

Article

Perception–Performance Gap in Generative AI: An Exploratory Study Across Two Engineering Education Contexts

Irida Shallari ^{1,*}, Vincenzo Gallo ², Marco Carratù ², Mazhar Hussain ¹, David Krapohl ¹
and Seyed Jaleleddin Mousavirad ¹

¹ Department of Computer and Electrical Engineering, Mid Sweden University, 85170 Sundsvall, Sweden; mazhar.hussain@miun.se (M.H.)

² Department of Industrial Engineering, University of Salerno, 84084 Fisciano, Italy; vgallo@unisa.it (V.G.); mcarratu@unisa.it (M.C.)

* Correspondence: irida.shallari@miun.se; Tel.: +46-(0)10-1428350

Abstract

Generative AI (GenAI) tools are increasingly used by students in higher education, including in technically demanding engineering courses. However, fluent AI-generated responses may still contain incorrect or incomplete information, creating a risk that students overestimate their reliability. This exploratory study investigates the relationship between students' perceived usefulness of GenAI and an instructor-benchmarked reference evaluation of model outputs in two digital systems design courses. The study involved voluntary survey responses from 32 students in an undergraduate course at MIUN and 20 students in a graduate-level course at UNISA. Student perception data were combined with teacher-side benchmarking of selected GenAI models on tasks categorized by cognitive depth. Findings indicate that prior GenAI familiarity was associated with interaction frequency and average perceived usefulness, whereas self-assessed subject knowledge showed limited association. A perception–performance gap emerged, with students often rating GenAI outputs as useful even when the instructor-side evaluation identified limitations in correctness or required substantial human scaffolding. The proposed framework should be interpreted as an exploratory guideline for studying and guiding GenAI use in engineering education, rather than as a definitive benchmark of model performance.

Keywords: Generative AI; student perceptions; engineering education; ChatGPT; cognitive depth; case study; multi-site; model comparison; hardware-oriented course; perception vs. performance



Academic Editor: Andrea C. Borowczak

Received: 15 April 2026

Revised: 13 May 2026

Accepted: 17 May 2026

Published: 20 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The emergence of GenAI has become one of the most significant technological developments in recent decades, with applications spanning creative content generation, programming assistance, and complex problem-solving (Kong & Yang, 2024; Luo et al., 2025; Martínez, 2024; Romera-Paredes et al., 2024). Large language models (LLMs), such as ChatGPT, GPT-4, and their successors, have introduced opportunities and challenges for higher education. These models are capable of generating coherent, contextually relevant responses to a wide range of prompts, enabling new forms of interactive learning. In technical disciplines such as digital systems design, problem-solving often integrates conceptual understanding with procedural precision. These tools offer potential benefits in accelerating learning by providing rapid feedback, alternative solution strategies, and debugging support (Cui et al., 2025; Noy & Zhang, 2023; Shailendra et al., 2024).

Despite these advantages, the pedagogical implications of GenAI remain complex. Key concerns involve academic integrity, overreliance on automated reasoning, and the erosion of analytical and metacognitive skills (Guedes et al., 2025; Qadir, 2023). Empirical studies have revealed performance limitations in tasks that demand multi-step reasoning, the integration of diverse concepts, or hardware-specific expertise, along with the risks of plagiarism and misplaced confidence in incorrect AI-generated solutions (Álvarez Ariza et al., 2025; Oh, 2025; Shallari & Hussain, 2024). These challenges highlight the need for systematic investigations that balance technological enthusiasm with a critical evaluation of GenAI's educational impact (Gong et al., 2025).

Within engineering and computing disciplines, GenAI also raises questions about trust, transparency, and educational validity. Students may perceive AI-generated outputs as authoritative due to their linguistic fluency, even when the underlying reasoning is flawed (Hussain et al., 2025; Zhao et al., 2025). Conversely, educators must find ways to integrate these tools to enhance learning outcomes without undermining assessment rigor or skill development. Addressing these tensions requires evidence-based approaches that quantify both model performance and student perception, allowing educators to better understand how learners interpret, evaluate, and apply AI-generated content (Corbin et al., 2025).

This study investigates the correspondence between students' perceptions of GenAI effectiveness and the instructor-benchmarked reference evaluation of the models, as assessed through systematic teacher interactions. Three OpenAI models, i.e., GPT-4.1, GPT-O3, and GPT-4.5, were assessed in two digital systems design courses from graduate and undergraduate degrees in Italy and Sweden, respectively. Based on a set of tasks categorized by cognitive depth, the study explores how learners engage with these systems, how prior familiarity shapes interaction patterns, and how perceived usefulness aligns with an instructor-benchmarked reference evaluation. By combining quantitative evaluation with qualitative perception data, the study aims to support evidence-based strategies for leveraging GenAI in engineering education, while safeguarding skill development, academic integrity, and equitable access. This study addresses four key gaps in the existing literature:

- **Perception–Performance Alignment**—We introduce a paired student–teacher framework that (i) captures students' perceptions for a given set of tasks and (ii) benchmarks the same tasks via teacher interactions, enabling an indirect perception–reference alignment analysis.
- **Cognitive Depth Analysis**—We operate on four cognitive-depth categories (conceptual, procedural, application, and synthesized) and a reduced-complexity protocol that links the assistance a model needs to the prior knowledge a student must supply.
- **Hardware-Oriented Context**—Focused on digital systems design that is a domain characterized by strict syntax, hardware-specific constraints, and engineering complexity, this research extends the scope of GenAI studies beyond general engineering and writing contexts.
- **Model Comparison**—We compare three OpenAI models (GPT-4.1, GPT-4.5, and GPT-O3), highlighting multimodality and capability differences and establishing a clear performance hierarchy relevant to educational deployment.

The remainder of this article is organized as follows: Section 2 provides an overview of the state-of-the-art research on GenAI in higher education, followed by Sections 3 and 4 presenting the methodological and statistical framework utilized in this study. In Section 5, the statistical results from the cross-examination of tasks from different cognitive depths are introduced. This is followed by a discussion on their implications for similar engineering education courses in Section 6 and summarized conclusions in Section 7.

2. Related Work

GenAI models have captured significant research interest, especially regarding their implications in higher education. Existing literature has examined the pedagogical opportunities and challenges of integrating GenAI into learning environments, focusing on areas such as personalized learning, intelligent tutoring, and automated assessment.

2.1. AI Chatbots and GenAI in Education

Early work on educational chatbots emphasized their potential to improve access to learning resources and streamline administrative support. [Allen et al. \(2024\)](#) introduced Q-Module-Bot, a GenAI-based Q&A system designed to integrate with virtual learning environments (VLEs) and address challenges in student engagement and educator workload. This system demonstrated the value of leveraging GPT-based architectures for delivering real-time, source-attributed responses, while highlighting the importance of stakeholder-centered design for adoption.

Recent reviews and position papers have shown that GenAI is rapidly becoming embedded in higher education as a tool for tutoring, content generation, feedback, explanation, and learning support ([Kasneci et al., 2023](#); [Lo, 2023](#); [Yan et al., 2024](#)). These studies highlight the potential of LLM-based systems to support personalized learning, provide immediate feedback, and assist students in complex academic tasks. At the same time, they also emphasize persistent risks, including hallucinated or incomplete answers, overreliance on automated reasoning, academic integrity concerns, and the need to redesign assessment practices ([Bittle & El-Gayar, 2025](#); [Yan et al., 2024](#)).

Therefore, GenAI should not be treated only as a productivity tool but as a socio-technical system that changes how students formulate questions, evaluate information, and demonstrate learning ([Kasneci et al., 2023](#); [Lo, 2023](#)). Prior reviews on ChatGPT and LLMs in education have consequently called for empirical studies that move beyond general attitudes and investigate how students use these systems in authentic disciplinary tasks ([Yan et al., 2024](#)).

2.2. Student Perceptions and Trust in GenAI

Understanding learners' attitudes toward AI tools is central to effective adoption. [Tossell et al. \(2024\)](#) examined student perceptions of ChatGPT use in a college essay assignment, finding that while learners valued AI for idea generation and feedback, they expressed reservations about accuracy and resisted fully autonomous AI grading. The study revealed an evolution in perception, from seeing ChatGPT as a "cheating tool" to a collaborative resource, provided that human oversight remains central.

A second relevant strand of literature concerns student perception, AI literacy, and trust calibration. Previous studies have shown that students may perceive GenAI tools as useful for idea generation, feedback, explanation, and problem-solving, while still expressing concerns about accuracy, transparency, and the need for human oversight. Recent work on GenAI literacy further suggests that effective use depends not only on disciplinary knowledge but also on students' ability to formulate prompts, interpret uncertainty, identify incorrect outputs, and decide when verification is required [Park \(2025\)](#).

This issue is particularly relevant because fluent LLM responses may create an impression of correctness even when the underlying reasoning is incomplete or technically wrong. Recent evidence shows that users may misjudge what LLMs know and may overestimate the reliability of model responses when these are expressed in a coherent and confident form ([Steyvers et al., 2025](#)). From this perspective, perceived usefulness and actual task performance should be investigated together, since positive student perceptions do not necessarily imply reliable learning support.

Similarly, [Hussain et al. \(2025\)](#) explored students' acceptance of AI-driven support tools in programming education, noting that trust and transparency strongly influence adoption. These findings align with broader literature emphasizing that sustained engagement depends on both perceived usefulness and ethical assurance.

2.3. Domain-Specific Implementations and Pedagogical Impact

Several studies have demonstrated domain-specific applications of GenAI in engineering and computing education. [Kirova et al. \(2024\)](#) reported the use of LLMs for generating assessment tasks in data science courses, while [Savelka et al. \(2023\)](#) evaluated GPT's effectiveness in solving complex programming exercises, identifying prompt engineering and domain fine-tuning as key factors for accuracy.

In engineering and computing education, these issues become especially relevant because tasks often require formal syntax, implementation constraints, simulation, debugging, and integration of multiple concepts. Recent reviews in engineering education indicate that LLMs can support explanation, ideation, coding, design activities, and formative feedback but also require careful instructional integration to avoid superficial learning and uncritical dependence on generated outputs ([Filippi & Motyl, 2024](#)).

In computing education, GenAI tools have been shown to support code generation, explanation, and debugging, but they also shift part of the learning process from producing a solution to evaluating, correcting, and validating AI-generated artifacts ([Prather et al., 2023](#)). Recent work has therefore proposed prompt-based programming exercises as a new instructional format in which students are explicitly required to formulate prompts, inspect AI-generated code, and reason about its correctness ([Denny et al., 2024](#)).

Nevertheless, less is known about how this dynamic operates in hardware-oriented domains such as FPGA, VHDL, Verilog, and digital systems design, where correctness depends not only on textual plausibility but also on functional constraints, simulation validity, and implementation feasibility.

Other works have applied GenAI to enhance inquiry-based and gamified learning environments. For example, [Huber et al. \(2024\)](#) used LLMs to generate interactive educational games, aiming to counter overreliance by fostering active problem-solving. In engineering education, [Nguyen et al. \(2024\)](#) showed that ChatGPT could assist with computational problem-solving but required structured oversight to ensure conceptual correctness.

2.4. Distinct Contributions of the Present Study

Prior work has documented the promise and pitfalls of GenAI in education, addressing student trust and acceptance alongside domain-specific uses in programming courses. However, to the best of our knowledge, there are no cross-studies of learners' perceptions with regard to instructor-benchmarked reference evaluation of task performance. This study addresses that gap by triangulating student-perceived usefulness with teacher-evaluated correctness on real field-programmable gate array (FPGA)/VHDL/Verilog tasks, benchmarking three OpenAI models across conceptual, procedural, application, and synthesized tasks, and doing so in a multi-site setting (Sweden and Italy). The proposed methodology offers a practical lens for designing AI-supported learning that calibrates trust, promotes AI literacy, and preserves rigor in engineering education.

Against this background, the present study contributes to the literature by explicitly comparing students' perceived usefulness with an instructor-benchmarked reference evaluation of GenAI outputs on the same family of technical tasks. This focus extends prior work on GenAI adoption in higher education ([Kasneci et al., 2023](#); [Lo, 2023](#); [Yan et al., 2024](#)), AI literacy and trust calibration ([Park, 2025](#); [Steyvers et al., 2025](#)), and domain-specific GenAI

use in engineering and computing education (Denny et al., 2024; Filippi & Motyl, 2024; Prather et al., 2023).

The specific contribution of the present study is therefore not only to examine whether students find GenAI useful but also to analyze whether such perceived usefulness remains aligned with task-level technical performance across cognitive-depth categories in a hardware-oriented engineering context.

3. Methodology

3.1. Study Design

The aim of this study is to examine the alignment between students' perceptions and an instructor-benchmarked reference evaluation of commercial GenAI models in domain-specific tasks. This is achieved by employing a mixed-methods approach to collect results from two courses in undergraduate (bachelor's) and graduate/postgraduate (master's) related to digital systems design for FPGA. Through tailored surveys for each set of tasks, we captured students' perceptions of the effectiveness of GenAI tools. This was later compared to teacher-based tests of these GenAI tools on the same tasks, associating scores with the answers obtained.

The flowchart in Figure 1 describes the methodological framework utilized in this study. Initially, a set of tasks under study, together with cognitive-depth categories, was defined to ensure accurate task categorization. In Figure 1, the upper part of the flowchart represents the researcher-defined setup phase, including task selection and cognitive-depth categorization. The left branch represents the student track, in which students report their GenAI use, perceived usefulness, perceived correctness, and further interaction. The right branch represents the teacher track, in which instructors benchmark GenAI models using the same task set and the reduced-complexity protocol. Rectangular boxes denote actions or evaluation steps, diamond-shaped boxes denote decisions, and rounded terminal boxes denote final outcomes. The student and teacher tracks were not intended to reproduce identical interaction conditions. In the student track, students reported their own GenAI use and perceptions in authentic course settings. In the teacher track, selected GenAI models were evaluated through a standardized reduced-complexity protocol. Therefore, the comparison should be interpreted as an indirect perception–reference alignment analysis, rather than as a controlled comparison under equivalent model, prompt, and interaction conditions. Cognitive depth refers to the level of cognitive processing necessary to adequately address specific tasks, ranging from basic recall to complex analytical and evaluative thinking. The four levels of cognitive depth utilized in this study are described in Table 1. Moreover, in Table 2, some representative examples of tasks for each cognitive-depth category are shown.

Table 1. Cognitive depth used in this study.

Cognitive Depth	Requirements
Conceptual	Applying a concept in a simplified context
Procedural	Applying a well-defined set of procedures
Application	Deploying on real-world setups
Synthesized	Integrating multiple concepts and critical thinking

Table 2. Representative examples of tasks for each cognitive-depth category.

Cognitive Depth	Representative Task Example
Conceptual	Simplify a Boolean function using a Karnaugh map and explain the resulting logic expression.
Procedural	Write a VHDL/Verilog module implementing an 8-bit adder and provide a basic test bench.

Table 2. Cont.

Cognitive Depth	Representative Task Example
Application	Implement and verify a memory-access component in an FPGA-oriented design flow, considering input/output constraints and simulation results.
Synthesized	Design a simplified CPU datapath integrating instruction fetch, branching, memory access, and arithmetic operations.

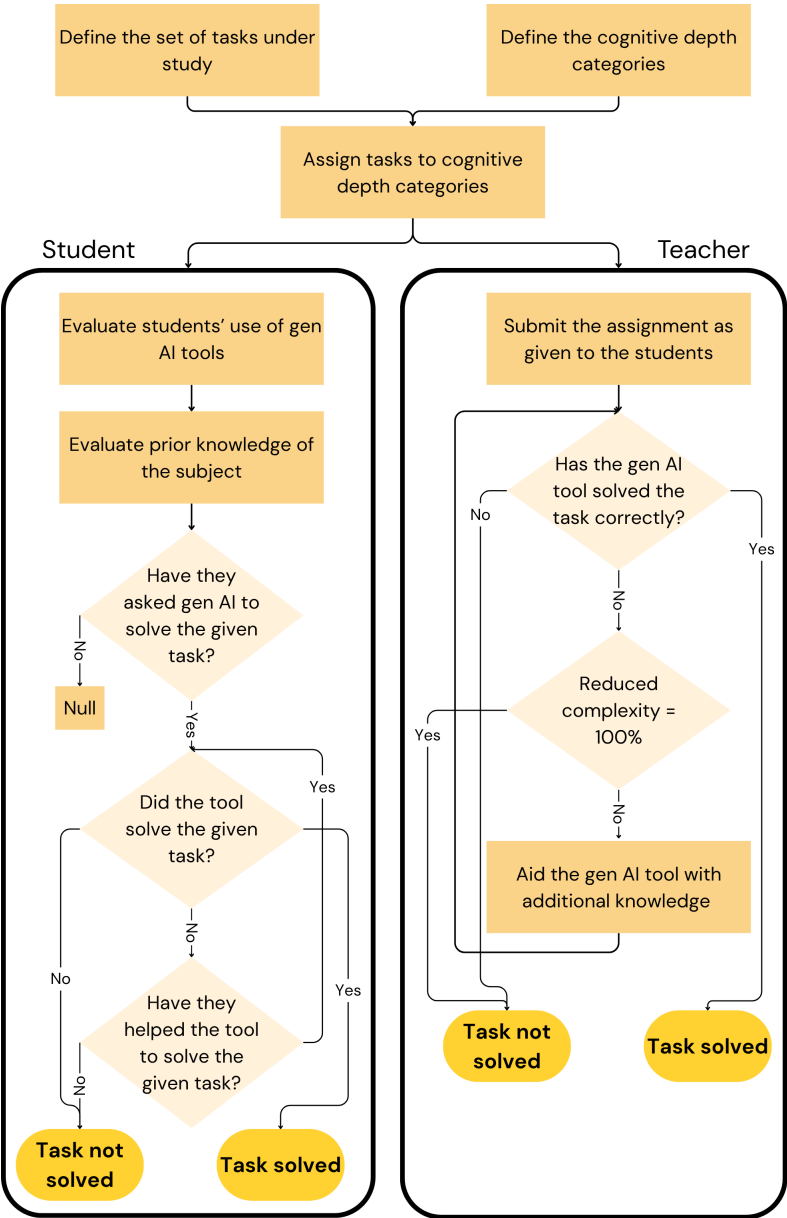


Figure 1. Flowchart of the study methodology. The upper part represents researcher-defined setup steps, while the left and right branches represent the student survey track and teacher benchmarking track, respectively. Rectangles indicate actions, diamonds indicate decision points, and rounded boxes indicate terminal outcomes.

The methodology was further bifurcated into two parallel evaluations, denominated as the student and teacher tracks. The student track captures students’ interactions with GenAI tools utilizing a set of surveys described in more depth in Section 3.2.2. In parallel, the teacher track provides an instructor-benchmarked reference evaluation of the outputs generated by three GenAI models (GPT-4.1, GPT-O3, and GPT-4.5) on the same task set.

The two tracks were designed to provide complementary insights into students' perceived effectiveness of GenAI tools and instructors' reference evaluation of model outputs on the same task set. In the iterations of solving a given task with a GenAI model, the teacher iteratively reduces the task complexity following the steps described in Table 3. The purpose of introducing these reduced-complexity steps is to associate the necessary knowledge a student would be required to have for the given task to help the tool solve the problem. The higher the cumulative percentage, the greater the knowledge the student needs in order to identify errors in the model and to guide it to a correct solution. The final step, providing the full answer, means that the tool was unable to solve the task throughout all the iterations performed.

Table 3. Reduced complexity of a given task represented as a cumulative percentage (%).

Complexity Reduction Steps	Cumulative Amount (%)
Full complexity task	0
Decomposed task	25
Generic error solving	45
Scoped error solving	70
Solution integration	90
Full solution provided	100

This structured reduction framework, shown in Table 3, allows for a quantitative mapping between AI performance limitations and the level of human expertise required for task completion. In particular, for each task, the instructor interacted with each GenAI model using a standardized starting prompt aligned with the official assignment specification. If the initial response was incorrect or incomplete, the instructor proceeded through the predefined reduced-complexity steps reported in Table 3 until a correct solution was obtained or the full solution had to be provided. The minimum step required to reach correctness was recorded as an operational measure of the amount of human guidance needed for task completion and used to derive the reference usefulness benchmark. Correctness was evaluated against the task requirements and reference solutions typically used in the course context, focusing on technical validity and completeness with respect to the specification. To improve reproducibility, the instructor-side evaluation followed a structured rubric. A GenAI response was considered correct only when it satisfied the functional requirements of the task, respected the required HDL or FPGA-related constraints, and provided a technically complete solution consistent with the reference solution used in the course. Responses were considered incomplete when they addressed the general direction of the task but omitted essential implementation details, violated relevant constraints, or required additional domain-specific clarification. Responses were considered incorrect when they failed to implement the requested functionality, contained substantial technical errors, or could not be mapped to a valid solution path. For each task, the instructor first submitted the full-complexity prompt. If the model response was not correct according to the rubric, the instructor moved to the next reduced-complexity step reported in Table 3. The process continued until either a correct solution was obtained or the full solution had to be provided. The minimum reduction step required to reach a correct answer was recorded as the reference indicator of the amount of human guidance needed for task completion. The teacher-side evaluation was performed by one instructor for each course. Each instructor evaluated the tasks belonging to the corresponding course and followed the same reduced-complexity protocol and rubric described above. The resulting scores should therefore be interpreted as an instructor-benchmarked reference within the study design, rather than as a multi-rater consensus evaluation.

3.2. Survey

Survey data were collected during the active delivery of the two selected courses in 2025. At the time when students received instructions for their respective assignments or projects, they were also provided with access to the accompanying survey. This approach ensured that students' perceptions and interactions were captured in real-time, closely aligning their responses with the actual tasks they were completing. To maintain methodological consistency, the study adopted a common survey logic, a shared set of response variables, the same cognitive-depth categories, and the same instructor-side reduced-complexity protocol across both institutions. However, task content, programming language, academic level, and local course implementation were course-specific. Accordingly, standardization should be understood at the level of the analytical framework and evaluation protocol, not as identical prompts, assignments, or student interaction conditions across the two cohorts. Participation in the surveys was voluntary and anonymous.

3.2.1. Courses

For this study, two courses focusing on digital systems design were selected to capture a comprehensive range of applications and instructional contexts. One undergraduate-level (bachelor's) course was delivered at Mid Sweden University (MIUN) in Sweden, focusing on FPGAs using VHDL as the hardware description language. This course provided suitable contexts for analyzing GenAI effectiveness in tasks related to hardware modeling and digital circuit design. There were 50 students enrolled in the course, of which 32 participated in the surveys.

The second course, a graduate-level course offered at the University of Salerno (UNISA) in Italy, centered around FPGA design using Verilog. This course provided an advanced instructional context, ideal for evaluating GenAI's performance in higher-complexity tasks, including sophisticated digital circuit modeling and verification processes. There were 25 students enrolled in this course, and the survey was answered by 20 students.

An important aspect of these courses is their integration of conceptual and procedural knowledge into applied knowledge through specific hardware devices and specialized software development environments. Evaluating GenAI tools in this context is particularly relevant, as their effectiveness may significantly depend on access to proprietary information and specific details of these devices and software environments. Understanding this dependency provides valuable insights into both the capabilities and limitations of GenAI in applied digital systems design contexts.

3.2.2. Survey Structure

The survey was structured into several distinct sections, each designed to capture specific dimensions of students' interactions and perceptions regarding GenAI tools. In the initial section, the survey aimed to understand students' overall approach towards using GenAI tools for university assignments in general. This included questions assessing their familiarity with such tools, frequency of their use, and the types of academic tasks for which students typically employ GenAI.

Subsequently, the survey collected data on students' prior knowledge of digital systems design via self-assessment. Students were specifically asked, "How would you rate your previous knowledge of digital electronics?" using a 5-point Likert scale to establish a baseline understanding of their existing competencies and confidence related to the course material.

The survey then progressed to task-specific questions aligned with the aspects of the assignments or project under study. For each critical aspect evaluated, the survey first established whether students employed a GenAI tool. If students indicated affirmative

usage, three consistent follow-up questions were posed across all tasks. The first follow-up question assessed perceived usefulness, asking students to rate the helpfulness of the AI-generated answer on a 5-point Likert scale. For the statistical analyses involving perceived usefulness, task-specific Likert-type ratings were aggregated into an average usefulness score. Therefore, perceived usefulness was analyzed as a bounded quasi-continuous variable rather than as an individual ordinal response or a count outcome. The second follow-up question evaluated correctness explicitly, inquiring whether the AI-provided solution was accurate (yes/no). Finally, the third question probed further interaction, seeking to determine whether students subsequently corrected the tool or provided additional input or information to refine the solution.

3.3. Surveyed Tasks

The tasks evaluated in this study originate from two academic levels—undergraduate and graduate—and were therefore presented differently to the respective student cohorts. In the undergraduate course at MIUN, the tasks were organized modularly into four sequential assignments, each increasing in both complexity and cognitive depth. The first assignment integrated conceptual and applied knowledge through exercises involving Karnaugh maps, Boolean optimization, and TTL-chip implementation of specified logic functions. The subsequent two assignments emphasized procedural cognitive depth, requiring students to develop HDL (hardware description language) code for 8-bit adders and other basic arithmetic circuits. The final assignment targeted synthesized cognitive depth, challenging students to design a CPU with predefined opcodes and operations.

Conversely, the graduate course at UNISA was structured as a comprehensive project encompassing all cognitive-depth categories. The project's objective was the design of a 32-bit CPU featuring the following elements:

- Dual RAM configurations for simultaneous access of instructions and data.
- Mechanism for storing results in the RAM.
- Support for floating-point operations (addition, subtraction, and fraction).
- Implementation of branching instructions.
- Single-cycle instruction fetch operation.

The combination of these tasks constitutes a robust and representative set for assessing the application of GenAI tools in digital systems design education.

3.4. GenAI Tools

To provide an instructor-benchmarked reference evaluation of GenAI model outputs across the two selected digital systems design courses, this study considered three OpenAI models: GPT-4.1, GPT-O3, and GPT-4.5. The rationale behind selecting multiple models stemmed from their varying capabilities and accessibility, which influence their effectiveness in academic settings. The teacher-side GenAI benchmarking was conducted in 2025, using the model versions available at the time of the experiments. Because GenAI systems evolve rapidly, the reported results should be interpreted as a time-specific evaluation rather than as a permanent characterization of these models. GPT-4.1, the free version, was included due to its widespread accessibility among students, making it an attractive option for educational use. However, it is the most lightweight model among the three, inherently limiting its problem-solving capabilities. Additionally, unlike the paid models, GPT-4.1 only supports text-based inputs, restricting its functionality in tasks requiring multimodal interaction or complex procedural input.

GPT-O3, a paid model, offers enhanced reasoning and problem-solving capabilities compared to GPT-4.1, supporting more sophisticated interactions and higher cognitive complexity tasks. Its ability to handle multimodal inputs and generate more nuanced responses

makes it suitable for tasks involving advanced procedural and conceptual integration, especially in synthesized knowledge.

GPT-4.5, also a paid model, represents the most advanced option utilized in this study. It provides superior performance in complex analytical and evaluative tasks, capable of managing intricate instructions and offering detailed, accurate solutions. Its robust performance makes it especially effective for tasks demanding precise technical knowledge. By comparing these three GenAI models, the study explores the implications of model accessibility, input limitations, and advanced cognitive processing capabilities on their effectiveness in educational contexts.

Based on these capability differences, the expected performance hierarchy was GPT-O3, followed by GPT-4.5 and GPT-4.1. This expectation was operationalized through the reduced-complexity protocol: a stronger model was expected to solve a larger fraction of tasks at full complexity or after fewer reductions.

4. Statistical Methodology

The statistical analysis explored three main relationships: students' prior familiarity with GenAI tools, their interaction behavior during the assignments, and their perceived usefulness of the AI-generated outputs. The aim was to assess whether prior GenAI experience or domain-specific knowledge was associated with how students used and evaluated these tools. Since the dataset included ordinal, categorical, count, and aggregated quasi-continuous variables, different statistical methods were selected according to the measurement level of each outcome.

The methodological framework consisted of four main analytical stages. Because perceived usefulness was assessed only when students reported actual GenAI use, the inferential analyses involving this variable were not based on the full respondent pool. First, a descriptive statistical analysis was conducted to summarize the dataset's main characteristics and provide an initial understanding of variable distributions. Measures of central tendency and dispersion (median, interquartile range, minimum, and maximum values) were used in place of means and standard deviations because normality assumptions were not met based on Shapiro–Wilk tests. This descriptive stage included the characterization of students' prior familiarity with GenAI tools, self-assessed subject knowledge, perceived usefulness of AI outputs, and the number of iterative interactions with the tools during assignments. Graphical representations such as boxplots and scatter plots were employed to visually compare patterns across the two institutions involved—MIUN and UNISA—and to identify potential trends warranting deeper inferential analysis.

Second, inferential testing was performed to assess statistically significant differences among categorical groups. Specifically, the Kruskal–Wallis test was applied to ordinal variables (for example, levels of GenAI familiarity and prior knowledge) to assess differences in the median number of tool interactions or perceived usefulness scores across groups. This nonparametric test was chosen because it does not assume normality or homoscedasticity, thereby ensuring validity for small, uneven sample sizes typical of survey-based educational studies. For numerical or quasi-continuous variables, monotonic relationships were further examined using Spearman's rank correlation coefficients, thereby quantifying associations between interaction counts and perceived usefulness.

Third, whenever statistically significant differences were detected in the Kruskal–Wallis analysis, post hoc regression models were fitted according to the measurement level of the dependent variable. For count outcomes, such as the number of interactions performed by students with GenAI tools, Poisson regression models were used. This approach is appropriate for count data and allows the effect of GenAI familiarity and university affiliation to be expressed through rate ratios. For perceived usefulness, however, Poisson

regression was not used, since this variable was collected through 1–5 Likert-type ratings and analyzed as an aggregated mean score. Therefore, a general linear model fitted by ordinary least squares was used for post hoc analysis of average perceived usefulness. In this case, coefficients were interpreted as average changes in perceived usefulness scores associated with the predictors, rather than as multiplicative rate ratios.

A fourth step examined the consistency of the observed associations across the UNISA and MIUN datasets. This comparison was used to assess whether similar effect directions emerged in two distinct educational contexts. Since the two courses differ in academic level, task structure, language context, and assignment design, the institutional indicator was interpreted as a contextual covariate rather than as evidence of specific cross-context mechanisms.

Overall, this structured multi-stage analysis enabled a comprehensive examination of both behavioral and perceptual dimensions of GenAI tool usage in engineering education. The combination of descriptive visualization, nonparametric inference, Spearman's rank correlation, and outcome-specific regression models provided complementary insights: descriptive analyses highlighted trends and variability; Kruskal–Wallis and correlation tests assessed statistically significant associations; Poisson regression quantified effects on count outcomes, whereas general linear modeling was used for aggregated perceived usefulness scores. The inclusion of a two-cohort comparison provided an additional perspective on the consistency of the observed associations presented in Section 5.

5. Results

5.1. Effects of Prior Familiarity with GenAI Tools on Subsequent Interactions with the Tool During the Assignment

Descriptive analysis was first conducted to examine the relationship between students' prior familiarity with GenAI tools and the number of subsequent interactions they performed with these tools during the completion of course assignments. Boxplots representing this relationship are shown in Figure 2a (UNISA) and Figure 2b (MIUN). Across both institutions, a consistent pattern emerged: students with greater prior familiarity with GenAI tools engaged more frequently in iterative interactions, such as refining prompts, contextualizing inputs, or revising partial outputs. In contrast, students with low or no prior exposure to these tools tended to accept initial answers without further prompting. This behavior suggests a progressive learning curve, with more experienced users demonstrating greater critical engagement in AI-assisted workflows.

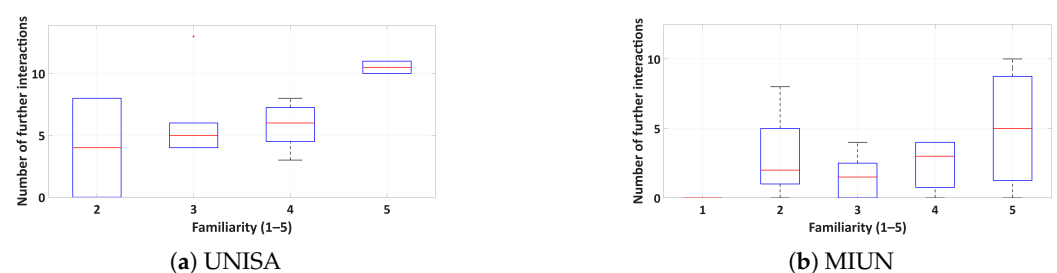


Figure 2. Comparison of students' previous familiarity with GenAI tools and the number of further interactions during the survey.

Particularly in UNISA, where tasks required higher technical specificity, the frequency of interactions increased noticeably in correlation with familiarity levels. While MIUN students displayed a similar trend, the variability of responses was less pronounced. This discrepancy between the two institutions may reflect differences in course and task structure (for example, assignment granularity) and should therefore be interpreted cautiously.

The observed differences were validated by a Kruskal–Wallis test, which yielded statistically significant results ($\chi^2 = 18.12$, $p < 0.001$), thereby supporting the hypothesis that prior familiarity influences subsequent usage patterns during assignment execution. To further substantiate this association, a post hoc generalized linear model (GLM) was employed, using a Poisson regression framework. This framework is particularly well-suited for count data, such as the number of interactions. The model incorporated two predictors: GenAI tool familiarity and university. Model results confirmed the significance of both predictors:

- GenAI tool familiarity (Estimate = 0.511, $SE = 0.0758$, $\exp(\beta) = 1.67$, $p < 0.001$) indicates that for each increase in familiarity level, the expected number of interactions grows by a factor of 1.67.
- University (UNISA vs. MIUN) was also a significant factor (Estimate = 1.417, $SE = 0.1648$, $\exp(\beta) = 4.12$, $p < 0.001$), suggesting contextual differences between the two cohorts. Since the two groups differ in academic level, course structure, language context, and assignment design, this coefficient should not be interpreted as a purely institutional effect.

The overall model fit was acceptable, as indicated by the Pseudo- R^2 of 0.465. Nonetheless, the findings robustly support the initial descriptive trend: prior familiarity with GenAI tools was significantly associated with richer, more interactive engagement patterns during assignment completion.

5.2. Effects of Previous Knowledge of the Subject on Subsequent Interactions with the Tool During the Assignment

After establishing the relationship between GenAI familiarity and engagement, we next examined whether students' subject-specific knowledge exerted a similar influence. The following analysis explores the correlation between students' self-assessed prior knowledge of digital system design and the number of subsequent interactions they engaged in with the GenAI tool during assignment completion. As illustrated in Figure 3a (UNISA) and Figure 3b (MIUN), the distribution of interactions as a function of the declared subject familiarity.

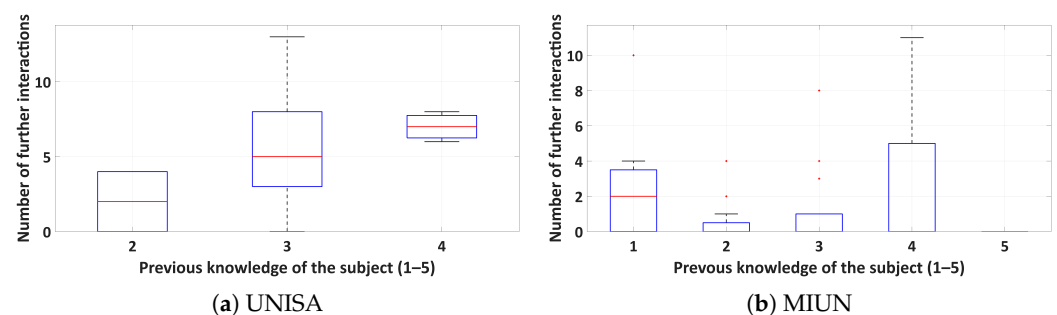


Figure 3. Comparison between students' previous knowledge of the subject and the number of further interactions during the survey.

The boxplots reveal a less structured trend in comparison to that observed with GenAI tool familiarity. In both institutions, students who self-assessed their knowledge as medium to high did not consistently exhibit higher interaction counts. At UNISA, a moderate increase in interaction count is seen among students who reported higher prior knowledge. However, the spread of data points is wide, indicating variability in behavior. At MIUN, the relationship appears to be even more attenuated, with interaction counts remaining relatively flat across different levels of familiarity.

This descriptive pattern suggests that technical confidence alone does not necessarily drive iterative interaction with AI tools. In fact, students with limited prior knowledge may have relied more passively on GenAI outputs. However, even those with higher self-assessed competence did not always engage critically with the models, potentially indicating overconfidence or a lack of metacognitive strategies.

The observed variations in the distribution of interactions across knowledge levels were not statistically significant, as indicated by the Kruskal–Wallis test ($\chi^2 = 4.00$, $p = 0.405$). This supports the interpretation that subject-specific knowledge does not independently predict the number of GenAI interactions. Due to the absence of statistical significance in the overall test, no post hoc model was conducted for this comparison.

5.3. Effects of Previous GenAI Familiarity on Perceived Usefulness of the Tool for the Assignment

The present study examined the relationship between students' prior familiarity with GenAI tools and their perceived usefulness of the generated answers. Figure 4a (UNISA) and Figure 4b (MIUN) illustrate the distribution of perceived usefulness scores across different familiarity levels.

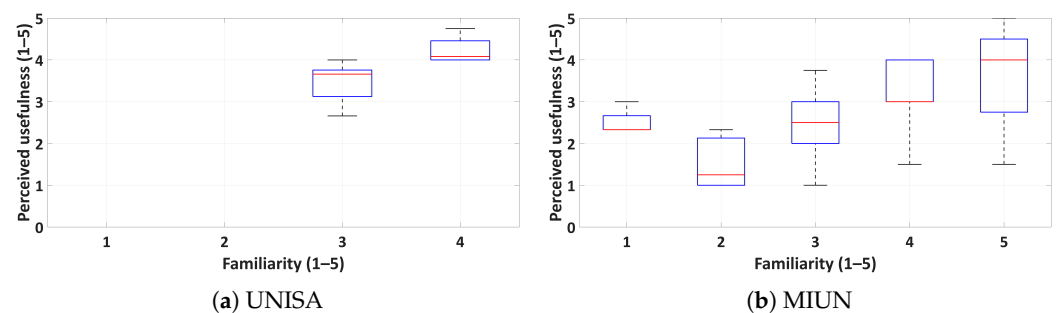


Figure 4. Comparison between students' previous familiarity and the perceived usefulness of the tool.

The descriptive analysis revealed a positive association: students who reported higher levels of familiarity tended to attribute greater usefulness to the GenAI outputs. At UNISA, the increase in perceived usefulness with familiarity is particularly evident, with the median score rising consistently across familiarity categories. This pattern suggests that familiarity not only influences the quantity of interaction, as previously discussed, but also shapes the way students evaluate the relevance and applicability of the AI-generated content. In MIUN, the trend follows a similar direction but at a more gradual pace, with a narrower gap between the lowest and highest familiarity groups.

These results suggest that students with more prior exposure to GenAI tools have developed more effective prompting strategies, enabling them to elicit higher-quality answers and thus perceive them as more useful. The Kruskal–Wallis test confirmed the significance of the observed differences ($\chi^2 = 18.15$, $p = 0.001$), indicating that prior familiarity significantly affects the perceived usefulness of GenAI outputs.

To further explore this relationship, a general linear model fitted by ordinary least squares was employed, using average perceived usefulness as the dependent variable and GenAI tool familiarity and university affiliation as predictors. This modeling choice was adopted because perceived usefulness was analyzed as an aggregated mean score rather than as a count outcome.

The overall model, fitted on the 42 students who reported GenAI use and provided valid average usefulness ratings, was statistically significant, $F(2, 39) = 18.6$, $p < 0.001$, with $R^2 = 0.488$, explaining 48.8% of the variance in average perceived usefulness scores. GenAI tool familiarity was a significant positive predictor, $\beta = 0.490$, $SE = 0.116$, 95% CI = [0.255, 0.725], $t = 4.22$, $p < 0.001$, indicating that each one-level increase in

prior familiarity was associated with an average increase of approximately 0.49 points in perceived usefulness. University affiliation was also significant, $\beta = 1.079$, $SE = 0.260$, 95% CI = [0.553, 1.605], $t = 4.15$, $p < 0.001$, suggesting contextual differences between the two cohorts.

Overall, higher prior familiarity with GenAI tools was associated not only with more frequent tool engagement but also with a more favorable evaluation of its outputs.

5.4. Effects of Students' Previous Knowledge of the Subject on Perceived Usefulness of the Tool for the Assignment

This comparison investigates the extent to which students' self-assessed prior knowledge of digital systems design affects their assessment of the usefulness of GenAI outputs during assignment completion. Figure 5a (UNISA) and Figure 5b (MIUN) show the distribution of perceived usefulness scores across different levels of prior subject knowledge. A descriptive analysis reveals an absence of consistent patterns across the two institutions. At UNISA, a slight upward trend can be observed for students reporting intermediate to high prior knowledge, but the differences between categories remain modest. At MIUN, the distribution is more uniform, with median usefulness scores showing minimal variation across the knowledge levels. In both contexts, the observed variability within each category suggests substantial individual differences, possibly reflecting differences in students' approaches to task formulation and AI tool prompting, regardless of their technical background.

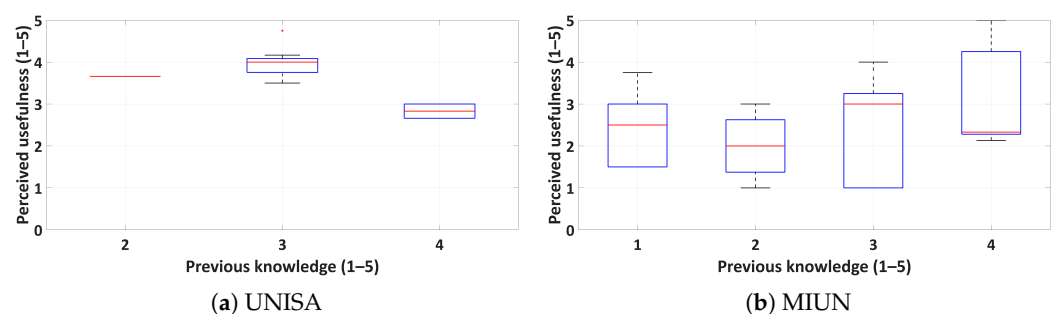


Figure 5. Comparison between students' previous knowledge of the subject vs. perceived usefulness of the tool.

This observation is consistent with the hypothesis that subject expertise alone does not necessarily enhance the perceived quality of AI-generated solutions. Even students with high prior knowledge may rely on the tool for confirmatory purposes rather than leveraging it to refine complex or ambiguous responses. The Kruskal–Wallis test results confirmed the absence of statistically significant differences ($\chi^2 = 3.26$, $p = 0.353$). Thus, the present study suggests that self-reported subject knowledge is not a strong predictor of perceived usefulness. Due to the absence of statistical significance, no post hoc general linear model was performed for this relationship.

5.5. Correlation Between the Number of Interactions of Students with GenAI Tools for the Assignment and Its Perceived Usefulness

This analysis examines the relationship between students' perceived usefulness of GenAI outputs and the number of subsequent interactions they engaged in with the tool during assignment completion. As illustrated in Figure 6a,b (UNISA, MIUN), the scatter plots with regression lines for each university demonstrate the alignment between usefulness scores and the number of follow-up interactions.

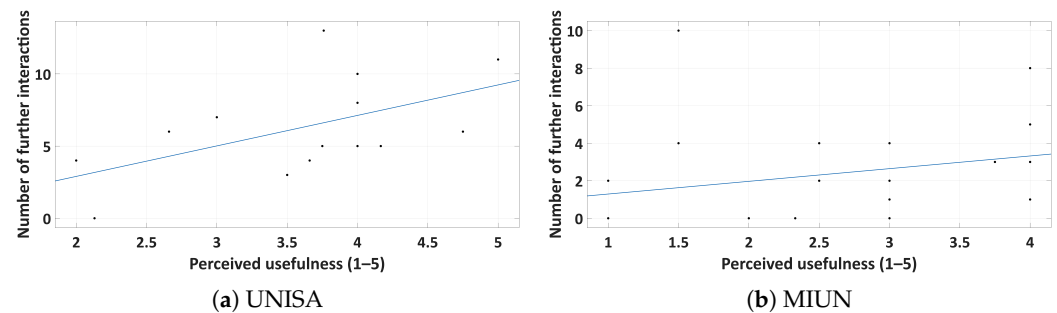


Figure 6. Regression between students' perceived usefulness of the tool vs the number of further interactions during the survey.

From a descriptive perspective, UNISA exhibits a clear positive trend: higher usefulness ratings are generally associated with more frequent interactions. This suggests that when students perceive the tool's responses as valuable, they are more inclined to refine and extend the exchange to obtain more precise or complete solutions. In contrast, MIUN data displays a weaker association, with greater dispersion and less evident alignment between usefulness ratings and interaction counts. This discrepancy may be due to differences in the structure of the assignment or in the degree of freedom students had when engaging with the AI tool.

When data from both institutions are aggregated, the trend becomes more evident, indicating that students who perceive higher usefulness are, on average, more proactive in interacting with the AI tool. The statistical analysis confirms these observations. At UNISA, the Spearman correlation between usefulness and further interactions is positive but not statistically significant in the UNISA subgroup ($\rho = 0.497$, $p = 0.060$), and is therefore reported as a trend, while at MIUN, the correlation is smaller and non-significant ($\rho = 0.221$, $p = 0.176$). When data from both universities are combined, the correlation becomes significant ($\rho = 0.429$, $p = 0.001$), reinforcing the interpretation of a general positive association across the overall student sample. Given that the primary analysis relied on correlation, no further post hoc modeling was conducted for this comparison.

5.6. Correlation Between the Number of Interactions of Students with GenAI Tools for the Assignment and Cognitive-Depth Category

This analysis explores the question of whether the cognitive-depth category of the task influences the number of additional interactions students perform with the GenAI tool after receiving the initial output. Figure 7 presents the distribution of further interactions for the four categories defined in Table 1.

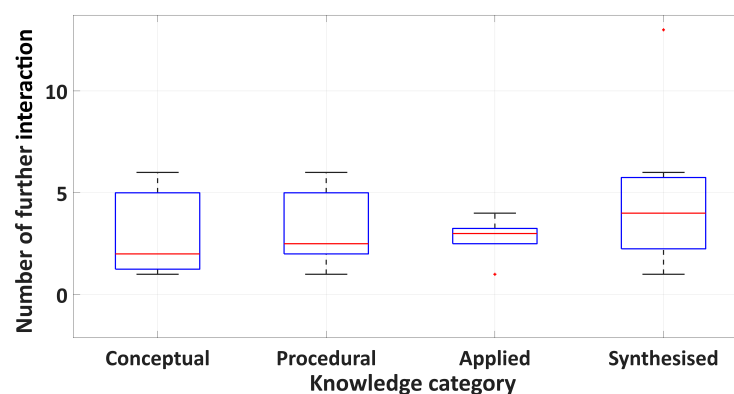


Figure 7. Comparison between the cognitive depth of a task and further interactions with the tool.

The descriptive analysis reveals only minor differences in the median and spread of interaction counts across categories. Tasks in the synthesized category, which require integrating multiple concepts, display slightly greater variability. However, the distributions overlap considerably across categories. Similarly, conceptual, procedural, and application tasks exhibit comparable ranges, with no category clearly associated with consistently higher or lower interaction counts.

These observations suggest that the cognitive depth of a task is not the sole determining factor in the extent of students' engagement in follow-up interactions with the AI tool. Instead, other factors, such as prior familiarity with GenAI or the perceived usefulness of the tool's output, are likely to have a greater influence on interaction patterns. The Kruskal–Wallis test confirmed the absence of statistically significant differences across the four cognitive-depth categories ($\chi^2 = 1.23$, $p = 0.746$). No post hoc analysis was performed due to the absence of statistical significance for this particular comparison.

5.7. Effects of Cognitive-Depth Category on the Perceived Usefulness of the GenAI Tool for the Assignment by the Student

This section investigates the impact of the cognitive-depth category of a task on the perceived usefulness of GenAI tool outputs to students. The four categories considered, as defined in Table 1, are conceptual, procedural, application, and synthesized. Figure 8 shows the distribution of perceived usefulness scores across different cognitive-depth categories. The descriptive analysis indicates modest differences: procedural and application tasks tend to receive slightly higher median ratings, while synthesized tasks show a greater degree of variability. Conceptual tasks present intermediate and relatively compact score distributions. The Kruskal–Wallis test did not detect statistically significant differences ($\chi^2 = 4.05$, $p = 0.256$), indicating that the cognitive depth of a task does not substantially influence students' perceived usefulness. No post hoc analysis was performed.

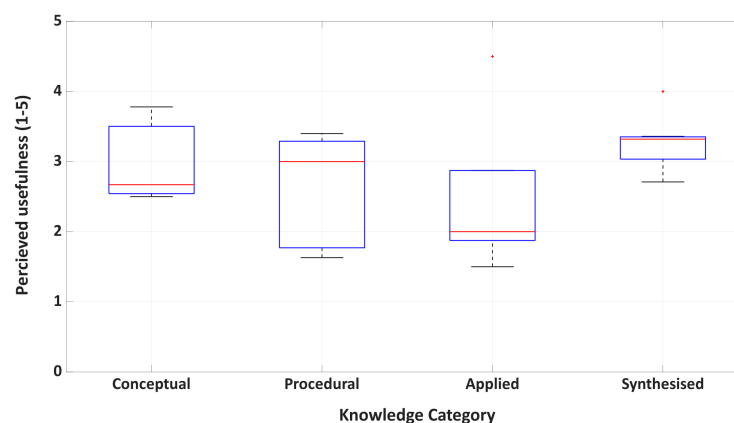


Figure 8. Comparison between the cognitive depth of the task and students' perceived usefulness of the tool.

5.8. Effects of Cognitive-Depth Category on Reference Usefulness

This analysis investigates how the cognitive-depth category of a task influences the reference usefulness of the GenAI tool. This is a benchmark measure derived from the instructors' evaluation of GPT-4.5 and GPT-4.1 responses to the course assignments, based on the four cognitive-depth categories defined in Table 1.

Figure 9a shows the reference usefulness distribution for GPT-4.5 across the cognitive-depth categories. The descriptive analysis highlights clear differences: The highest median reference usefulness was achieved by procedural tasks, reflecting the model's strong performance on structured, step-by-step problem-solving tasks. Application tasks also obtained relatively high scores, while synthesized tasks displayed greater variability, indicating

mixed performance when multiple concepts had to be integrated. Conceptual tasks were typically situated in an intermediate position. The Kruskal–Wallis test confirmed that these differences are statistically relevant ($\chi^2 = 15.6$, $p = 0.001$).

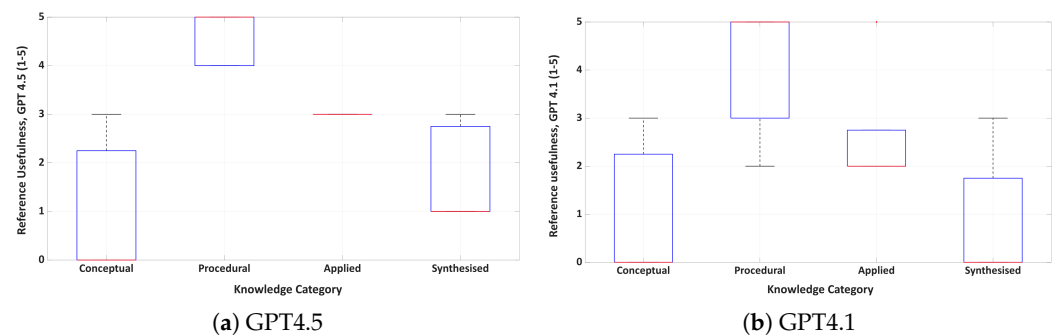


Figure 9. Comparison between the cognitive depth of the task and the reference usefulness of the tool.

As illustrated in Figure 9b, the reference usefulness distribution for GPT-4.1 shows a similar trend to GPT-4.5, albeit with generally lower scores. Procedural tasks, once again, had the highest median values, while synthesized tasks exhibited the widest spread, with some notably low evaluations. The Kruskal–Wallis test also indicated significant differences among categories ($\chi^2 = 13.2$, $p = 0.004$).

GPT-O3 showed the strongest instructor-benchmarked performance among the evaluated models. In the undergraduate tasks, GPT-O3 solved all tasks at full complexity, without requiring any reduced-complexity step. In the graduate-level tasks, GPT-O3 generally required only minimal complexity reduction, approximately corresponding to the decomposed task level, with applied tasks solved without additional simplification. This near-ceiling performance indicates that GPT-O3 required substantially less instructor scaffolding than GPT-4.5 and GPT-4.1 across the evaluated cognitive-depth categories.

In summary, GPT-O3 required the least instructor scaffolding within the adopted reduced-complexity protocol, followed by GPT-4.5 and GPT-4.1. This result should be interpreted as a study-specific reference evaluation based on the selected tasks and model versions, rather than as a definitive or general ranking of GenAI systems. This aligns with known LLM limitations in multi-concept synthesis (Chi et al., 2024).

6. Discussion

To maintain a direct link with our research questions, the Discussion Section is structured around the perception–performance gap and the observed behavioral predictors of GenAI engagement. Unless explicitly stated as a pedagogical implication, all statements below are grounded in the Results Section presented in Section 5. Broader considerations are provided as potential implications and directions for future work.

6.1. Model Comparison

In our analysis, we evaluated three GenAI models from OpenAI: GPT-4.1, GPT-4.5, and GPT-O3, representing varying problem-solving capabilities and accessibility (free or subscription-based). We systematically assessed the performance of each model on the set of course-specific tasks following the teacher evaluation pathway described in Figure 1. Consistent with the quantitative results, a clear performance hierarchy was observed.

GPT-4.1, the most lightweight and freely available model, showed the lowest overall performance. It demonstrated difficulties, particularly in graduate-level tasks requiring conceptual optimization, such as logic-function minimization using Karnaugh maps. Additionally, the model showed limitations in tasks categorized under applied cognitive depth, largely attributable to the need for precise, technology-specific information. How-

ever, the model performed better in procedural tasks, although even here, graduate-level assignments required higher complexity reductions due to intrinsic task complexity.

A critical limitation of GPT-4.1 stems from its text-only input capability, significantly hindering its performance on tasks demanding synthesized knowledge presented in multimodal formats. In both undergraduate and graduate assignments, transforming task requirements originally presented in combined textual, tabular, and schematic formats into purely text-based inputs posed considerable challenges. Merely accurately translating the tasks necessitated domain-specific expertise, equating to an implicit complexity reduction of approximately 25%, thereby shifting part of the cognitive load from the AI tool to the human operator during query formulation. Consequently, GPT-4.1 consistently required a complexity reduction of about 90% and even then did not always achieve a full solution of the task.

Conversely, GPT-4.5 showed a marked improvement across all cognitive depth domains but continued to experience difficulties in tasks involving conceptual optimization. For procedural and applied knowledge categories, GPT-4.5 reliably completed tasks following moderate complexity reductions ranging from 25–45. Furthermore, the complexity reductions required for synthesized knowledge tasks decreased notably as compared to GPT-4.1, now ranging between 45–90, with the model consistently producing complete solutions.

GPT-O3 demonstrated distinctly superior performance compared to the other two models. It successfully completed all undergraduate tasks without any reduction in complexity. For graduate-level tasks, GPT-O3 required only minimal complexity reduction ($\approx 25\%$) across all cognitive categories, with the exception of applied knowledge tasks, where it consistently achieved optimal performance without further simplification.

These findings confirm that effective technical problem-solving depends not only on linguistic fluency alone but also on the depth of model reasoning and its capacity for contextual integration. The variability observed in the synthesized task indicates that even GPT-4.5 struggles with higher-order abstraction.

6.2. Student Familiarity, Interaction, and Perceived Usefulness

Besides model capabilities, student behavior influenced the perceived usefulness. As shown in Sections 5.1 and 5.3, prior familiarity with GenAI tools was significantly associated with both the number of interactions and the perceived usefulness of the responses ($p < 0.001$ in both analyses). Students who were more familiar with GenAI tools engaged in longer prompting sessions and more refinements, reporting higher perceived usefulness. Self-assessed subject knowledge (Sections 5.2 and 5.4) did not significantly correlate with interaction frequency or usefulness ratings ($p > 0.3$), suggesting that AI literacy, rather than disciplinary expertise alone, may play a central role in effective engagement. This result is consistent with previous studies emphasizing that effective GenAI use depends on trust calibration, transparency, and structured human oversight. Prior work on student perceptions of ChatGPT reported that students may value AI support while still requiring human evaluation and guidance. Similarly, studies in programming and engineering education have shown that GenAI can support problem-solving but that its educational value depends on students' ability to formulate prompts, identify errors, and verify the correctness of outputs. In this sense, the present findings extend the literature by showing that AI-specific familiarity was more closely associated with interaction behavior and perceived usefulness than self-assessed disciplinary knowledge alone. The moderate positive correlation between perceived usefulness and interaction count in the combined data ($\rho = 0.429$, $p = 0.001$) indicates that iterative exploration is associated with higher perceived learning value. Consequently, familiarity might not only be a cognitive support

mechanism but also a motivational factor, thus leading to users being more reflective when engaging with AI tools.

6.3. Perception–Performance Alignment

Figure 10 illustrates the discrepancy between students' perceptions and teachers' evaluations regarding the required level of interaction with the GenAI tool across different cognitive depths. The figure includes only cases in which students engaged in at least one additional interaction with the tool, thereby confirming the absence of a correlation between the frequency of further interactions and the cognitive depth of a given task, as also observed in Figure 7.

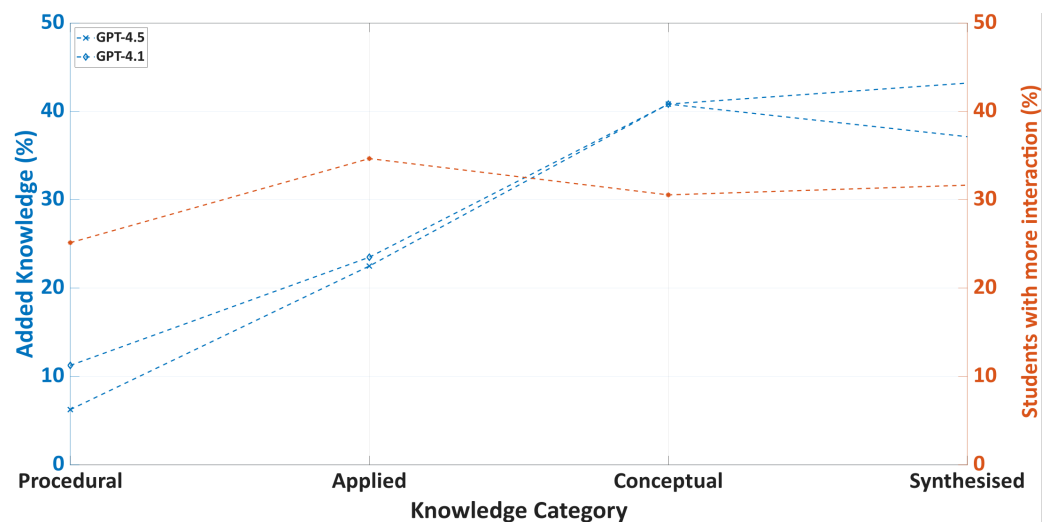


Figure 10. Comparison of student further interaction and added knowledge for each category.

In light of the positive correlation between students' subsequent interactions and perceived usefulness shown in Figure 6, the current results suggest that students did not report statistically distinguishable usefulness differences across cognitive-depth categories. However, this should not be interpreted as direct evidence that students accurately assessed model correctness at the task level. Rather, when considered together with the instructor-benchmarked reference evaluation, the pattern suggests a possible perception–reference miscalibration across cognitive-depth categories.

Importantly, these findings reflect an association between perceived usefulness and teacher-evaluated quality in this specific survey-based setting. They do not establish a causal link between students' trust and downstream academic performance. Rather, the observed miscalibration suggests that highly fluent outputs may be accepted without sufficient verification, which can increase the likelihood that errors remain unnoticed.

Students frequently rated the tool's outputs as highly useful even when teachers identified factual errors or incomplete reasoning. This misalignment aligns with prior findings by Tossell et al. (2024) and Hussain et al. (2025), who reported that students tend to exhibit a positive bias, often placing unwarranted trust in responses that appear coherent and well-articulated despite underlying inaccuracies.

Although these pedagogical implications were not directly tested in this study, bridging the perception–performance gap is critical for the educational integration of GenAI tools. Furthermore, requiring manual validation or incorporating peer review mechanisms may reduce overreliance and foster more critical engagement with AI-generated content.

6.4. Cognitive Depth and Implications on Education

Our analysis further highlights a divergence between student perception and instructor-benchmarked model evaluation. Teacher assessments varied significantly with task complexity (Section 5.8), whereas student-rated usefulness did not (Section 5.7). This suggests that students may not perceive increases in conceptual difficulty or recognize performance variability across depth categories.

From a pedagogical perspective, clearly defining and discussing the levels of reasoning required in each task could improve reflective understanding when using GenAI tools, although this will have to be tested in a future study.

Guided reflection questions, explicit evaluation criteria, or post-task discussions can help students identify which types of reasoning (conceptual, procedural, application, synthesis) benefit most from AI tools and which still demand human judgment. Because subject expertise did not predict interaction quality, instructional design could emphasize AI-specific competencies such as prompt formulation, verification, and debugging over mechanical content repetition. Such training aligns with emerging human-AI co-learning models in engineering education (Lang et al., 2025).

6.5. Implications for Assignment Design

The observed perception–performance gap suggests that assignments involving GenAI should not only ask students to obtain an answer but also require them to evaluate, justify, and validate the answer. Instructors can address this issue by designing tasks in which students must submit the original prompt, the GenAI output, the subsequent refinements, and a short technical justification of which parts of the output were accepted, corrected, or rejected. A second implication concerns the role of verification. For hardware-oriented courses, students can be required to validate AI-generated solutions through simulation results, test benches, synthesis reports, timing checks, or comparison with expected truth tables and functional specifications. This shifts the learning objective from passive acceptance of fluent AI output to active technical validation. A third implication concerns task scaffolding. Since students may not reliably perceive when task complexity increases, instructors can explicitly label tasks according to cognitive depth and discuss what kind of reasoning is required at each level. Conceptual and synthesized tasks may require additional reflection prompts, while procedural tasks may be suitable for controlled GenAI support followed by verification exercises. Finally, assessment design should reward critical engagement with GenAI rather than mere use of the tool. Rubrics may include criteria such as prompt quality, error identification, correctness of validation, and the ability to explain why an AI-generated solution is or is not technically acceptable. In this way, GenAI can be integrated as an object of critical evaluation rather than treated only as a solution generator.

7. Limitations and Conclusions

This study investigated the relationship between students' perceptions of generative AI effectiveness and an instructor-benchmarked reference evaluation of selected OpenAI models in digital systems design education. By integrating voluntary student surveys with teacher-side reference evaluations, a perception–performance gap was observed, where students frequently rated AI-generated outputs as useful even when the instructor-side benchmarking identified limitations in correctness or required substantial human scaffolding. The study is observational and based on voluntary survey data combined with instructor benchmarking; therefore, it can identify associations but cannot establish causal relationships between prior GenAI familiarity, interaction behavior, and perceived usefulness. This study should also be interpreted as an exploratory and practice-oriented framework rather than as a definitive benchmark of GenAI performance. The teacher-side

evaluation does not constitute an absolute ground truth but an instructor-benchmarked reference within the boundaries of the adopted protocol. This distinction is important because GenAI systems evolve rapidly, model outputs may vary across prompts and model versions, and students are already using these tools in their learning activities despite their known unreliability. Therefore, the value of the proposed framework lies in making student perceptions, interaction patterns, and instructor-side reference evaluations comparable and discussable in realistic educational settings. Within this perspective, lower reduced-complexity percentages should be interpreted as indicating lower dependence on instructor-provided scaffolding in the specific tasks analyzed, rather than as a universal measure of model quality. Similarly, on the student side, successful engagement should not be understood as the passive acquisition of a complete AI-generated solution. Rather, meaningful use of GenAI involves prompting, refinement, verification, and critical evaluation of the generated output. The observed perception–performance gap should therefore be read as a pedagogical warning: students may perceive GenAI outputs as useful even when instructor-side benchmarking indicates that additional expertise or validation is required.

Controlled follow-up studies are warranted to test causality, for example, by introducing a short AI-literacy training module and comparing trained versus untrained groups or by randomizing students to structured prompting and verification strategies while measuring objective learning outcomes in addition to perceptions. While the respondent pool was necessarily limited (32 at MIUN and 20 at UNISA), this study provides an initial multi-site benchmark of perception–performance alignment in an engineering setting. The observed effect directions and their replication across two cohorts motivate larger multi-site replications to increase statistical power, enable finer-grained subgroup analyses, and provide more precise estimates of associations. Future studies will also be able to pre-register primary hypotheses and incorporate additional outcome measures, strengthening the consistency across the two cohorts of the present findings. Moreover, because the study did not track objective learning outcomes, no conclusions can be drawn regarding the impact of perceived usefulness or trust on academic achievement; this should be examined in future controlled and longitudinal studies. Moreover, teacher benchmarking was performed using a structured protocol, but it may still be affected by rater-dependent judgment in borderline cases, tasks admitting multiple valid implementations, or alternative evaluation criteria for partial correctness. The instructor-side benchmark should therefore be interpreted as a structured expert reference internal to the present study design. It is not an external validation standard and does not provide evidence of inter-rater reliability. Although the use of a shared rubric and a reduced-complexity protocol improved procedural consistency, the absence of multiple independent raters limits the robustness of borderline judgments, especially for tasks admitting more than one technically valid implementation. Future work should include multiple raters per task and report inter-rater agreement to quantify the reliability of the reference evaluation. Moreover, because the two cohorts differ in academic level, course structure, language context, and assignment design, the institutional variable should be interpreted as a contextual covariate rather than as evidence of a specific institutional mechanism. On the other hand, GenAI outputs can also vary due to model stochasticity and prompt sensitivity. Quantifying output variability through repeated prompt runs and confidence intervals is a valuable extension, but it addresses a different question than the present survey-based perception–performance comparison; we therefore include repeated-run benchmarking as future work.

Across both institutions, prior familiarity with GenAI was significantly associated with interaction frequency and was positively associated with average perceived usefulness, whereas domain knowledge showed no significant influence. Because the two cohorts differ in course level and task design, this study does not aim to explain cross-context differences

but rather to examine whether the main associations replicate across two settings. These findings suggest that AI literacy may play an important role in supporting meaningful and critical engagement with GenAI tools, although this interpretation should be tested in larger and controlled studies. Moreover, these findings are specific to digital systems design; broader disciplinary generalization requires replication by domain experts. Model comparison further demonstrated that while advanced systems such as GPT-O3 and GPT-4.5 handle procedural and application-level tasks effectively, their performance declines in conceptual synthesis, revealing current limitations of large language models in higher-order reasoning. Accordingly, the pedagogical implications proposed in this manuscript should be interpreted as design-oriented guidelines rather than definitive prescriptions. Future studies should test these guidelines through controlled interventions, larger multi-site samples, and objective learning-outcome measures.

In addition, the teacher-side evaluation was based on one instructor per course. Although the same structured rubric was used in both contexts, future work should include multiple raters per task and inter-rater agreement analysis.

These findings extend prior literature on GenAI in higher education by showing that the educational value of these tools cannot be inferred from perceived usefulness alone. Consistent with previous studies on GenAI adoption, AI literacy, and trust calibration, our results suggest that students may value GenAI outputs even when substantial verification or instructor scaffolding is still required (Kasneci et al., 2023; Lo, 2023; Park, 2025; Steyvers et al., 2025).

At the same time, the findings add a domain-specific perspective to engineering and computing education research (Denny et al., 2024; Filippi & Motyl, 2024; Prather et al., 2023). In hardware-oriented digital systems design, the key pedagogical challenge is not only whether GenAI can generate plausible answers but also whether students can critically validate those answers against formal specifications, HDL constraints, simulation evidence, and implementation requirements. Therefore, the study supports the growing literature calling for assessment designs that incorporate prompt documentation, output verification, technical justification, and explicit reflection on the limits of AI-generated solutions (Bittle & El-Gayar, 2025; Denny et al., 2024; Yan et al., 2024).

Overall, these results highlight the need for structured educational strategies that help students calibrate trust, evaluate plausibility, and understand the boundaries of automated reasoning. Integrating guided reflection, validation exercises, and peer evaluation can support the development of critical thinking and reduce overreliance on GenAI outputs. Future research should include controlled or multi-institutional designs, standardized learning outcomes, and tracking of long-term effects on metacognition and calibration. Moreover, expanding the analysis to additional disciplines and AI models (including tool-enabled and multimodal systems) could better capture real-world use. Finally, interventions to improve students' ability to critically assess AI output should be developed and empirically tested.

Author Contributions: Conceptualization, I.S.; methodology, I.S.; software, V.G.; validation, I.S. and V.G.; formal analysis, I.S. and V.G.; investigation, I.S. and V.G.; resources, I.S., M.H. and V.G.; data curation, I.S. and V.G.; writing—original draft preparation, I.S. and V.G.; writing—review and editing, M.C., M.H., D.K. and S.J.M.; visualization, V.G.; supervision, I.S.; project administration, I.S.; funding acquisition, I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded within the HEaD project by Mid Sweden University.

Institutional Review Board Statement: The study was conducted using fully anonymous questionnaires, and no personally identifiable information was collected. The research involved no risk to participants. According to the policies of Mid Sweden University and the University of Salerno, studies based on fully anonymous survey data are exempt from Institutional Review Board approval.

Therefore, formal ethics approval was not required for this study. All participants were informed about the purpose of the study, and their participation was voluntary.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GenAI	Generative Artificial Intelligence
LLM	Large Language Model
VLE	Virtual Learning Environment
FPGA	Field Programmable Gate Array
HDL	Hardware Description Language
VHDL	VHSIC Hardware Description Language

References

- Allen, M., Naeem, U., & Gill, S. S. (2024). Q-module-bot: A generative AI-based question and answer bot for module teaching support. *IEEE Transactions on Education*, 67(5), 793–802. [\[CrossRef\]](#)
- Álvarez Ariza, J., Benitez Restrepo, M., & Hernández Hernández, C. (2025). Generative AI in engineering and computing education: A scoping review of empirical studies and educational practices. *IEEE Access*, 13, 30789–30810. [\[CrossRef\]](#)
- Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information*, 16(4), 296. [\[CrossRef\]](#)
- Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., Liu, T., & Han, B. (2024). Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37, 96640–96670.
- Corbin, T., Bearman, M., Boud, D., & Dawson, P. (2025). The wicked problem of AI and assessment. *Assessment & Evaluation in Higher Education*, 50(8), 1234–1245. [\[CrossRef\]](#)
- Cui, Y. L., Zeng, M. L., Du, X. K., & He, W. M. (2025). What shapes learners' trust in AI? A meta-analytic review of its antecedents and consequences. *IEEE Access*, 13, 164008–164025. [\[CrossRef\]](#)
- Denny, P., Leinonen, J., Prather, J., Luxton-Reilly, A., Amarouche, T., Becker, B. A., & Reeves, B. N. (2024). Prompt problems: A new programming exercise for the generative AI era. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1* (pp. 296–302). Association for Computing Machinery. [\[CrossRef\]](#)
- Filippi, S., & Motyl, B. (2024). Large language models (LLMs) in engineering education: A systematic review and suggestions for practical adoption. *Information*, 15(6), 345. [\[CrossRef\]](#)
- Gong, L., Chen, J., & Wu, F. (2025). Is ChatGPT a competent teacher? Systematic evaluation of large language models on the competency model. *IEEE Transactions on Learning Technologies*, 18, 530–541. [\[CrossRef\]](#)
- Guedes, P., Abranches Silva Lopes, E., Ribeiro, P. F., & Zambroni de Souza, A. C. (2025). The impact of artificial intelligence on learning and teaching of engineering. *IEEE Transactions on Education*, 68(5), 417–425. [\[CrossRef\]](#)
- Huber, S. E., Kiili, K., Nebel, S., Ryan, R. M., Sailer, M., & Ninaus, M. (2024). Leveraging the potential of large language models in education through playful and game-based learning. *Educational Psychology Review*, 36(1), 25. [\[CrossRef\]](#)
- Hussain, M., Pietrosanto, A., Liguori, C., Paciello, V., & De Santo, M. (2025, January 7–10). *ChatGPT in the engineering classroom: A pre-study on students' perceptions and experience*. Hawaii International Conference on System Sciences 2025 (HICSS-58), Waikoloa Village, HI, USA. [\[CrossRef\]](#)
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. [\[CrossRef\]](#)
- Kirova, V. D., Ku, C. S., Laracy, J. R., & Marlowe, T. J. (2024). Software engineering education must adapt and evolve for an LLM environment. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1* (pp. 666–672). ACM. [\[CrossRef\]](#)
- Kong, S.-C., & Yang, Y. (2024). A human-centered learning and teaching framework using generative artificial intelligence for self-regulated learning development through domain knowledge learning in K–12 settings. *IEEE Transactions on Learning Technologies*, 17, 1562–1573. [\[CrossRef\]](#)

- Lang, Q., Wang, M., Yin, M., Liang, S., & Song, W. (2025). Transforming education with generative AI (GAI): Key insights and future prospects. *IEEE Transactions on Learning Technologies*, 18, 230–242. [\[CrossRef\]](#)
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. [\[CrossRef\]](#)
- Luo, X., Rechartd, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., Marinazzo, D., Nicholas, J., Salatiello, A., Sucholutsky, I., Minervini, P., Razavi, S., Rocca, R., Yusifov, E., Okalova, T., ... Love, B. C. (2025). Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 9(2), 305–315. [\[CrossRef\]](#)
- Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, 32(3), 581–604. [\[CrossRef\]](#)
- Nguyen, S., Babe, H. M., Zi, Y., Guha, A., Anderson, C. J., & Feldman, M. Q. (2024). How beginning programmers and code LLMs (mis)read each other. In *Proceedings of the CHI conference on human factors in computing systems* (pp. 1–26). ACM. [\[CrossRef\]](#)
- Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6667), eadh2586. [\[CrossRef\]](#)
- Oh, S. (2025). Evaluating mathematical problem-solving abilities of generative AI models: Performance analysis of o1-preview and GPT-4o using the Korean college scholastic ability test. *IEEE Access*, 13, 1227–1235. [\[CrossRef\]](#)
- Park, J. (2025). A systematic literature review of generative artificial intelligence (GenAI) literacy in schools. *Computers and Education: Artificial Intelligence*, 9, 100487. [\[CrossRef\]](#)
- Prather, J., Denny, P., Leinonen, J., Becker, B. A., Albluwi, I., Craig, M., Keuning, H., Kiesler, N., Kohn, T., Luxton-Reilly, A., MacNeil, S., Petersen, A., Pettit, R., Reeves, B. N., & Savelka, J. (2023). The robots are here: Navigating the generative AI revolution in computing education. In *Proceedings of the 2023 working group reports on innovation and technology in computer science education* (pp. 108–159). Association for Computing Machinery. [\[CrossRef\]](#)
- Qadir, J. (2023). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE global engineering education conference (EDUCON)* (pp. 1–9). IEEE. [\[CrossRef\]](#)
- Romera-Paredes, B., Barekatin, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J., Ellenberg, J. S., Wang, P., Fawzi, O., Kohli, P., & Fawzi, A. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625, 468–475. [\[CrossRef\]](#)
- Savelka, J., Agarwal, A., Bogart, C., Song, Y., & Sakr, M. (2023). Can generative pre-trained transformers (GPT) pass assessments in higher education programming courses? In *Proceedings of the 2023 conference on innovation and technology in computer science education V. 1* (pp. 117–123). ACM. [\[CrossRef\]](#)
- Shailendra, S., Kadel, R., & Sharma, A. (2024). Framework for adoption of generative artificial intelligence (GenAI) in education. *IEEE Transactions on Education*, 67(5), 777–785. [\[CrossRef\]](#)
- Shallari, I., & Hussain, M. (2024, January 3–6). *Assignments in the ChatGPT-era: Case study on PLAGIARISM in digital systems design courses*. Hawaii International Conference on System Sciences 2024 (HICSS-57), Honolulu, HI, USA.
- Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 7, 221–231. [\[CrossRef\]](#)
- Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & De Visser, E. J. (2024). Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Learning Technologies*, 17, 1069–1081. [\[CrossRef\]](#)
- Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90–112. [\[CrossRef\]](#)
- Zhao, X., Chen, X., Huang, V., Rollins, M., Carratù, M., & Shallari, I. (2025, January 7–10). *Students' use and attitudes toward generative artificial intelligence: A comparative study between the UK and China*. Hawaii International Conference on System Sciences 2025 (HICSS-58), Waikoloa, HI, USA [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.