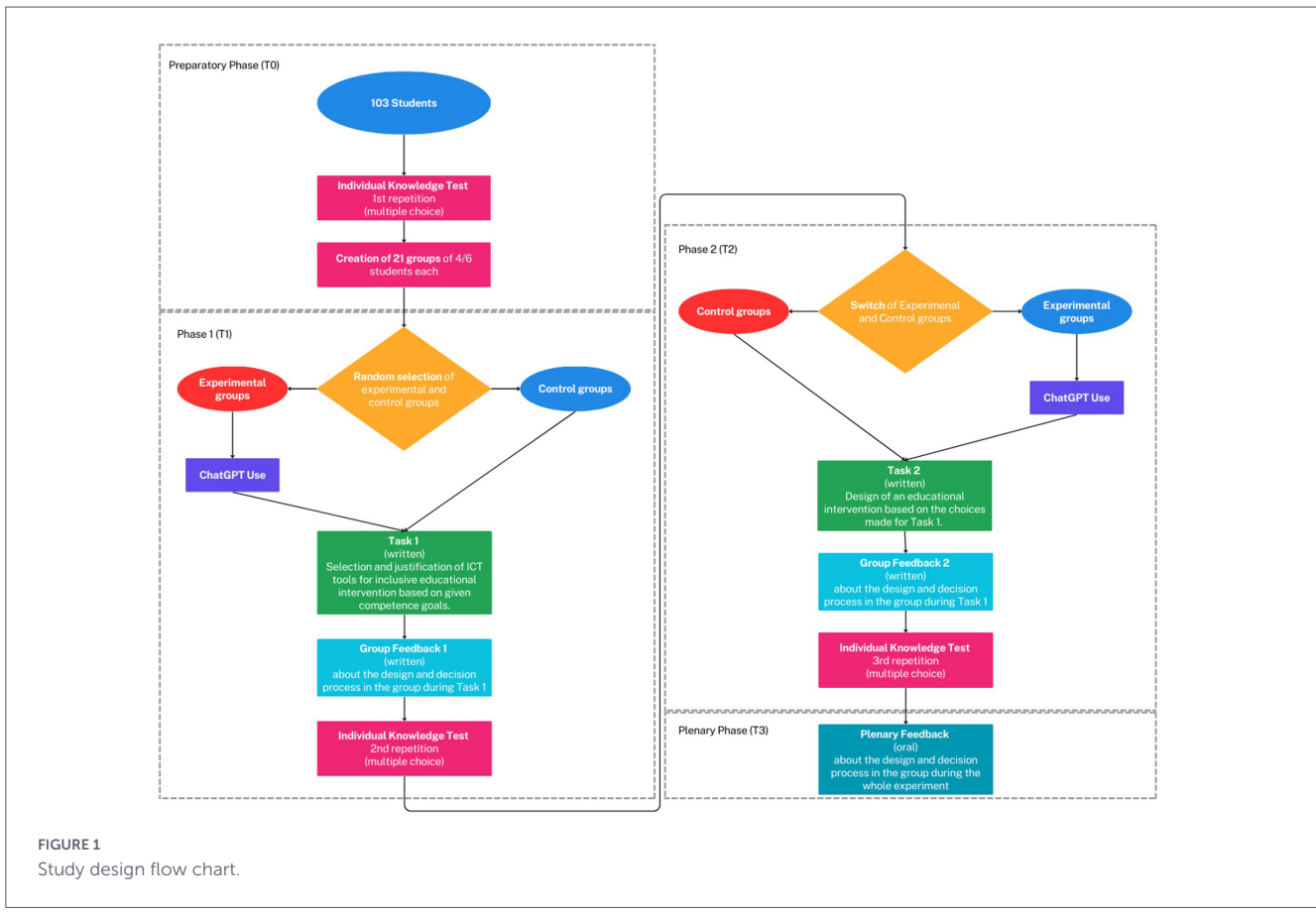


FIGURE 1 Study design flow chart.

1. Pre-Experiment (T0 – 0–15 min): Before any group work began, all participants completed an individual knowledge pre-test.
2. Task 1 Phase (T1 – 15–120 min): Groups worked collaboratively on Task 1 (Selection and justification of ICT tools). Upon completion, and each group submitted written feedback reflecting on their process, and all participants, in the last 15 min, took the individual knowledge test the second time.
3. Task 2 Phase (T2 – 120–225 min): Following the crossover of conditions, groups proceeded to Task 2 (Design of an educational intervention). Afterwards, a second round of group written feedback was collected, and, in the last 15 min, participants completed the final individual knowledge test.
4. Plenary Phase (T3 – 225–240 min): The study concluded with a plenary session where students shared reflections on the entire process with the whole class, moderated by the teacher.

The specific instruments used for the knowledge tests, task evaluation, and feedback collection are detailed in the following section.

4.3 Measures and tasks

To address our research questions, we collected data using three primary instruments: an individual multiple-choice knowledge test to measure learning gains, group-produced artifacts to assess performance, and written feedback to understand the collaborative

process. The multiple-choice test assesses conceptual knowledge; skill-related outcomes are operationalized via artifacts (products of Task 1 and Task 2) quality/rubric and qualitative evidence.

A custom-designed multiple-choice knowledge test was administered to all participants at three time points (T0, T1, and T2) to assess conceptual knowledge aligned with the course learning objectives. The instrument included a short demographic section followed by 13 multiple-choice items: seven focused on educational technology (D1–D7) and six on teaching methodology and instructional design (D8–D13). Item cognitive demand was characterized using Anderson and Krathwohl’s revised Bloom taxonomy (Anderson and Krathwohl, 2001). Consistent with the test’s purpose as a brief content-sampling knowledge check, items predominantly target the lower cognitive processes of Remembering and Understanding (e.g., recognizing key principles, or matching an instructional goal to an appropriate technological or methodological choice), rather than scenario-based application or higher-order reasoning. A full item-by-item mapping to the taxonomy and content domains is provided in [Supplementary Data Sheet 1](#).

Each item presented five response options plus an “I don’t know” option to discourage random guessing; for scoring, responses were coded dichotomously as correct versus non-correct. Internal consistency (KR-20/Cronbach’s alpha on dichotomous scoring) was modest across time points (T0 $\alpha = 0.387$, T1 $\alpha = 0.320$, T2 $\alpha = 0.278$) and did not improve when splitting items by content domain (technology: $\alpha = 0.347/0.230/0.097$; methodology: $\alpha = 0.090/0.142/0.131$). Given the instrument’s intentionally

heterogeneous content coverage and its use as a knowledge check rather than a unidimensional scale, these coefficients are reported just as a description; nonetheless, we complement them with item-level statistics (e.g., difficulty/discrimination) reported in [Supplementary Data Sheet 1](#). As a robustness check, we repeated the Wilcoxon analyses excluding the only item showing negative item-total correlations at multiple waves (D11); the substantive pattern of results was unchanged (see [Supplementary Data Sheet 1](#)).

The core of the experiment involved two authentic tasks performed in groups:

- Task 1 where participants were asked to define the pedagogical objectives of a proposed teaching intervention and to select suitable inclusive technologies to achieve those aims, offering a written rationale for their selections (see [Table 1](#)). Instructions for Task 1 were as follows (translated from Italian): “Identify a learning objective (conducive to a competence) to be developed. Select a digital technology to support that objective, and justify (a) why this technology is appropriate for the objective, and (b) why and how its use is inclusive.”
- Task 2 where participants were asked to use those selected technologies to design a detailed, ready to use, educational intervention (see [Table 1](#)). Instructions for Task 2 were as follows (translated from Italian): “Develop the full design of the inclusive lesson/intervention: target group, learning objectives, school level, class, title, contents, activities, tools/technologies, and any other relevant elements. Provide a well-structured lesson plan.”
- Groups using ChatGPT in either Task had the additional instruction: “During both tasks, treat and address ChatGPT as if it were a group member. Use it in a peer-like way: keep it in the loop of the conversation, ask its view and suggestions when needed, challenge or ask for clarification of its proposals, interact with the answers as you would do with a colleague.”

The final written artifacts (i.e., the justifications and the lesson designs) produced by each group served as the primary measure of task performance.

To capture insights into the collaborative experience, each group was required to submit written, semi-structured feedback after completing both Task 1 and Task 2 with the following instructions (translated from the Italian): “Describe how you worked as a group. If you used ChatGPT, describe how you used it and your impressions of the group dynamics with ChatGPT as a peer. If you did not use ChatGPT, describe how you worked among yourselves. Please address: (i) how we organized the work; (ii) strengths; (iii) weaknesses.”

The open-ended prompts were designed to elicit reflections on their group’s internal dynamics, decision-making processes, the perceived role and influence of ChatGPT (in the experimental condition), and their overall design experience. This qualitative data was essential for understanding the students’ workflow and their evolving perceptions of AI-human collaboration. We prioritized groups’ immediate post-task feedback/reflections because our primary interest was the collaborative meaning-making process (e.g. how prompts and AI outputs were negotiated, interpreted, and integrated into the team’s work) and how these dynamics related to individual knowledge development and group-produced artifact quality; these links would have been only partially observable from interaction transcripts alone.

The quality of the group-produced artifacts from Task 1 and Task 2 was assessed by two independent expert evaluators using the rubric detailed in [Table 1](#). The instrument used two distinct criteria for each task to align with its specific objectives, plus a transversal one. For Task 1, criteria included “Depth of Reflection,” “Argumentation,” and “Originality.” For Task 2, criteria were “Completeness,” “Relevance of Methodological Choices,” and “Originality.” Each criterion was rated on a three-level scale, yielding a maximum score of 9 points per evaluator for each task.

TABLE 1 Evaluation rubrics of Task 1 and Task 2 artifacts.

Dimension	Level 1	Level 2	Level 3
Task 1: reflection and argumentation			
Depth of reflection (C1)	The group develops disconnected reflection points	The group develops deep reflection with disconnected themes	The group develops deep reflection with the ability to connect themes
Argumentation (C2)	The group partially argues the themes emerging from brainstorming with basic vocabulary	The group argues the themes emerging from brainstorming with adequate vocabulary	The group thoroughly argues the themes emerging from brainstorming with rich and detailed vocabulary
Originality (C3)	The group proposes the introduction of limited approaches/tools, without an innovation perspective	The group proposes the introduction of some approaches/tools with a partially innovative perspective	The group proposes the introduction of multiple tools and approaches through innovative thinking
Task 2: design and implementation			
Completeness (C1)	The group develops a design with limited activities described summarily	The group develops a design with adequate activities well described and articulated	The group develops a rich design in terms of proposed activities, which are described comprehensively
Methodological choices relevance (C2)	The group selects methodologies and tools that are barely relevant to the objectives	The group selects methodologies and tools partially aligned with the objectives	The group selects methodologies and tools fully aligned with the objectives
Originality (C3)	The group proposes the introduction of limited approaches/tools, without an innovation perspective	The group proposes the introduction of some approaches/tools with a partially innovative perspective	The group proposes the introduction of multiple tools and approaches through innovative thinking

Originality (C3) intentionally uses identical descriptors across Task 1 and Task 2 because we treat it as a transversal learning-design quality (i.e., the extent to which groups propose and integrate approaches and tools in an innovative way). See [Boxes 1, 2](#) for task-specific anchors.

The scores from both evaluators were summed, resulting in a final task score out of a possible 18 points.

While the two tasks produced different artifacts (tool selection/justification vs. full lesson design), originality was conceptualized as a transversal construct capturing the breadth and novelty of approaches/tools introduced in the group product. Task-specific instantiations of originality are illustrated in the rubric anchors (Boxes 1,2), which show how the same construct was interpreted for tool selection/justification (Task 1) versus lesson-plan integration (Task 2).

To improve interpretability of the rubrics, we include de-identified anchoring excerpts illustrating typical Level 1 versus Level 3 performance for Task 1 and Task 2 (see Boxes 1,2).

4.4 Data analysis

To answer our research questions, we employed the following approach. Quantitative data from the knowledge tests and task evaluation rubrics were analyzed using JASP statistical software. Qualitative data from the group feedback sessions were analyzed using thematic content analysis in Atlas.ti. The analysis was structured to directly address our two research questions.

4.4.1 RQ1: the impact of ChatGPT on knowledge and skills acquisition

Research Question 1 sought to understand the impact of using ChatGPT as a peer-like feedback provider on the acquisition of knowledge and skills. This was assessed using both quantitative and qualitative data.

- **Quantitative Analysis of Knowledge Acquisition:** To measure changes in students' content knowledge, we analyzed scores from the 13-item multiple-choice knowledge test administered at three time points (T0, T1, T2). Each score was computed as the number of correct responses (0–13), with non-correct responses (including “I don't know” ones) scored as 0. Because the test was designed as a brief content-sampling check across heterogeneous course topics rather than a unidimensional scale, internal consistency coefficients are reported descriptively and item-level difficulty/discrimination statistics are provided in [Supplementary Data Sheet 1](#). For each counterbalanced sequence (GT1 and GT2), we tested within-group changes using Wilcoxon signed-rank tests between the adjacent phases relevant to that sequence (e.g., T0 vs. T1 for GT1; T1 vs. T2 for GT2). This non-parametric approach was chosen due to deviations from normality, and scores that are bounded and discrete. Effect sizes were quantified using the rank-biserial correlation (r_{rb}). These analyses estimate whether performance on the knowledge check changes after the phase in which students had access to AI and whether performance is maintained in the subsequent phase. In addition to the within-sequence Wilcoxon tests, we performed a model-based crossover validity check using interval change scores ($\Delta T1-T0$ and $\Delta T2-T1$). We conducted a 2×2 mixed-design ANOVA on interval change scores, with Treatment (AI vs. no-AI interval; within-subject) and Sequence (GT1 vs. GT2;

between-subject), focusing on the Treatment \times Sequence interaction as the explicit test of order/carryover dependence.

- **Qualitative Analysis of Skill and Process Impact:** To explore the perceived impact on skills, collaborative processes, and student experience, we conducted a thematic content analysis of the written group feedback. Using a bottom-up, inductive coding approach in Atlas.ti, we identified key themes related to students' reflections on AI use, its effect on group work dynamics, and the development of specific skills like prompting and critical evaluation. This analysis provided rich, contextual insights into how and why ChatGPT influenced the learning process. Second, the oral feedback from the final plenary session (T3) was transcribed and reviewed. While not formally coded in the same manner, key quotes and overarching themes from this session were used as a standpoint to triangulate and interpret the findings from the written feedback and quantitative results in our Discussion section. In the qualitative results section, direct quotations from student groups are cited using a (Code:Quote) format, where the first number corresponds to the code from [Table 8](#) and the second is a unique identifier for the quotation.

4.4.2 RQ2: differences in outcomes between groups

Research Question 2 aimed to determine how outcomes differed between groups that used ChatGPT during the initial task (GT1) versus the design task (GT2). This was evaluated by comparing the quality of their final products and their described processes. Our analyses are design-aligned and deliberately conservative: bounded rubric scores and a modest number of independent groups argue for simple contrasts with minimal assumptions.

- **Quantitative Analysis of Task Outcomes:** The quality of groups' work on Tasks 1 and 2 was assessed with the evaluation rubrics ([Table 1](#)) by two independent expert raters. For each group and task, we computed a total product score by summing the three rubric dimensions and summing the two raters' scores. Given the small sample size and the limited/ordinal nature of the scoring scale, our primary inferential approach was non-parametric. We analyzed (i) the within-group change from Task 1 to Task 2 using Wilcoxon signed-rank tests (overall and within each sequence). Effect sizes are reported as paired-samples Cohen's *d*. For transparency, we also report a 2×2 mixed ANOVA (Task as within-group factor; sequence as between-group factor) as a robustness check, while retaining the non-parametric analyses as the primary basis for inference.
- **Qualitative Analysis of Workflow and Process Differences:** We further analyzed the coded qualitative data by comparing themes that emerged from groups in the GT1 condition versus the GT2 condition. This involved examining how the timing of AI integration influenced reported workflows, the nature of collaboration, the challenges faced, and the perceived trade-offs between efficiency and creativity. This comparative analysis allowed us to understand the different experiential

outcomes of using ChatGPT at different stages of the design process.

knowledge test's overall score ($p = 0.023$, $r_{rb} = -0.371$) and in the technological part ($p = 0.022$, $r_{rb} = -0.442$).

5 Results

Data from 82 students who completed all three tests were included in the final analysis. Of these, 38 were in the GT1 condition and 44 were in the GT2 condition.

5.1 Quantitative analysis

5.1.1 GT1 knowledge test results

Results of the students using ChatGPT during the first group task (GT1) (Figure 2 and Table 2) suggest that they had a significant increment in the knowledge test performance after this task ($p = 0.015$, $r_{rb} = -0.429$) and in particular in the technological part ($p = 0.001$, $r_{rb} = -0.692$).

As shown in Figure 2 and Table 3, the results of the knowledge test administered after Task 1 and the one after Task 2, where Group GT1 was not permitted to use ChatGPT, remained essentially stable. This suggests that the absence of ChatGPT in Task 2 did not significantly impact the knowledge retention or performance of GT1, indicating consistency in their learning outcomes across the two tasks.

5.1.2 GT2 knowledge test results

The results for students who used ChatGPT during the second group task (GT2) (Figure 2 and Table 4) indicate that their performance on the knowledge test remained essentially stable between the pre-test (T0) and the post-test after the first group task (T1). This is to be expected since they did not use ChatGPT during the first task.

On the other hand, after the second group task, when they had to use ChatGPT as a group member, the results (Figure 2 and Table 5) suggest that learners had a significant increment in the

5.1.3 Robustness check: sequence x treatment interaction (knowledge test)

To explicitly test whether the AI-related change depended on exposure order, we conducted a 2×2 mixed-design ANOVA on interval change scores ($\Delta T1-T0$ and $\Delta T2-T1$), with Treatment (AI interval vs. no-AI interval) as a within-subject factor and Sequence (GT1 vs. GT2) as a between-subject factor (Table 6). The Treatment \times Sequence interaction was not significant for the overall knowledge score or for either sub-score (all $ps \geq 0.387$). The Treatment main effect was significant for the technological sub-score, whereas it was not significant for the overall score or the methodological sub-score (Table 6).

5.1.4 Work groups product analysis

To compare the quality of educational designs across groups, each artifact was scored independently by two expert raters using the three-dimension rubrics (Table 1). For each group and task, we computed a total product score by summing the three dimension of each rubric and summing the two raters' scores [$J_total = (J1_C1+J1_C2+J1_C3) + (J2_C1+J2_C2+J2_C3)$]; theoretical range 6–18]. Interrater reliability was estimated using a two-way random-effects intraclass correlation coefficient with absolute agreement [ICC(2,1)]. Rubric anchors and representative artifact excerpts are illustrated in Boxes 1, Box 2. Reliability for the summed rubric score was excellent for both Task 1 [ICC(2,1) = 0.934; ICC(2,2) = 0.966] and Task 2 [ICC(2,1) = 0.944; ICC(2,2) = 0.971], supporting the use of the sum of the two assessors' scores in subsequent analyses.

Given the small sample and the limited/ordinal nature of the scoring scale the primary inferential analyses were non-parametric. We first examined overall change from Task 1 to Task 2 using a Wilcoxon signed-rank test on paired group scores, which indicated a reliable increase in Task 2 relative to Task 1 (Table 7B). Planned within-sequence Wilcoxon tests showed that the early-AI (GPT in Task 1) sequence did not exhibit evidence of a decrease when ChatGPT was removed in Task 2 (two-sided $p = 0.236$; one-tailed test for decrease $p = 0.882$; $d = 0.43$), whereas the late-AI (GPT in Task 2) sequence showed a positive Task1→Task2 change that was supported under a directional hypothesis (two-sided $p = 0.056$; one-tailed test for increase $p = 0.028$; $d = 0.64$) (Figure 3 and Table 7).

For transparency and as a robustness check, we additionally report a 2×2 mixed ANOVA with Task (Task 1 vs. Task 2) as a within-group factor and sequence (early-AI vs. late-AI) as a between-group factor. The ANOVA showed a main effect of Task, $F(1, 19) = 6.33$, $p = 0.021$, but neither a main effect of sequence, $F(1, 19) = 0.27$, $p = 0.612$, nor a Task \times sequence interaction, $F(1, 19) = 0.08$, $p = 0.776$ (Table 7C). Interpretation of these effects is presented in the Discussion.

Finally, the final product quality in Task 2 did not differ between sequences [early-AI: $M = 14.75$, $SD = 2.92$; late-AI: $M = 14.38$, $SD = 4.43$; Welch's $t(18.84) = 0.23$, $p = 0.82$, $d = 0.09$], indicating comparable end performance regardless of whether ChatGPT had been used in Task 1 or Task 2 (Table 8).

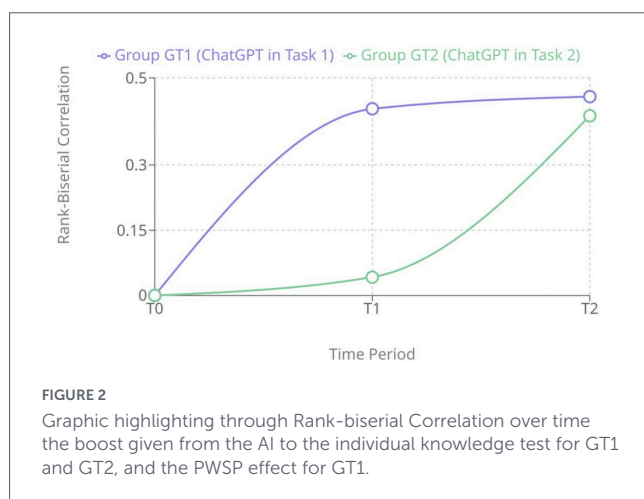


TABLE 2 Paired samples Wilcoxon signed-rank test for knowledge scores in Group 1 (GT1) before and after Task 1.

Measure	W	Z	p	Rank-biserial correlation (r_{rb})	95% CI for r_{rb}
T0 vs. T1					
Total knowledge score	170.00	-2.18	0.015	-0.43	(-0.80, -0.06)
Technology knowledge subscore	54.00	-3.09	0.001	-0.69	(-1.12, -0.26)
Methodology knowledge subscore	110.00	-0.85	0.200	-0.20	(-0.65,0.25)

This analysis is for Group 1 (GT1; $n = 38$), which used ChatGPT for Task 1. T0, Pre-test scores; T1, Post-Task 1 scores. Technology and Methodology scores are subscales of the total knowledge test. The test is one-tailed, hypothesizing that scores at T1 are higher than at T0. The column “df” from the original was removed as it is not typically reported for Wilcoxon tests; n is provided instead. p values are rounded to three decimal places.

TABLE 3 Paired samples Wilcoxon signed-rank test for knowledge scores in Group 1 (GT1) before and after task 2.

Measure	W	z	p	Rank-biserial correlation (r_{rb})	95% CI for r_{rb}
T1 vs. T2					
Total knowledge score	226.00	-0.134	0.451	-0.028	(-0.43,0.38)
Technology knowledge subscore	142.50	0.137	0.561	0.033	(-0.43,0.49)
Methodology knowledge subscore	86.00	-0.709	0.244	-0.181	(-0.67,0.31)

This analysis is for Group 1 (GT1; $n = 38$), which used ChatGPT for Task 1. T1, Post-Task 1 scores; T2, Post-Task 2 scores. Technology and Methodology scores are subscales of the total knowledge test. The test is one-tailed, hypothesizing that scores at T2 are higher than at T1.

TABLE 4 Paired samples Wilcoxon signed-rank test for knowledge scores in group 2 (GT2) before and after Task 1.

Measure	W	z	p	Rank-biserial correlation (r_{rb})	95% CI for r_{rb}
T0 vs. T1					
Total knowledge score	373.50	-0.230	0.412	-0.042	(-0.40, 0.31)
Technology knowledge subscore	244.50	0.247	0.602	0.052	(-0.035, 0.46)
Methodology knowledge subscore	176.00	-0.897	0.186	-0.191	(-0.60, 0.22)

This analysis is for Group 2 (GT2; $n = 45$), which used ChatGPT for Task 2. T0, Pre-test scores; T1, Post-Task 1 scores. Technology and Methodology scores are subscales of the total knowledge test. The test is one-tailed, hypothesizing that scores at T1 are higher than at T0.

TABLE 5 Paired samples Wilcoxon signed-rank test for knowledge scores in Group 2 (GT2) before and after Task 2.

Measure	W	z	p	Rank-biserial correlation (r_{rb})	95% CI for r_{rb}
T1 vs. T2					
Total knowledge score	233.00	-1.994	0.023	-0.371	(-0.73, -0.01)
Technology knowledge subscore	105.50	-2.006	.022	-0.442	(-0.87, -0.02)
Methodology knowledge subscore	174.00	-0.360	.363	-0.079	(-0.50,0.35)

This analysis is for Group 2 (GT2; $n = 45$), which used ChatGPT for Task 2. T1, Post-Task 1 scores; T2, Post-Task 2 scores. Technology and Methodology scores are subscales of the total knowledge test. The test is one-tailed, hypothesizing that scores at T2 are higher than at T1.

TABLE 6 Robustness check (2 × 2 mixed ANOVA on interval change scores).

Outcome	Δ during AI interval, M (SD)	Δ during no-AI interval, M (SD)	Treatment main effect	Treatment × Sequence (carryover/order check)
Total (13 items)	0.044 (0.171)	0.011 (0.167)	$F(1, 82) = 0.95, p = 0.333, \eta_p^2 = 0.011$	$F(1, 81) = 0.76, p = 0.387, \eta_p^2 = 0.009$
Technological (7 items)	0.079 (0.215)	-0.003 (0.201)	$F(1, 82) = 4.28, p = 0.042, \eta_p^2 = 0.050$	$F(1, 81) = 0.50, p = 0.480, \eta_p^2 = 0.006$
Methodological (6 items)	0.009 (0.211)	0.026 (0.201)	$F(1, 82) = 0.20, p = 0.656, \eta_p^2 = 0.002$	$F(1, 81) = 0.65, p = 0.422, \eta_p^2 = 0.008$

Δ scores are computed per participant for the two adjacent intervals ($\Delta T1-T0$ and $\Delta T2-T1$). For each participant, the AI interval is the interval corresponding to their sequence (GT1: T0→T1; GT2: T1→T2), and the no-AI interval is the other interval. The Treatment × Sequence interaction tests whether the AI-related change differs by order (i.e., order-dependence/carryover). $N = 83$ (GT1 = 38, GT2 = 45).

Box 1 Representative artifact excerpts and Task 1 rubric anchors.

Notes: Excerpts are translated from Italian and lightly edited for readability. The first excerpt is from a group that received Level 1 on all three Phase 1 criteria; the second excerpt is from a group that received Level 3 on all three criteria.

Anchor A—Level 1 across criteria (Phase 1/Task 1)

“We chose to develop the competence of preparing and deliver an oral presentation or a debate on an advanced topic, using multimedia aids because it helps develop European key competences and citizenship competences.”

“We chose Canva because it is easy to use, accessible, inclusive, intuitive, free, and already known by most students. Using Canva supports debate.”

“We proposed Canva because it has already been used in class and received positive feedback from students.”

Why this corresponds to Level 1:

- Depth of reflection (L1): The reflection consists of brief statements with limited elaboration and weak integration across ideas.
- Argumentation (L1): Justification remains generic (e.g., “easy,” “inclusive,” “already used”) with minimal supporting reasoning and limited conceptual detail.
- Originality (L1): The proposal centers on a single familiar tool previously used in class, with little evidence of a broader or innovative design perspective.

The product rated by both judges is not strictly incorrect but, in relation to the specific rubric criteria, it presents fragmented and poorly connected ideas, develops the identified themes only superficially using limited language, and proposes few basic approaches or tools without demonstrating innovation or critical reflection.

Anchor B—Level 3 across criteria (Phase 1/Task 1)

Competence goal: preparing and delivering an oral presentation or debate on an advanced topic using multimedia supports.

Topic: “The death penalty: pros and cons” (interdisciplinary: history, philosophy, law, civic education).

“To select the most suitable technologies, we relied on ChatGPT. We asked both for supporting materials for the two positions and for multimedia technologies to use.”

“Based on its suggestions, we selected: (i) Google Forms for a diagnostic pre-test of prior knowledge; (ii) PowerPoint and video to support information and argument construction; (iii) debating platforms for the debate itself; (iv) a final online poll to examine whether students changed their initial view after engaging with new information.”

(Continued)

BOX 1 (Continued)

Why this corresponds to Level 3:

- Depth of reflection (L3): The submission provides a coherent rationale connecting the competence goal, topic choice, and a sequenced set of learning and assessment activities.
- Argumentation (L3): The design decisions are justified in a specific and detailed way (diagnostic → multimedia support → debate tools → post-debate polling), using richer and more precise language than generic tool claims.
- Originality (L3): The group integrates multiple tools and approaches into an innovative instructional sequence (interdisciplinary framing, diagnostic + formative elements, and change-of-mind check).

This Level 3, connected to the judges' assessment procedure, is related to a product which demonstrates a high level of elaboration, with deep and well-connected reflections, thorough and well-argued development of the brainstorming themes using rich language, and the proposal of multiple innovative tools and approaches.

5.2 Qualitative analysis

A qualitative analysis was carried out to explore the relationship developed between the use of ChatGPT and its influence on students' work (Table 9).

5.2.1 Students' feedback

This feedback was collected via written, open-ended prompts submitted by each group after the completion of Task 1 and Task 2. Students proposed general reflection on the use of AI during their formative activities in class: specifically, they reflected on the group dynamics during the design activity and the AI use during the process. For example students from the Group GT1 affirmed that "Unlike the previous work, in this case we did not use any research technology, nor did we feel the lack of it. There was an interaction based on sharing personal experiences" (Code 2: Citation 1, from now on 2:1) and again in relation to the specific use of ChatGPT "In this second phase of work, the group benefited from a broader and more personal sharing of the various points elaborated in the outline, as they did not use ChatGPT" (2:2).

They highlighted both pros and cons related to the use of AI: for example, they underlined the power of ChatGPT's response that allowed them to reduce the discussion time by making the choice concrete, in fact, "the use of ChatGPT facilitated the work in terms of timing and structuring the learning units/modules" (1:7). Specifically, the use of ChatGPT facilitated the work in terms of timing and structuring the teaching design: in general, students affirmed that "We found it difficult at first to understand how to ask ChatGPT questions. ChatGPT's response allowed us to reduce the discussion time by making the choice concrete. It was also difficult to stay within the allotted time because of the different degrees of precision each of us wanted to use in the work. The enthusiasm in designing the activity grew during the processing as did the satisfaction in using AI" (1:2).

In connection with the previous reflection theme concerning tool management, students emphasized the importance of carefully selecting and planning the activities suggested by ChatGPT. They highlighted the need to iteratively create and revise prompts, as well as critically evaluate the generated outputs, in order to obtain accurate and useful information for instructional design tasks. On the negative side, students reported that the use of

AI during group activities reduced opportunities for discussion and debate. Additionally, they initially found it challenging to formulate effective prompts and understand how to interact with ChatGPT productively.

For the category "AI use modalities," in relation to the code "ChatGPT as "oracle" for selecting TIC," groups use ChatGPT as a collaborative information partner to sustain the selection process of the most appropriate technology to be implemented in their formative design project, to achieve the learning outcomes. For example, "after choosing the objective, we asked ChatGPT which was the most effective tool to achieve the objective. After collecting ChatGPT feedback, after a comparison, we chose the most suitable tool for our initial objective" (4:1). To better emphasize ChatGPT use as a collaborative information partner, for example it is possible to report a strategy adopted by one of the group in which one member proposed consulting ChatGPT to identify a suitable technology for their purpose. The group agreed, collaboratively refining the prompt before submitting it. They then reviewed ChatGPT's response together: "One member of the group suggested asking ChatGPT what technology might be suitable for achieving the objective. Since we all agreed, we decided together what the best request to write would be. We then read ChatGPT's response together. While going through the response, one member suggested replacing the word *technology* in the request with *innovative and inclusive technologies* to see whether ChatGPT would propose different technologies" (3:2).

In another instance, two group members independently submitted the same prompt to ChatGPT to assess the consistency and quality of the responses. After comparing the outputs, they selected the one they judged to be the most comprehensive: "Two of us entered the same request into ChatGPT to evaluate the quality of the response. We then selected the one we considered the most complete" (1:14).

ChatGPT was sometimes seen as an oracle, a very expert peer, and a collaborative resource and therefore able to guide the generative group flow and to choose the most suitable technology with a specific connection with the specificity of the design. The students' comments show that ChatGPT is used as a specific tool to support decision-making within the group in relation to ICT selection, but at the same time, each Chat proposal/output is analyzed and critically discussed within the team in order to adopt the best design strategy and tools at an educational level. Groups

BOX 2 Representative artifact excerpts and Task 2 rubric anchors.

Notes: Excerpts are translated from Italian and lightly edited for readability (e.g., bulleting/spacing only). The first excerpt is from a group that received Level 1 on all three Phase 2 criteria; the second excerpt is from a group that received Level 3 on all three criteria.

Anchor A—Level 1 across criteria (Phase 2/Task 2)

Objective: “Identify and name the continents and oceans of the world.”

“The required points are addressed in the chat.” [list of unreasoned lessons names follows in the chat]

“Critical issue: we did not understand that we had to keep the same objective as in Phase 1.”

Why this corresponds to Level 1:

- Completeness (L1): The submission does not provide a structured lesson plan; activities and instructional sequence are missing or only alluded to.
- Relevance of methodological choices (L1): Methods/tools are not specified and are not explicitly aligned to the stated objective.
- Originality (L1): There is no evidence of multiple approaches or innovative design choices; the content remains minimal.

As for Task 1, the product rated by both judges is not strictly incorrect but, in relation to the specific rubric criteria, it shows a minimal level of planning, with few and superficially described activities, weak alignment between objectives, methodologies, and tools, and a limited number of approaches lacking innovation, which justifies a Level 1 (minimum score) assessment.

Anchor B—Level 3 across criteria (Phase 2 Task 2)

Lesson: “Responsible Use of Technology for 14-year-olds (AI).”

Objectives: understand responsible technology use; identify challenges for this age group; develop practical skills for navigating the digital world responsibly.

Introduction (10 min): discussion on students’ uses of technology (positive experiences and challenges).

Activity 1 (15 min): identify age-relevant challenges (e.g., social media pressure, time management, privacy).

Activity 2 (20 min): digital etiquette and respect online, using scenarios and concrete examples.

Activity 3 (15 min): online safety and privacy (e.g., strong passwords, privacy settings, recognizing risky situations).

Activity 4 (20 min): critical evaluation of online content (reliability, manipulation).

(Continued)

BOX 2 (Continued)

Conclusion (10 min): recap and Q&A; emphasize responsible use.

Homework: students produce a short guide for peers on responsible technology use.

Assessment: formative; group (60%) and individual (40%); self-assessment and brief satisfaction questionnaire; optional meeting with families facilitated by an expert.

Why this corresponds to Level 3:

- Completeness (L3): The lesson plan is structured and detailed, with multiple activities, timing, progression, assessment, and follow-up work.
- Relevance of methodological choices (L3): Activities and tools are coherently aligned to the objectives (responsible use, privacy/safety, critical evaluation), with a clear instructional sequence.
- Originality (L3): The plan combines multiple approaches (discussion, scenarios, practical skill-building, critical analysis, peer-facing product, family involvement), indicating broader and more innovative design thinking.

The products related to Task 2 and rated with Level 3 by the two independent judges demonstrate an advanced level of design, with well-developed and clearly described activities, full alignment between objectives, methodologies, and tools, and the integration of multiple innovative approaches.

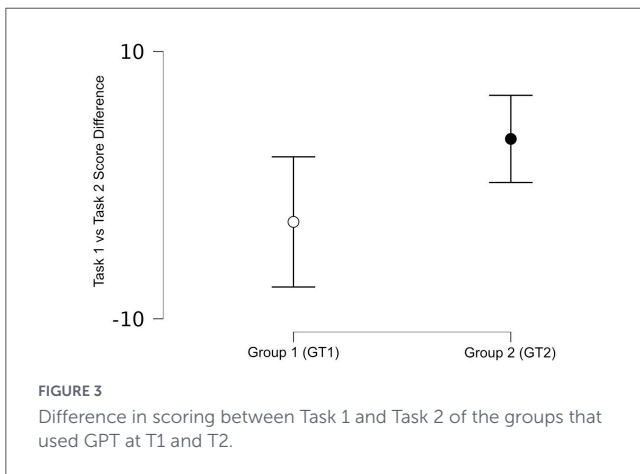


FIGURE 3
Difference in scoring between Task 1 and Task 2 of the groups that used GPT at T1 and T2.

frequently blended multiple interaction modes with ChatGPT (e.g., information-seeking alongside more dialogic processes of refinement and justification). Accordingly, it was introduced the term “peer-like” as the intended pedagogical positioning, while characterizing the observed enactment as hybrid and dynamic. In fact, the percentage of ChatGPT use as “an oracle” (43%) is related to the fact that the groups that mentioned this type of adoption declare to use it as an expert guide at list one time during their product development. The column dedicated to “Percentage of Groups (%)” is related to the proportion of the 21 groups that mentioned a theme in that code and this is applied also for the code discussed.

For the code “Chat GPT for formative design” groups reported the use of AI to identify a specific topic for their formative design process but also suggestions and ideas to build the design and programming process itself. In fact, for example, one group affirmed that “ChatGPT helped us identify macro areas for the chosen topic, find reference sites and find material suitable for the lesson to present to the students” (5:5).

For the code “Chat GPT chain of thoughts,” a few groups reported the importance of revising and enhancing the prompting process in groups. For example they affirmed that “after a discussion within the group, we defined clear instructions for ChatGPT, also taking into account the learning objectives, without neglecting “feedback” and evaluation” (3:1). It is crucial to discuss and to try to specify better and better the information for ChatGPT in order to receive the most accurate output and to guide the workflow. It is important to be aware of the clarity, efficacy and accuracy of the information produced by ChatGPT, then the prompting process has to be shared and well-structured with a specific alignment with the other design parts.

5.2.2 Students’ interactions during the group work

A detailed analysis of the students’ feedback in relation to the use of AI during specific moments of the in-class activity proposed (group task 1 and group task 2) was developed. The process of design in Task 1 and Task 2 was explored through a content analysis process by using Atlas.Ti.

TABLE 7 Change in product quality from Task 1 to Task 2 (sum of two raters' scores; range = 6–18).

(A) Descriptive statistics (cell means and change scores).						
Sequence (first use of ChatGPT)	<i>n</i>	Task 1: M (SD) (95% CI)		Task 2: M (SD) (95% CI)	Δ = Task2-Task1: M (SD) [95% CI]	
Task 1 (early-AI)	8	12.25 (3.99) (8.91, 15.59)		14.75 (2.92) (12.31, 17.19)	2.50 (5.86) (-2.40, 7.40)	
Task 2 (late-AI)	13	11.19 (4.30) (8.59, 13.79)		14.38 (4.43) (11.71, 17.06)	3.19 (5.01) (0.17, 6.22)	
(B)Planned within-sequence tests (paired Wilcoxon signed-rank).						
Sequence	<i>n</i>	Wilcoxon statistic (W)	<i>p</i> (two-sided)	<i>p</i> (one-tailed, directional)	Cohen's <i>d</i> (paired)	Direction tested
Early-AI (GPT in Task 1)	8	7.0	0.236	0.882	0.43	Task 2 < Task 1 (decline test)
Late-AI (GPT in Task 2)	13	11.5	0.056	0.028	0.64	Task 2 > Task 1
(C)Robustness check: 2 × 2 mixed ANOVA (Task within; Sequence between).						
Effect	<i>F</i> (df1, df2)	<i>p</i>	Partial η^2			
Sequence (between)	<i>F</i> (1, 19) = 0.27	0.612	0.014			
Task (within)	<i>F</i> (1, 19) = 6.33	0.021	0.250			
Task × Sequence	<i>F</i> (1, 19) = 0.08	0.776	0.004			

Early-AI, groups used ChatGPT in Task 1 (and not in Task 2). Late-AI = groups used ChatGPT in Task 2 (and not in Task 1). One-tailed *p*-values match the planned hypotheses (no decline for Early-AI; improvement for Late-AI).

Table 10 presents a thematic analysis of the codes that emerged from the students' written feedback about their workflow process. T1 and T2 are used to describe the task of reference in which students integrated or not the use of AI.

In Task 1, groups working without AI demonstrated higher engagement with traditional workflows (46%) and showed substantial reflection on digital tools (69%). These groups typically began with collaborative brainstorming sessions, focusing on technology assessment and methodological discussions. The non-AI groups exhibited stronger emphasis on collective decision-making and creative expression, as evidenced by Group 11's observation of their strength in "heterogeneous thinking and good level of knowledge, diverse professional and personal training experiences, abundant production of ideas" (T1, no AI, Strength). A little less than half the groups that used AI in the first phase demonstrate to reflect about it (44%) (Table 10).

A significant shift occurred in Task 2, where the groups that now cannot use AI think back about it (67%) feeling the contrast, while AI-using groups showed markedly higher engagement with the instructional design process (100%) and increased reflection on AI integration (75%). However, this came with notable trade-offs. As Group 16 pointed out, AI usage sometimes led to "less motivated to introduce personal creativity, adopting more standard solutions," (T2, AI, Weaknesses AI) while Group 19 noted "less cooperation among group members and mental commitment" (T2, AI, Weaknesses AI).

Interestingly, groups that transitioned from AI to non-AI approaches provided particularly insightful reflections. Group 9 acknowledged AI's utility while emphasizing human agency: "it was very useful to be able to benefit from ChatGPT's input in the first phase, but we recognized that it was fundamental and necessary for us to be able to select and plan in detail the activities suggested by Chat GPT." (T2, no AI, Reflection on AI use). Similarly, Group 2

noted that traditional approaches were "more stimulating and more confronting because it was more creative, divergent and generative" (T2, AI, Reflection on AI use).

The analysis revealed distinct patterns in the evolution of group workflows, particularly concerning the incorporation of artificial intelligence (AI). Groups not utilizing AI exhibited notably stronger collaborative dynamics and higher levels of creative engagement, as Group 14 which reported that "Compared to the teaching unit carried out without artificial intelligence, where we discussed things much more and tried to respect everyone's ideas, with AI there was less discussion and we all immediately agreed with the unit proposed by the computer" (T2, AI use, Reflection on Ai use). In contrast, groups incorporating AI tools demonstrated increased efficiency, though this often appeared to come at the expense of interpersonal interactions and group cohesion, as Group 16 mentioned as the input of AI use promoted "greater comparison of experiences in order to arrive at the choice of intervention." (T2, AI use, Group workflow with AI).

A notable learning curve emerged in relation to AI utilization. Initially, groups encountered difficulties in effectively formulating prompts, as exemplified by Group 13's reflection: "We found it difficult at first to understand how to pose questions to ChatGPT" (T2, AI, Reflection on AI use). Despite these initial challenges, participants progressively recognized the value of AI as a practical, time-saving tool, while simultaneously becoming more aware of its inherent limitations.

Differences in the quality and focus of group reflections were also evident between AI and non-AI groups. Non-AI groups tended to concentrate more thoroughly on aspects related to the collaborative process and group dynamics. Conversely, groups using AI developed insights that were technologically richer and more precise, yet these groups reported comparatively diminished levels of interpersonal engagement and interaction.

TABLE 8 Descriptive statistics and independent-samples Welch *t*-test for Task-2 product quality (sum of the two raters' scores; possible range = 6–18).

First use of ChatGPT	<i>n</i>	<i>M</i>	<i>SD</i>	95% CI for <i>M</i>	<i>t</i> (<i>df</i>)	<i>p</i>	Cohen's <i>d</i>	95% CI for <i>d</i>
Task 1 (early-AI)	8	14.75	2.92	(12.31, 17.19)				
Task 2 (late-AI)	13	14.38	4.43	(11.71, 17.06)	0.23 (18.84)	0.822	0.09	(-0.85, 1.03)

Welch's *t*-test indicates no reliable difference in Task-2 product quality between sequences. Cohen's *d* is positive when the Task-1 group mean is higher.

TABLE 9 Frequencies of categories and codes from qualitative feedback on AI use.

Category	Code	Frequency (n)	Percentage of groups (%)
Students' reflections	1. Reflection on AI use	18	86
	2. General reflection about group work	4	19
	3. ChatGPT chain of thoughts	3	14
AI use modalities	4. ChatGPT as "oracle" for selecting ICT	9	43
	5. ChatGPT for formative design	7	33
	6. ChatGPT for assessment selection	2	10
	7. ChatGPT as peer for ICT selection	2	10

Analysis is based on written feedback from 21 student groups collected after Tasks 1 and 2. "Frequency (n)" refers to the number of quotations coded into each category from a total of 45 relevant quotations. "Percentage of Groups (%)" is the proportion of the 21 groups that mentioned a theme in that code. ICT, Information and Communication Technology.

TABLE 10 Thematic analysis of student group feedback on workflow and process, by task and AI condition.

Task	AI condition	Theme/Code	Frequency (n)	Percentage of Groups (%)
1	No AI	Group workflow	6	46
		Instructional design process without AI	6	46
		Reflection on digital tools	9	69
		Strengths	3	23
		Weaknesses	3	23
	AI used	Group workflow	2	22
		Instructional design process with AI	4	44
		Reflection on AI use	1	11
		Strengths	2	22
		Weaknesses	2	22
2	No AI	Group workflow without AI	1	11
		Instructional design process without AI	7	78
		Reflection on AI use	6	67
		Reflection on group work	2	22
		Strengths	1	11
		Weaknesses	2	22
	AI used	Group workflow with AI	3	25
		Instructional design process with AI	13	100
		Learning outcomes description with AI	1	8
		Reflection on AI use	9	75
		Reflection on digital tools	4	33
		Strengths of AI	5	42
		Weaknesses	2	17
Weaknesses of AI	2	17		

This table presents a thematic analysis of the written feedback submitted by the 21 student groups after each task. The analysis is based on a total of 96 coded quotations. "Frequency (n)" indicates the number of quotations coded for each theme. "Percentage of Groups (%)" refers to the proportion of groups *within that specific condition* that mentioned a theme (e.g., for "Task 1, No AI," the percentage is out of the 13 groups in that condition). Codes in the "Theme/Code" column that specify "with AI" or "without AI" are presented as they were coded from the source data.

Certain quotes captured during the research process highlight these observations effectively. Group 5, which worked initially without AI, reflected strategically on technology integration, stating: “We thought of integrating this technology to be administered at the beginning and end of the lesson to assess how the perception of the subject matter and the content assimilated changes.” (T1, non-AI, Reflection on digital tools). Group 13, which engaged directly with AI, emphasized practical advantages and growing enthusiasm: “ChatGPT’s response allowed us to reduce the discussion time by making the choice concrete. The enthusiasm in designing the activity grew during the process as did the satisfaction in using AI.” Finally, Group 7, transitioning from prior AI exposure to a session without AI, noted the absence of technological reliance positively: “Unlike the previous work, in this case we did not use any research technology, nor did we feel the lack of it. There was an interaction based on sharing personal experiences” (T2, non-AI, Reflection on group work).

6 Discussion

The goal of this study was twofold: to understand what is the impact of using ChatGPT as a peer-like feedback provider on the acquisition of knowledge and skills during collaborative activities, particularly in the execution of specific tasks, and to understand how the outcomes differ between groups that use ChatGPT during the selection of technological tools for inclusive teaching design and those that use it during the actual design process.

6.1 Impact of ChatGPT as a peer-like feedback provider

6.1.1 Post-withdrawal sustained performance effect

The findings are consistent with what we refer to as a post-withdrawal sustained performance (PWSP) effect, defined here as improved performance observed during the AI-available phase that is not followed by a detectable decline in the immediately subsequent AI-withdrawn phase within the study timeframe. In collaborative instructional design, this pattern is reflected in the absence of poorer artifact quality when AI support was removed, alongside a knowledge-check pattern in which gains occur in the AI-allowed phase and remain stable thereafter. Given our timing constraints (240 min total, with knowledge tests spaced approximately 90 min apart and the two design tasks separated by roughly 25 min from the end of the first to the beginning of the second, or ~105 min end-to-end), we can only speak to short-term dynamics.

First, in terms of artifact quality, groups that used ChatGPT during the initial task (GT1; early-AI) showed no evidence of deterioration when moving to Task 2 without AI. This is reflected in the paired non-parametric test within this sequence: the Task1→Task2 change was not negative, and the one-tailed Wilcoxon test specifically targeting a decline was non-significant. The mixed ANOVA robustness check (Table 7C) similarly showed no evidence that the Task 1→Task 2 change differed by sequence

(no Task × sequence interaction), consistent with the absence of deterioration after AI withdrawal. While the small sample limits precision, the pattern is inconsistent with a “dependency” interpretation and instead supports the possibility that early AI-supported activity helped groups internalize design strategies that remained available when AI was no longer accessible (at least in the short term). This aligns with recent quantitative syntheses suggesting that LLM-supported learning can yield durable gains when used as a learning aid rather than as a shortcut (Deng et al., 2024; Wang and Fan, 2025).

Second, the knowledge test results converge with this interpretation. Within-sequence Wilcoxon tests show that improvements are phase-specific: GT1 increased from T0 to T1 (immediately following the AI-allowed task), with no significant change from T1 to T2 (after AI withdrawal), while GT2 showed no significant change from T0 to T1 (no AI in Task 1) and a significant increase from T1 to T2 (immediately following the AI-allowed task). This phase-locked pattern is especially clear on the technological sub-score, consistent with the notion that the AI-supported activity was most closely aligned with technology-focused content and decisions. Importantly, the post-AI phase in GT1 does not show a detectable drop, which is consistent with short-term maintenance of the gain rather than a transient boost limited to AI availability.

Third, because crossover and counterbalanced designs can be sensitive to order-related artifacts, we explicitly tested whether the AI-related change depended on exposure order in the knowledge data. Using a 2 × 2 mixed-design ANOVA on interval change scores (AI interval vs. no-AI interval × sequence), the Sequence × Treatment interaction was non-significant for the overall score and both sub-scores (all $ps \geq 0.387$), providing no evidence that the AI-related change depended on receiving AI first versus second. In the same robustness check, the Treatment main effect was significant only for the technological sub-score [$F(1, 82) = 4.28, p = 0.042, \eta_p^2 = 0.050$], whereas it was not significant for the overall score or the methodological sub-score. Taken together with the Wilcoxon pattern, this supports the interpretation that the observed improvements are linked to the AI-allowed phase without indicating order-dependent artifacts, and that gains were maintained when AI was removed (in the short window of this study).

To situate these findings relative to prior work, it is useful to highlight that several influential studies observing either harms or performance parity with LLM support operate within similarly short, contiguous time windows. Bastani et al. report that the harmful effect of “vanilla” ChatGPT (and, by contrast, performance parity with the control condition when ChatGPT is constrained via tutor-like prompting) emerges within three consecutive phases delivered in a single 90-min classroom session (teacher review, assisted practice, and an unassisted, closed-book/closed-laptop evaluation), although the paper does not specify the exact minutes allocated to each phase. Lehmann et al., in turn, find no statistically significant learning advantage (i.e., parity with the control condition) across two lab experiments conducted in a single 85-min sitting: a 20-min pre-test, a 45-min learning phase (with AI access in the treatment condition), followed immediately by a 20-min post-test. In this respect, despite our own timing constraints and the short interval between

AI-allowed and AI-withdrawn phases, our study speaks to the same “within-session” dynamics that have been central in recent experimental work.

Against that backdrop, our results strengthen the idea that the conditions under which learners engage with generative AI—especially whether its use is explicitly scaffolded—may be a key moderator of outcomes. Even on a short timescale, structuring AI as a team-member that offers candidate ideas for discussion, alongside explicit metacognitive scaffolds (rules, roles, instructions, guidelines, and environmental/technical supports), may help shift generative AI from oracular answer-giving toward procedural and strategic support (Tankelevitch et al., 2024), potentially reducing the risks reported when access is comparatively unstructured (e.g., Bastani et al., 2025; Lehmann et al., 2025).

6.1.2 Knowledge development pattern

The data suggest that this PWSP pattern differed between technological and methodological domains, with technological knowledge showing larger gains than methodological knowledge. This was echoed in the groups’ reflections: several described using ChatGPT in Task 1 primarily to surface, compare, and select candidate tools. In the same reflections, some groups reported that when AI was not available in Task 2, they proceeded “the old way” and/or felt freer working without it, and at least one explicitly noted that they “did not miss it.” Our qualitative coding of the groups’ written reflections supports this interpretation: reported AI use often centered on requesting tool options and making selections, more than on pedagogical rationales (e.g., running the same request twice to judge response quality and then selecting the most complete output; asking which tool would be most effective and then choosing the most suitable option after discussion; Table 9; Codes 4 and 7). One plausible mechanism is near transfer: Task 1 required selecting and justifying technologies, which overlaps closely with technological knowledge and may therefore transfer more readily than methodological reasoning. This aligns with classic accounts suggesting that transfer is strongest when a subsequent task is highly similar to what was practiced (Perkins and Salomon, 1992). The task design may have amplified this short-term PWSP pattern.

6.1.3 Reflection and meta-learning

The qualitative data reveals substantial reflection on AI use (40% of coded responses, 86% of groups). Notably, groups engaged in reflection about AI even during phases where they weren’t actively using it (code “T2 0 Reflection on AI use” with 67%), suggesting that exposure to AI prompted continuous critical evaluation of tool selection and design processes. When students engage in resource comparison activities, they develop enhanced self-feedback abilities that extend beyond the immediate task, and our participants demonstrated metacognitive awareness about when and why to use AI tools (Nicol and Kushwah, 2024).

On a similar line, Group 7 reflected: “Unlike the previous work, in this case we did not use any research technology, nor did we feel the lack of it. There was an interaction based on sharing personal experiences” (T2, No AI use, Reflection on AI use). This

type of reflection exemplifies what Lipnevich and Smith (2022) describe as behavioral processing of feedback, a critical component where learners actively decide which strategies to employ based on cognitive and affective evaluations. In their Student-Feedback Interaction Model, this represents the “What am I going to do with the feedback?” stage, where students make deliberate choices about tool utilization rather than passively accepting technological assistance. The students’ critical evaluation of when AI was unnecessary demonstrates the self-regulatory processes that both Lipnevich and Smith (2022) and Nicol and Kushwah (2024) identify as essential for developing learner agency and effective feedback processing.

Finally, student’s feedbacks suggest a learning curve in AI interaction. They reported that “the initial difficulty encountered by the group was to clearly identify which specific questions to ask the AI in order to receive a comprehensive answer” (T2, AI use, Reflection on AI use) reflecting the necessity of an initial phase of exploration that lead to more effective strategies. Additionally, students recognize the prompting creation as an iterative process, and this is highlighted from the fact that they refined their prompts for better alignment with learning objectives. Similar findings in higher education have been reported by Carrasco-Sáez et al. (2025) that noted how students adapt, progressively refining their prompting approaches and adopting more purposeful strategies.

The ability to critically evaluate and modify AI suggestions seems to be an emergent one in the groups during the experimentation, as noted in other studies where it appears to develop as part of a broader metacognitive engagement with the tool (Teng, 2025). In collaborative contexts, this evaluation capacity often becomes collective, with groups co-constructing and refining prompts to improve AI responses, an emergent skill that strengthens through shared experimentation (Perifanou and Economides, 2025).

6.1.4 Trade-offs in collaborative learning

Looking again at the qualitative data, a tension emerges about the use of the AI as a group member. While ChatGPT improved task performance and knowledge retention, some groups noted reduced peer interaction. For example, Group 16 noted that “through the use of AI, groups are less motivated to introduce a personal reworking/creativity” (T2, AI use, Weaknesses AI) and Group 19 observed that AI “seems to prompt less cooperation among the group members and less mental commitment” (T2, AI use, Weaknesses AI). These observations align with challenges identified by Zheng et al. (2024), who emphasized that learners often face difficulties in collaborating with one another and co-regulating their learning in online collaborative environments, which can lead to diminished group performance. Their research specifically highlights that effective collaborative knowledge building requires not just technological support, but also carefully designed feedback and feed forward mechanisms that promote co-regulation among participants.

Similar trade-offs have been observed elsewhere. In a recent computer-supported asynchronous collaborative learning study, introducing ChatGPT reshaped students’ interaction patterns, sometimes reducing the depth of peer exchange when the AI

became a dominant conversational partner (Kim et al., 2024). Likewise, research on interdisciplinary collaborative learning found that while ChatGPT enhanced efficiency and output quality, students also reported a decline in creativity and self-discipline (Zhu et al., 2023).

While our findings show that AI can enhance task outcomes, they also suggest that without proper scaffolding for collaboration, AI integration might inadvertently reduce the social regulation processes that are crucial for successful collaborative learning. This tension can be solved by keeping the balance between technological enhancement and maintaining the essential human elements of co-regulation and collaborative knowledge building in technology-mediated learning environments (Zheng et al., 2024).

6.2 Effect of AI on group interactions and final product quality

Each group produced one AI-assisted and one non-AI artifact. We summarized the within-group AI impact as the difference between these two products and compared that impact between sequences (GT1 vs. GT2). This tests whether AI changes product quality differently depending on whether groups encountered it first or second (see section 3.4).

Looking at the data, synthesized in Table 8, the effect was small and the CI crossed zero ($d = 0.24$, 95% CI $[-0.63, 1.08]$), indicating no reliable difference in final product quality (the product of Task 2) between groups that used ChatGPT during the selection of ICT tools for inclusive teaching design (Task 1) and those that used it later during the actual design process (Task 2).

Moreover, looking at the process highlighted by the qualitative analysis, this study results do not provide a simplistic AI-positive or AI-negative perspective, highlighting both strengths (time-saving, structured output, technology insights) and challenges (decreased discussion, initial difficulty in prompting, potential loss of creativity and motivation) as discussed in the previous section about trade-offs (Zheng et al., 2024; Kim et al., 2024; Zhu et al., 2023).

In terms of comparison between AI and non-AI approaches used for the activity development, we identified workflow shifts. For example, non-AI groups seemed to be engaged in richer collaborative design processes, while AI-assisted groups supported efficiency and structured output. Through this lens, AI, at the same time, speeds up work but may limit deeper students' engagement during the group work. These findings are coherent with broader research showing that AI can streamline tasks (Noy and Zhang, 2023) while simultaneously reducing opportunities for co-regulation that underpin deeper learning (Järvelä and Hadwin, 2013).

Nonetheless, those reflections were shared by most of the groups at a certain time, not only by the ones that used ChatGPT in Task 1 or Task 2. On the other hand, three quarters of those who used ChatGPT in Task 1, did reflect about it, or about its absence, even during Task 2. This illustrates how technological mediation influences learners' regulation strategies beyond its immediate use, reinforcing the need for scaffolding that preserves co-regulation while harnessing AI's efficiency (Zheng et al., 2024; Zhu et al., 2023).

6.3 Why do we have different results compared to other literature about knowledge transfer and retention using LLMs?

Recent research has raised significant concerns about the potential for generative AI to hinder authentic learning. For instance, studies such as Bastani et al. (2024) and Lehmann et al. (2025) have found that when students use AI tools individually, they often fall into a pattern of passive acceptance, taking the AI's output at face value without critical evaluation. This behavior, where the AI is treated as an infallible oracle, can lead to reduced cognitive effort, an "illusion of understanding" or "crutch effect" and ultimately, harm the development of durable knowledge. A key distinction in these studies is the context of use: students worked alone and were not prompted to collaboratively discuss or critically deconstruct the AI-generated content, a stark contrast to the methodology employed in our study. In our study, improvements were observed in the AI-available phase, and we did not observe a decline in the subsequent AI-withdrawn phase; this pattern was clearest on technology-related knowledge and was consistent with stable artifact quality after AI removal. This aligns with qualitative feedback where Group 9 noted: "it was very useful to be able to benefit from ChatGPT's input in the first phase, but we recognized that it was fundamental and necessary for us to be able to select and plan in detail the activities suggested by ChatGPT."

The preservation of knowledge appears tied to students actively processing and critically evaluating AI suggestions, rather than passively accepting them. This finding parallels Escalante et al.'s (2023) research, which demonstrated that ENL students could achieve comparable learning outcomes with AI-generated feedback as with human tutor feedback, particularly when students engage critically with the AI suggestions.

The request of using ChatGPT as a group member, keeping it in the loop of the conversation, and discussing its feedback as one would with a human group member, seems to have contributed critically to the positive outcome of these interactions. This highlights the importance of a blended feedback model that leverages the strengths of AI feedback, such as its clarity, specificity, and availability, while encouraging student engagement through discussion and critical evaluation of the feedback received (Escalante et al., 2023).

7 Implications

This study set out to understand how positioning ChatGPT as a "feedback team-mate" during collaborative lesson-design tasks shapes students' knowledge, skills, and group dynamics. Drawing on a quasi-experimental, mixed-methods design with 102 student-teachers across two authentic design challenges, we compared groups that worked with or without ChatGPT. The evidence gathered provides a rich picture of the benefits and tensions that emerge when an LLM becomes an active participant in learning.

The results extend the current debate on the impact of AI on learning by showing that AI-supported work does not necessarily produce fragile performance that collapses once the tool

is removed. In our study, improvements were observed in the AI-available phase, and we did not observe a decline in the subsequent AI-withdrawn phase; this pattern of Post-Withdrawal Sustained Performance—clearest on technology-related knowledge—was consistent with stable artifact quality after AI removal. Although the present design only speaks to short-term dynamics, the findings are compatible with the idea that AI-mediated collaboration can seed strategies and conceptual understanding that continue to serve students in later, tool-free contexts. This invites longitudinal research to trace how long such benefits endure and whether repeated alternations between AI-supported and unsupported tasks strengthen or dilute them.

Equally significant is the shift in group workflow. Across our qualitative evidence, ChatGPT often streamlined production and helped groups move forward, yet in some cases it appeared to dampen spontaneous, peer-to-peer idea-building exchanges that are central to collaborative creativity. Educators face a design dilemma: how can they harness the efficiency gains without compromising the generative dialogue that fuels creativity? Alternating phases with and without AI, assigning explicit roles such as “AI checker” or “human synthesizer,” and scheduling reflective debriefs may help restore conversational balance.

Prompt development emerged from the qualitative analysis as an important metacognitive skill. Students attempted to refine prompts, evaluate AI output, and decide when to accept, adapt, or discard suggestions. Embedding explicit instruction on prompt-craft and, more in general, critical AI interaction into curricula could therefore enhance both digital literacy and self-regulated learning.

Methodologically, the study illustrates the value of mixed-methods approaches for capturing the often invisible processes of AI-human interaction. Quantitative gains told only part of the story; only by triangulating scores, expert ratings, and coded reflections could we surface the tension between speed and interaction.

At the policy level, a binary “ban or allow” framing is no longer adequate. Institutions should instead cultivate a culture of critical adoption, providing safe AI sandboxes, professional-development time, and assessment redesign guidance. When AI is invited into, rather than over, the learning conversation, it can amplify efficiency and catalyze metacognitive growth—provided that educators orchestrate its participation deliberately and remain attentive to the social fabric of collaboration.

7.1 Limitations

A limitation of this study is that the post-test was administered immediately following the tasks, also, ideally, there would have been a final follow up where none of the groups could use AI. Therefore, while our results indicate sustained performance immediately after AI withdrawal within the study timeframe, future research should incorporate delayed post-tests to assess longer-term knowledge retention. Furthermore, participants were not formally trained in prompt crafting and AI interaction. Although the instruction was to “treat it like a team-member,” so to let the participants use a known type of interaction, individual differences in their ability to query the AI may have influenced the outcomes.

Future studies could include a prompt-crafting/AI-interaction tutorial to control for this variable.

We did not collect verbatim ChatGPT interaction transcripts; therefore, our characterisation of AI interaction modes is based on groups’ immediate post-task reflections and feedback rather than direct log-based coding. Future work should complement these accounts with automatic capture of ChatGPT interaction traces (to code prompt types and prompting strategies), while retaining process-oriented measures to interpret how inputs and outputs were negotiated and taken up within the group.

A small number of items in the knowledge test showed weak/negative discrimination at specific waves, consistent with a short heterogeneous content-sampling knowledge check; future iterations will refine or replace these items. Finally, additional studies are needed for validating the results in other tasks, disciplines, cultures and in general to widen the sample. Because of the small group numbers several intervals are broad (e.g., Δ -score CI spans -2.20 to -0.09); true effects may therefore be smaller (or larger) than our point estimates.

In conclusion, this study suggests that AI tools such as ChatGPT can support immediate performance gains and may also foster the development of design skills with potential for transfer, particularly when used as a team member in small-group settings. Despite the study’s limitations, these findings add to the emerging evidence on AI-supported collaborative learning and point to clear directions for future empirical research and instructional design.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The requirement of ethical approval was waived by the University of Trento Research Ethics Committee for the studies involving humans (2025-082ESA). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

DA: Investigation, Writing – review & editing, Conceptualization, Writing – original draft, Formal analysis, Data curation, Methodology, Visualization. AS: Methodology, Project administration, Supervision, Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition, Resources. FP: Formal analysis, Data curation,

Methodology, Writing – original draft, Writing – review & editing. AL: Validation, Supervision, Resources, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research has been possible partly thanks to the PNRR (National Plan of Recovery and Resilience) Next-Generation EU funding. This work was supported by the Italian Ministry of University and Research (MUR) through the Department of Excellence grant awarded to the Department of Psychology and Cognitive Science, University of Trento.

Acknowledgments

We authors acknowledge the use of large language models for editing, rephrasing and polishing parts of the manuscript to enhance the general fluency and readability.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Acosta-Enriquez, B. G., Arbulú Ballesteros, M. A., Arbulú Perez Vargas, C. G., Orellana Ulloa, M. N., Gutiérrez Ulloa, C. R., Pizarro Romero, J. M., et al. (2024). Knowledge, attitudes, and perceived ethics regarding the use of ChatGPT among Generation Z university students. *Int. J. Educ. Integr.* 20:10. doi: 10.1007/s40979-024-00157-4
- Agostini, D., and Picasso, F. (2024). Large language models for sustainable assessment and feedback in higher education: Towards a pedagogical and technological framework. *Intell. Artif.* 18, 121–138. doi: 10.3233/IA-240033
- Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., et al. (2019). “Guidelines for human-AI interaction,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, 1–13.
- Anderson, L. W., and Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives: Complete Edition*. Boston, MA: Addison Wesley Longman, Inc.
- Arum, R., Calderon Leon, M., Li, X., and Lopes, J. (2025). ChatGPT early adoption in higher education: Variation in student usage, instructional support, and educational equity. *Aera Open* 11:23328584251331956. doi: 10.1177/23328584251331956
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, Ö., and Mariman, R. (2024). *Generative AI can Harm Learning*. The Wharton School Research Paper.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, Ö., and Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proc. Natl. Acad. Sci. U. S. A.* 122:e2518204122. doi: 10.1073/pnas.2518204122
- Bektik, D., Edwards, C., Whitelock, D., and Antonaci, A. (2025). *Use of LLM Tools within Higher Education: Report 2 (Project Deliverable)*. Geneva: Zenodo. doi: 10.5281/zenodo.17747509
- Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ.* 5, 7–74. doi: 10.1080/0969595980050102

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. The authors acknowledge the use of large language models (such as OpenAI GPT O3 and 5-class models) for editing, rephrasing and polishing parts of the manuscript to enhance the general fluency and readability.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2026.1751618/full#supplementary-material>

- Carless, D., and Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assess. Eval. High. Educ.* 43, 1315–1325. doi: 10.1080/02602938.2018.1463354
- Carrasco-Sález, J. L., Contreras-Saavedra, C., San-Martín-Quiroga, S., Contreras-Saavedra, C. E., and Viveros-Muñoz, R. (2025). Analyzing higher education students’ prompting techniques and their impact on ChatGPT’s performance: An exploratory study in Spanish. *Appl. Sci.* 15:7651. doi: 10.3390/app15147651
- Chandler, C., Raju, R., Reitman, J. G., Penuel, W. R., Ko, M., Bush, J. B., et al. (2025). *Improving the Generalizability of Models of Collaborative Discourse*. Italy: International Educational Data Mining Society.
- Chen, L., Zaharia, M., and Zou, J. (2023). How is ChatGPT’s behavior changing over time? *arXiv [Preprint]* doi: 10.48550/arXiv.2307.09009
- Chen, Y., Pan, X., Li, Y., Ding, B., and Zhou, J. (2024). A simple and provable scaling law for the test-time compute of large language models. *arXiv [Preprint]* doi: 10.48550/arXiv.2411.19477
- Corbett, A. T., and Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User Adapt. Interact.* 4, 253–278. doi: 10.1007/BF01099821
- Deng, R., Jiang, M., Yu, X., Lu, Y., and Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Comput. Educ.* 227:105224. doi: 10.1016/j.compedu.2024.105224
- Ebel, P., Söllner, M., Leimeister, J. M., Crowston, K., and de Vreede, G. J. (2021). Hybrid intelligence in business networks. *Electron. Mark.* 31, 313–318. doi: 10.1007/s12525-021-00481-4

- Escalante, J., Pack, A., and Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *Int. J. Educ. Technol. High. Educ.* 20:57. doi: 10.1186/s41239-023-00425-2
- European Commission (2019a). *Building Trust in Human-Centric Artificial Intelligence [COM(2019) 168 final]*. Brussels: European Commission.
- European Commission (2019b). *High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI*. Luxembourg: Publications Office of the European Union. doi: 10.2759/346720
- Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487
- Heffernan, N. T., and Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Int. J. Artif. Intell. Educ.* 24, 470–497. doi: 10.1007/s40593-014-0024-x
- Hobert, S., and Berens, F. (2023). Developing a digital tutor as an intermediary between students, teaching assistants, and lecturers. *Educ. Technol. Res. Dev.* 71, 2631–2652. doi: 10.1007/s11423-023-10293-2
- Holmes, W., Bialik, M., and Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching & Learning*. Boston, MA: The Center for Curriculum Redesign.
- Holmes, W., Bialik, M., and Fadel, C. (2023). *Artificial Intelligence in Education*. Geneva: Globethics Publications.
- Holmes, W., and Tuomi, I. (2022). State of the art and practice in AI in education. *Eur. J. Educ.* 57, 542–570. doi: 10.1111/ejed.12533
- Järvelä, S., and Hadwin, A. F. (2013). New frontiers: Regulating learning in CSCL. *Educ. Psychol.* 48, 25–39. doi: 10.1080/00461520.2012.748006
- Kasneji, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274
- Kim, H. K., Nayak, S., Roknaldin, A., Zhang, X., Twyman, M., and Lu, S. (2024). Exploring the impact of ChatGPT on student interactions in computer-supported collaborative learning. *arXiv [Preprint]* doi: 10.48550/arXiv.2403.07082
- Koedinger, K. R., and Corbett, A. (2006). “Cognitive tutors: Technology bringing learning sciences to the classroom,” in *The Cambridge Handbook of the Learning Sciences*, ed. R. K. Sawyer (Cambridge: Cambridge University Press), 61–78.
- Kovari, A. (2025). A systematic review of AI-powered collaborative learning in higher education: Trends and outcomes from the last decade. *Soc. Sci. Humanit. Open* 11:101335. doi: 10.1016/j.ssaho.2025.101335
- La Scala, J. A., Sahli, S., and Gillet, D. (2025). “Stimulating brainstorming activities with generative AI in higher education,” in *Proceedings of the 2025 IEEE Global Engineering Education Conference (EDUCON)*, London. doi: 10.1109/EDUCON62633.2025.11016340
- Lee, S. S., and Moore, R. L. (2024). Harnessing Generative AI (GenAI) for automated feedback in higher education: A systematic review. *Online Learn.* 28, 82–104. doi: 10.24059/olj.v28i3.4593
- Lehmann, M., Cornelius, P. B., and Sting, F. J. (2025). AI meets the classroom: When do large language models harm learning? *arXiv [Preprint]* doi: 10.2139/ssrn.4941259
- Lipnevich, A. A., Berg, D. A., and Smith, J. K. (2016). “Toward a model of student response to feedback,” in *Handbook of Human and Social Conditions in Assessment*, eds G. T. L. Brown and L. R. Harris (Abingdon: Routledge), 169–185.
- Lipnevich, A. A., and Smith, J. K. (eds). (2018). *The Cambridge Handbook of Instructional Feedback*. Cambridge: Cambridge University Press.
- Lipnevich, A. A., and Smith, J. K. (2022). Student-feedback interaction model: Revised. *Stud. Educ. Eval.* 75:101208. doi: 10.1016/j.stueduc.2022.101208
- Luckin, R., and Holmes, W. (2016). *Intelligence Unleashed: An Argument for AI in Education*. Available online at: <https://discovery.ucl.ac.uk/id/eprint/1475756/> (accessed November 19, 2025).
- Lynch, S. (2022). *AI in the Loop: Humans Must Remain in Charge*. Stanford, CA: Stanford Institute for Human-Centered Artificial Intelligence.
- Maadi, M., Akbarzadeh Khorshidi, H., and Aickelin, U. (2021). A review on human–AI interaction in machine learning and insights for medical applications. *Int. J. Environ. Res. Public Health* 18:2121. doi: 10.3390/ijerph18042121
- Martin, P. P., Kranz, D., Wulff, P., and Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *J. Res. Sci. Teach.* 60, 1883–1918. doi: 10.1002/tea.21903
- Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., et al. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students’ text revision, motivation, and positive emotions. *Comput. Educ. Artif. Intell.* 6:100199. doi: 10.1016/j.caeai.2023.100199
- Mohammadi, M., Tajik, E., Martinez-Maldonado, R., Sadiq, S., Tomaszewski, W., and Khosravi, H. (2025). Artificial intelligence in multimodal learning analytics: A systematic literature review. *Comput. Educ. Artif. Intell.* 8:100426. doi: 10.1016/j.caeai.2025.100426
- Monarch, R. M. (2021). *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*. Shelter Island, NY: Manning Publications.
- Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artif. Intell. Rev.* 56, 3005–3054. doi: 10.1007/s10462-022-10246-w
- Natarajan, S., Mathur, S., Sidheekh, S., Stammer, W., and Kersting, K. (2025). Human-in-the-loop or AI-in-the-loop? Automate or collaborate? *Proc. AAAI Conf. Artif. Intell.* 39, 28594–28600. doi: 10.1609/aaai.v39i27.35083
- Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assess. Eval. High. Educ.* 46, 756–778. doi: 10.1080/02602938.2020.1823314
- Nicol, D., and Kushwah, L. (2024). Shifting feedback agency to students by having them write their own feedback comments. *Assess. Eval. High. Educ.* 49, 419–439. doi: 10.1080/02602938.2023.2236219
- Nicol, D. J., and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Stud. High. Educ.* 31, 199–218. doi: 10.1080/03075070600572090
- Noy, S., and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 187–192. doi: 10.1126/science.adh2586
- OECD (2024). Explanatory Memorandum on the Updated OECD Definition of an AI System. Paris: OECD. Available online at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf (accessed March 1, 2026).
- OpenAI (2025). *Introducing GPT-OSS [Press Release]*. San Francisco, CA: OpenAI.
- Pack, A., Barrett, A., and Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Comput. Educ. Artif. Intell.* 6:100234. doi: 10.1016/j.caeai.2024.100234
- Panadero, E., Lipnevich, A., and Broadbent, J. (2019). “Turning self-assessment into self-feedback,” in *The Impact of Feedback in Higher Education: Improving Assessment Outcomes for Learners*, eds M. Henderson, R. Ajjawi, D. Boud, and E. Molloy (Abingdon: Routledge), 147–163.
- Perifanou, M., and Economides, A. A. (2025). Students collaboratively prompting ChatGPT. *Computers* 14:156. doi: 10.3390/computers14050156
- Perkins, D. N., and Salomon, G. (1992). “Transfer of learning,” in *International Encyclopedia of Education*, 2nd Edn, eds T. Husén and T. N. Postlethwaite (Oxford: Pergamon Press), 6452–6457.
- Roll, I., and Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *Int. J. Artif. Intell. Educ.* 26, 582–599. doi: 10.1007/s40593-016-0110-3
- Shaer, O., Cooper, A., Mokryn, O., Kun, A. L., and Ben Shoshan, H. (2024). “AI-augmented brainwriting: Investigating the use of LLMs in group ideation,” in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, New York, NY, 1–17. doi: 10.1145/3613904.3642414
- Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., and Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* 5, 46–57. doi: 10.1038/s42256-022-00593-2
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum. Comput. Interact.* 36, 495–504. doi: 10.1080/10447318.2020.1741118
- Shute, V. J. (2008). Focus on formative feedback. *Rev. Educ. Res.* 78, 153–189. doi: 10.3102/0034654307313795
- Stahl, G., Koschmann, T., and Suthers, D. (2006). “Computer-supported collaborative learning: An historical perspective,” in *Cambridge Handbook of the Learning Sciences*, ed. R. K. Sawyer (Cambridge: Cambridge University Press), 409–426.
- Tamkin, A., Brundage, M., Clark, J., and Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv [Preprint]* doi: 10.48550/arXiv.2102.02503
- Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., et al. (2024). “The metacognitive demands and opportunities of generative AI,” in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, New York, NY, 1–24.
- Teng, M. F. (2025). Understanding EFL student writers’ metacognitive awareness in utilizing ChatGPT. *System* 135:103848. doi: 10.1016/j.system.2025.103848
- Thorsteinsson, G., and Page, T. (2012). Encouraging innovativeness through computer-assisted collaborative learning. *J. Sch. Educ. Technol.* 7, 22–30. doi: 10.26634/jsch.7.3.1671
- To, J., Gutterman, J., and Lipnevich, A. A. (2023). “Students’ emotions in feedback engagement,” in *Unpacking Students’ Engagement with Feedback* (Routledge), 26–40.
- Tomazin, L., Lipnevich, A. A., and Lopera-Oquendo, C. (2023). Teacher feedback vs. annotated exemplars: Examining the effects on middle school students’ writing performance. *Stud. Educ. Eval.* 78:101262.
- van Allen, P. (2018). Prototyping ways of prototyping AI. *Interactions* 25, 46–51. doi: 10.1145/3274566

- VanLehn, K. (2006). The behavior of tutoring systems. *Int. J. Artif. Intell. Educ.* 16, 227–265. doi: 10.3233/IRG-2006-16(3)02
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* 46, 197–221. doi: 10.1080/00461520.2011.611369
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). “Attention is all you need,” in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, 5998–6008.
- Volta, S., and Di Stefano, N. (2024). Using wearable sensors to study musical experience: A systematic review. *Sensors* 24:5783. doi: 10.3390/s24175783
- Vygotsky, L. S. (1966). Igra i ee rol'v psikhicheskoy razvitiy rebenka. *Voprosy Psikhologii* 62–76.
- Wang, H., Wang, C., Chen, Z., Liu, F., Bao, C., and Xu, X. (2025). Impact of AI-agent-supported collaborative learning on the learning outcomes of university programming courses. *Educ. Inf. Technol.* 30, 17717–17749. doi: 10.1007/s10639-025-13487-8
- Wang, J., and Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Hum. Soc. Sci. Commun.* 12:621. doi: 10.1057/s41599-025-04787-y
- Wisniewski, B., Zierer, K., and Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Front. Psychol.* 10:3087. doi: 10.3389/fpsyg.2019.03087
- Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Gener. Comput. Syst.* 135, 364–381. doi: 10.1016/j.future.2022.05.011
- Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions* 26, 42–46. doi: 10.1145/3328485
- Yan, Y.-M., Chen, C.-Q., Hu, Y.-B., and Ye, X.-D. (2025). LLM-based collaborative programming: Impact on students' computational thinking and self-efficacy. *Humanit. Soc. Sci. Commun.* 12:149. doi: 10.1057/s41599-025-04471-1
- Yousif, J. H. (2025). Artificial intelligence revolution for enhancing modern education using zone of proximal development approach. *Appl. Comput. J.* 5, 386–398. doi: 10.52098/acj.2025.5239
- Zheng, L., Fan, Y., Chen, B., Li, X., and Li, H. (2024). An AI-enabled feedback-feedforward approach to promoting online collaborative learning. *Educ. Inf. Technol.* 29, 11385–11406. doi: 10.1007/s10639-023-12292-5
- Zhou, Q., Hashim, H., and Sulaiman, N. A. (2025). Integrating AI chatbots in informal digital English learning: Impacts on listening competencies in Chinese higher education. *Educ. Inf. Technol.* doi: 10.1007/s10639-025-13811-2 [Epub ahead of print].
- Zhu, G., Fan, X., Hou, C., Zhong, T., Seow, P., Shen-Hsing, A. C., et al. (2023). Embrace opportunities and face challenges: Using ChatGPT in undergraduate Students' collaborative interdisciplinary learning. *arXiv [Preprint]* doi: 10.48550/arXiv.2305.18616