Computers and Education: Artificial Intelligence 5 (2023) 100172

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances

Margherita Bernabei [a], Silvia Colabianchi [b], Andrea Falegnami [c], Francesco Costantino [b,*]

[a] Department of Mechanical and Aerospace Engineering, Sapienza University of Rome, Via Eudossiana 18, 00184, Rome, Italy
[b] Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome, Via Ariosto 25, 00185, Rome, Italy
[c] Management Engineering Faculty, Uninettuno University, Corso Vittorio Emanuele II 39 00186, Rome, Italy

ARTICLE INFO

Keywords:
LLM
ChatGPT
Higher education
Essay generation

ABSTRACT

The accessibility of advanced Artificial Intelligence-based tools, like ChatGPT, has made Large Language Models (LLMs) readily available to students. These LLMs can generate original written content to assist students in their academic assessments. With the rapid adoption of LLMs, exemplified by the popularity of OpenAI's ChatGPT, there is a growing need to explore their application in education. Few studies examine students' use of LLMs as learning tools. This paper focuses on the application of ChatGPT in engineering higher education through an in-depth case study. It investigates whether engineering students can generate high-quality university essays with LLMs assistance, whether existing LLMs identification systems can detect essays produced with LLMs, and how students perceive the usefulness and acceptance of LLMs in learning. The research adopts a deductive/inductive approach, combining conceptualization and empirical evidence analysis. The study involves mechanical and management engineering students, who compose essays using LLMs. The essay assessment showed good results, but some recommendations emerged for teachers and students. Thirteen LLMs detectors were tested without achieving satisfactory results, suggesting to avoid LLMs ban. In addition, students were administered a questionnaire based on constructs and items that follow the technology acceptance models available in the literature. The results contribute to qualitative evidence by highlighting possible future research and educational practices.

1. Introduction

The application of LLMs in education is a new research topic, especially if we consider the most recent LLMs with outstanding emerging capabilities (Wei et al., 2022). An extensive collection of contributions on the use of ChatGPT in education helps define the opportunities, challenges, and implications related to using LLMs in the context of education (Ji et al., 2022; Kasneci et al., 2023). Some authors measured the capability of LLMs to pass specific exams, but mostly to measure the LLMs' power to mimic human intelligence: according to OpenAI GPT-4 passes LSAT, GRE, the Bar Exam, US Medical Licencing Examination, and other exams (Gilson et al., 2023; Katz et al., 2023; OpenAI, 2023). Few studies are available on the use of LLMs by students as a tool to ease the learning process. LLMs can provide effective translations, summarizations of complex text, and LLMs can generate text by replacing the work of reading, synthesis, integration of sources, and paraphrasing of texts.

While some schools and universities have banned the use of LLMs (Johnson, 2023), these tools can be seen as an opportunity to rethink some traditional learning processes. Examples that may be similar are the following: the calculator is allowed to facilitate work in complex mathematical problems, no longer placing importance on the ability to perform calculations on paper; and CAD software is used to design more quickly, without requiring students to draw by hand.

Several teachers already use LLMs in education to create syllabi (Cribben & Zeinali, 2023), produce exercises (Sarsa et al., 2022), provide basic educational materials (Dwivedi et al., 2023), and more (Lesage et al., 2023). Jeon and Lee (2023) examined the perspectives and experiences of teachers who deployed ChatGPT in their instruction. In a certain way, it is fair to make available to students the same tools used by teachers, in this case, the LLMs.

Few experiments are available in the field of engineering education.

* Corresponding author. Via Ariosto 25, 00185, Rome, Italy.
E-mail address: francesco.costantino@uniroma1.it (F. Costantino).

https://doi.org/10.1016/j.caeai.2023.100172
Received 6 July 2023; Received in revised form 5 October 2023; Accepted 6 October 2023
Available online 10 October 2023

Some authors have tried the performance of OpenAI tools to produce written text from laboratory notes taken in a senior fluid course in mechanical engineering, without involving real students, just simulating the application of the LLM, and assessing the results (Lesage et al., 2023). Other authors have introduced ChatGPT to aid students in problem-solving processes (Tsai et al., 2023). The students' perception of LLM is a field of research, with applications in programming trainin (Yilmaz & Karaoglan Yilmaz, 2023), and embedded systems design (Shoufan, 2023). The first work suggested overcoming the assessment of the efficacy of students in simple programming tasks; we focused on complex assignments. The second work applied ChatGPT as a support tool to gather knowledge and respond to quizzes, without considering the generation of text by the students. The interest in the use of LLMs by students is evidently current.

In this paper, we study the application of LLMs in the education process of engineers using an in-depth case study in which students used an LLM to produce a university essay, as a step in the expected learning process. The purpose of this paper is to explore the research area of education through students' use of LLM (in particular ChatGPT) in university engineering education.

The research questions of this study are as follows:

- RQ1. Can engineering students produce good university essays with the help of LLMs?
- RQ2. Can available LLM identification systems identify university essays produced using LLMs?
- RQ3. How do students perceive the use of LLMs in terms of learning usefulness and adoption acceptance?

The research approach was deductive/inductive: the deduction started from the conceptualization of how to observe the perception of LLM technologies in the study activities of engineering students; the induction process studied the observations of evidence in the case studies to highlight notable information to be shared with educators and researchers in the field of education.

The contribution of this paper is to provide evidence from a qualitative research approach, based on an in-depth case study of a master's class in engineering studies. Indeed, the literature frequently cites the case study strategy as one of the most prevalent qualitative research strategies (Glette & Wiig, 2022; Lavarda & Bellucci, 2022; Rashid et al., 2019).

This work is especially pertinent to educators in the engineering field seeking to explore the application of LLMs by students, and it is noteworthy that there is a notable scarcity of existing research with similar analyses. It is particularly relevant as it delves into the increasingly prevalent integration of Large Language Models (LLMs) in educational settings, illuminating their effectiveness in aiding engineering students in essay compositions and assessing their detectability. Additionally, it investigated the pivotal aspect of student perceptions regarding LLM utilization. This study contributes by offering empirical evidence regarding the practical utility of LLMs in higher education and the efficiency of current identification systems, thereby advancing our comprehension of the potential advantages and challenges linked to their adoption in academic environments.

The study involved students from the mechanical engineering and management engineering programs of the Smart Factory master's course at the University of Rome "Sapienza". The teacher provided a list of topics and asked the students to compose a personal essay on a topic of their choice using a Large Language Model. A questionnaire was administered to students before and after the use of the LLM to analyze constructs measuring new technology acceptance. The paper presents the details of these steps and the experiment.

Section 2 describes the experimental design to develop the case study and analyze it. Specifically, it outlines the process to construct the questionnaire, conduct the experiment, and analyze the results. Section 3 identifies recommendations for students and teachers, along with future research.

2. Methods

After defining the research objective and questions, the study was conducted in five steps, as shown in Fig. 1. Step 1 focused on defining the sample and its characteristics. Step 2 involved a comprehensive review of technology acceptance models and dimensions to assess university students' perceptions. The structure and items of the questionnaire were developed based on theoretical foundations and group discussion. Step 3 was dedicated to conducting the experiment, including administering the questionnaire, carrying out, and evaluating the assignment. In Step 4, the questionnaire was administered for the second time to evaluate students' perception post-LLMs utilization, and essays produced as assignments were examined to identify LLM-related content. In Step 5, an analysis and discussion of the results took place. To enhance understanding of the paper, this paragraph describes step-by-step the research process following the blocks depicted in Fig. 1.

3. Step-by-step description

3.1. Sample setup

In the first phase, a representative sample of university students was prepared to ensure valid results. A quality sample is crucial for accurate results, and homogeneity in terms of preparation and knowledge was achieved.

3.1.1. Sample selection

31 students (61.3% male, 38.7% female) are recruited from the master's degree course in Management Engineering (67.7%) and Mechanical Engineering (32.3%) at the University of Rome La Sapienza. All participants have a bachelor's degree in engineering, which confirms the shared basic scientific knowledge of all participants. 51.6% of the students had taken at least one course in computer science fundamentals and programming. The experiment is conducted within the Smart Factory course, a master's degree program focused on the application of Artificial Intelligence (AI) in the industrial context. All participants are students enrolled in the Smart Factory course and are invited by their instructors to voluntarily take part in the experiment.

3.1.2. Delivery of the assignment rules

Students must complete an assignment and deliver a presentation as part of the course requirements. The assignment involves the submission of a written essay (10 pages in length) with an in-depth study of an innovative Industry 4.0 technology and a case study of a real-life application. Afterward, students are invited to present their work to the class. Students receive an evaluation for the written assignment and an evaluation for the oral presentation. For the development of the written assignment, the teachers presents an evaluation grid and allows the use of ChatGPT (OpenAI, 2022). ChatGPT is an advanced chatbot capable of generating human-like text, answering questions, and assisting in tasks such as summarization or grammar checks through human-like completion. ChatGPT is built upon a Generative Pre-Trained Transformer large language model developed by OpenAI. The instrument is free to use.

3.1.3. Sample preparation

In this phase, the professor ensures that all participants have a similar level of knowledge about the ChatGPT tool. Each student undergoes training on how to use ChatGPT and its underlying machine learning models to gain awareness of the tool. The training covers the workings of LLMs, accessing ChatGPT, and its functionalities. Moreover, students receive guidance on formulating questions and requests, emphasizing the importance of clear context and precise keywords. They are also encouraged to follow up, rework, and verify ChatGPT's

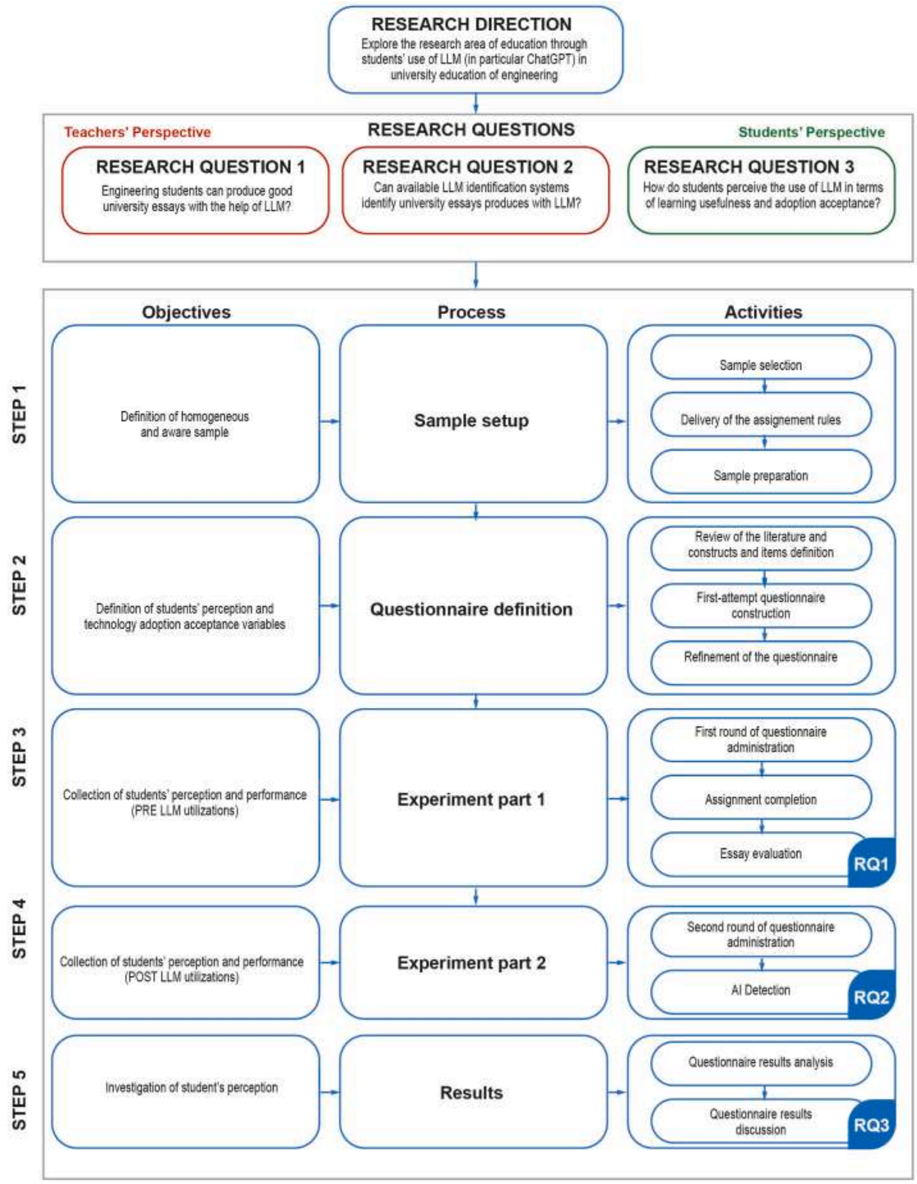

Fig. 1. Research method.

responses. The professor allows students to freely utilize ChatGPT to enhance their understanding of topics, rephrase scientific articles, or generate parts of their writing. In the case study, all students decided to use the tool. Finally, students are trained on how to approach the administration of questions.

3.2. Questionnaire definition

Constructs and items are identified through a literature analysis, focusing on the recurring constructs found in studies examining the implementation of innovative technologies. This analysis started from the well-known Technology Acceptance Model (TAM) (Davis, 1989), going through the UTAUT (Y. S. Wang & Shih, 2009), to further models described in the following paragraph about constructs definition. To define the items of each construct are investigated Benefits, Opportunities, Costs, and Risks (BOCR) related to the implementation of innovative LLM technologies in education. BOCR investigation is an established methodology for analyzing factors related to technology implementation and their application potential (Lee et al., 2011; Osmani et al., 2021; Tabatabaee et al., 2019; Wijnmalen, 2007; Zakeri et al., 2023). The identified benefits, opportunities, costs, and risks pertain to the identified constructs. Based on these results, we define the questionnaire items, as shown in the following paragraphs.

3.2.1. Constructs definition

An established theory to investigate the implementation of innovative technology tools is the Technology Acceptance Model (TAM) (Davis, 1989). TAM models how users accept and use technology. This model was upgraded by Vendkatesh's model, which integrates the fragmented theory and research on individual acceptance of information technology into a unified theoretical model, to capture the essential elements of the previously established models (Venkatesh et al., 2003). Researchers are increasingly testing, modifying, and revisiting this model in many field and for different purposes (Afonso et al., 2012; Ayaz & Yanartaş, 2020; Kabanda & Brown, 2017; Mosweu et al., 2017; Sezer & Yilmaz, 2019; Y. S. Wang & Shih, 2009). The analysis of such models reveals that behavioral intention leads people to use technology. Behavioral intention can be influenced by many internal and external factors. Firstly, by the attitude, i.e., the impression users have about the technology. Attitude is focal to understanding people's perceptions toward technology, whether there is a positive or negative inclination and what feelings and preconceptions exist (Okonkwo & Ade-Ibijola, 2021). The acceptance of technology relies on both the benefits it provides to users and their level of trust while using the system (Wu et al., 2011), and on the social influence that the context generates on the user (Ayaz & Yanartaş, 2020; Venkatesh et al., 2003). The user judges the fairness of the technology, based on the correlation between input and expected output, as well as the provision of accurate, unbiased, correctable, and representative information in accordance with ethical or moral standards (Tuyet Mai et al., 2013). As such, ethic directly affects performance expectancy, effort expectancy, and social influence (Ayaz & Yanartaş, 2020; Aziz et al., 2021). Besides ethical and moral perception, the choice to use technology is also driven by perceived usefulness and perceived performance expectancy (Ayaz & Yanartaş, 2020; Davis, 1989; Dizon et al., 2022; Keržič et al., 2019), which significantly impact user satisfaction. As such, users assess the effort of using a technology by comparing it to performing the task manually and consider the ease of use (Davis, 1989; Dizon et al., 2022; Keržič et al., 2019).

This analysis leads to the identification of the following six constructs: (1) Attitude; (2) Trust; (3) Social Influence; (4) Fairness & Ethics; (5) Usefulness & Performance Expectancy; (6) Effort & Ease of Use. Considering the RQ3 *"How do students perceive the use of LLM in terms of learning usefulness and adoption acceptance*," the Usefulness & Performance Expectancy construct is expanded with additional items in the second round of questionnaire administration, post-LLM utilization.

3.2.2. Items definition

The definition of the items required detailing the 6 identified constructs. Each construct was the subject of a research team discussion. The discussion started from the literature cited above to define specific items, and then added some more considering elements of benefit, opportunity, cost, and risk. More in detail, the literature highlights the Benefits, Opportunities, Costs, and Risks (BOCR) related to the use of LLM technologies in education. These were investigated to define the items within each construct. Benefits are related to the integration of contents, quick access to information, motivation and engagement associated with the new learning modes. Namely, the innovative learning environment can generate intrinsic motivation (Yin et al., 2021; Colabianchi, Bernabei, & Costantino, 2022). Furthermore, benefits related to the opportunity to allow multiple users immediate assistance (Okonkwo & Ade-Ibijola, 2021)**,** speeding up the response time, availability 24 h a day of information, and the possibility to reduce stress and increase the willingness to learn (Mageira et al., 2022)**.** Opportunities arise from the possibility of LLM technologies to generate hybrid-learning/teaching approaches (J. Wang et al., 2021). Therein, new pedagogical approaches are referenced, to modernize the education system and transform education via innovation (Mhlanga, 2023), after defining specific areas of education that could benefit from digitalization (Okonkwo & Ade-Ibijola, 2021). The main costs to mitigate potential negative effects are related to the design of appropriate user-interfaces (Kasneci et al., 2023) and to the effort required to be aware of the technologies' limitations and capabilities (Mageira et al., 2022). Further costs arise from the continuous and in-depth verification of the information accuracy and integrity (Mhlanga, 2023) and to the efforts to maintain these up-to-date (Kasneci et al., 2023). Risks are mostly in ethical and equity. Recurrent is the dimension related to ethical issue, which highlights risks of data privacy and security, dissemination of false or misleading information, bias, fairness and discrimination (Kasneci et al., 2023; Mhlanga, 2023; Okonkwo & Ade-Ibijola, 2021). Then, risks associated with the difficulty to distinguish model-generated from student– generated answers (Kasneci et al., 2023), and in the generation of inappropriate or misunderstanding knowledge (Mageira et al., 2022). Moreover, LLM technologies risk undermining creativity, critical thinking, and problem solving skill, traits exclusive to human beings (Kasneci et al., 2023).

Based on these activities, several items were defined and were input of the first attempt questionnaire construction.

3.2.3. First attempt questionnaire construction

In this step, the research team defines the questionnaire to evaluate students' perceptions. Items are assigned to specific constructs, selected, and adapted based on theory and questionnaires from the literature, as well as team evaluations to best reflect the university context and students' learning environment. Table 1 provides a detailed list of all constructs and items. The questionnaire uses a 4-point Likert scale (1 = strongly disagree; 4 = strongly agree) (Warmbrod, 2014). Additionally, it includes items regarding participating students' demographic information, such as age, gender, and bachelor and master details.

3.2.4. Refinement of the questionnaire

Once the constructs and items are defined, the team consisting of two professors and two researchers discuss each item. The questions are ordered logically. Subsequently, the team conduct a refinement process to eliminate jargon terminology, smooth out language differences minimize the chance of misunderstanding, and make the assessment tool as homogeneous as possible. In addition, various biases are minimized related to the survey process (acquiescence, authority, politeness, herd behavior, etc.) (Choi et al., 2005). Finally, the team ensures objectivity and neutral tone for each question, avoiding leading phrasing and removing absolutes. The final version of the questionnaire is in appendix.

3.3. Experiment part 1

3.3.1. First round of the questionnaire administration

The first round of the questionnaire is administered to the students before the start of the assignment and, therefore, before they use ChatGPT. The students are invited to complete the questionnaire during class hours. All 31 students enrolled in the course, who have completed the preliminary training, complete the questionnaire.

3.3.2. Assignment completion

After the first administration of the questionnaire, the students select the technology to be explored in the assignment. The students have 10 days to complete and submit their assignments. All 31 students submitted the assignment (a written essay). All students presented their work to the class through an oral presentation.

3.3.3. Assignment evaluation

The course professor and the two researchers discuss the evaluation of each assignment. To assign an overall value to the assignment, the essay and oral presentations are evaluated (see paper appendix). A qualitative grid is built for the evaluation. The grid takes into account aspects such as quality and completeness of the content, quality of bibliographic references, level of discussion and reworking, critical analysis skills, and proposed SWOT analysis. Each assignment is given a grade from 18 to 30. A grade lower than 18 results in a non-passed assignment. The essay evaluation and the oral presentation evaluation contribute equally to the final mark of the assignment. In the essay evaluation, of the 31 submitted, 30 out of 31 passed. All the students passed the oral presentation. Fig. 2 shows all the grades reported in the two sections and the final grade for each assignment. Fig. 3, on the other hand, shows the grade distribution of the assignments. It is interesting to note that overall the students received good evaluations. The evaluations covered a wide range of grades as in previous years or other exams. This result shows how the use of LLM has not altered students' performance but only changed the way they approach the assignment. This observation is partly confirmed by the graph in Fig. 4. Course teachers evaluated the essays using a qualitative evaluation grid and for each essay evaluated the presence or absence of certain elements. The frequency analysis of these elements shows a general tendency to follow the assignment rules and to provide complete work in the required parts.

Table 1
LLM-DT analysis.

2. LLM-DTs employ NLP techniques and sometimes LLM models themselves to generate a probability of text being written by an LLM. This incurs a computational cost proportional to the number of words. The number of words is often constrained both as a minimum value to ensure sufficient accuracy and as a maximum value to limit the computational cost. Therefore, the LLM-DTs are evaluated to highlight these limitations expressed in terms of the number of words, sometimes in terms of the number of characters or tokens. In the context of LLM, tokens represent the fundamental units of text or code that an LLM utilizes for language processing and generation. Tokens can be characters, words, subwords, or other segments of text or code, depending on the chosen tokenization method or scheme. OpenAI suggests, as a rough rule of thumb, that 1 token is approximately equivalent to 4 characters or 0.75 words for English text.

Id	Detector name	Detector weblink	Limitations	Type of answer	API availability
1	A.I. Text Classifier	https://platform.openai.com/ai-text-classifier	Min 1000 characters	Narrative result (e.g. "Your text is unlikely to be written by ChatGPT")	No
2	Content @ Scale	https://contentatscale.ai/ai-content-detector/	Max 25000	Score: probability for ai/human; Predictability: percentage, Probability: percentage, Pattern: percentage.	Yes
3	Copyleaks	https://copyleaks.com/ai-content-detector	Not declared	Score: probability for ai/human	Yes
4	Crossplag	https://crossplag.com/	Max 3000 words	Score: probability for ai/human and a narrative result (e.g. "This text is mainly written by an AI.")	No
5	GLTR	http://gltr.io/	Not declared	Score: count of unexpected words	No
6	GPT-2 Output Detector	https://openai-openai-detector.hf.space/	Max 510 tokens	Score: percentage of fake/real	No
7	GPTKit	https://gptkit.ai/?ref=theresanaiforthat	Max 2048 characters	Score: percentage of fake/real	Yes
8	GPTZero	https://gptzero.me/	5000 char	Narrative result ("Your text is likely to be written entirely by a human")	Yes
9	Hive Moderation	https://hivemoderation.com/ai-generated-content-detection	Max 8192 characters	Score: probability for ai/human	No
10	Originality.AI	https://originality.ai/free-ai-content-detector-chrome-extension/	Not declared	Score: probability for ai/human	Yes

Table 1 (*continued*)

Id	Detector name	Detector weblink	Limitations	Type of answer	API availability
11	Sapling.ai	https://sapling.ai/ai-content-detector	Max 2000 characters	Score: probability for ai/human	Yes
12	Smodin - AI Detection	https://smodin.io/ai-content-detector	Max 50000 characters	Score: probability for ai/human	Not specific for detection
13	Writer - AI content detector	https://writer.com/ai-content-detector/	Max 1500 characters	Score: percentage of fake/real	Yes
14	ZeroGPT	https://www.zerogpt.com/	Not declared	Score: probability for ai/human	Yes

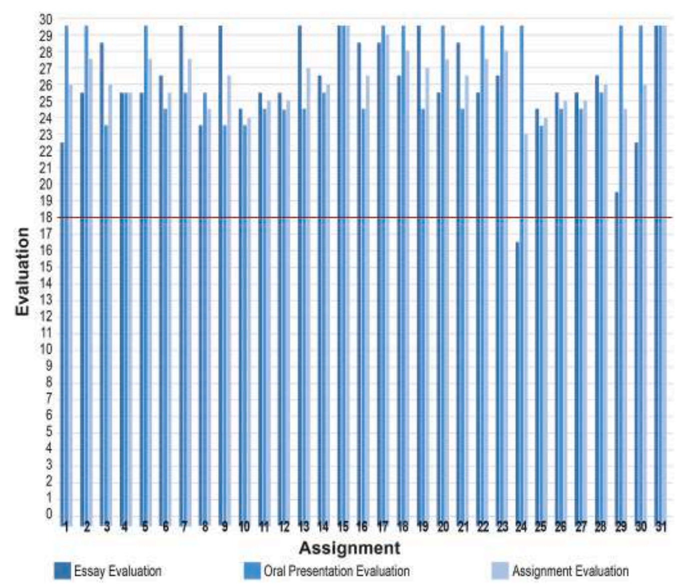

Fig. 2. Assignment evaluations.

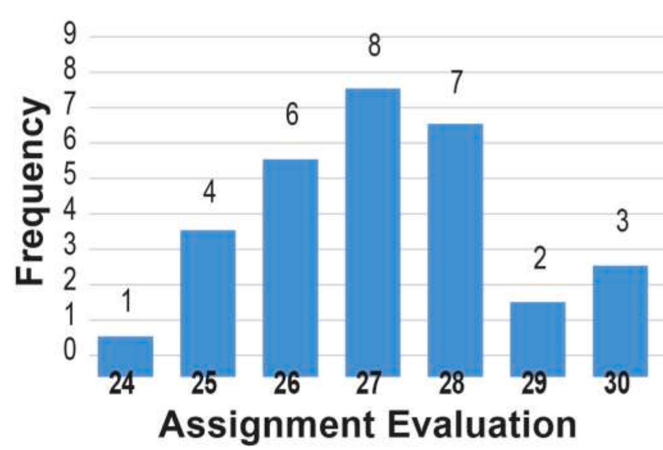

Fig. 3. Distribution of assignment evaluations.

However, when the professor asked the students to provide their analysis and discussion the results were worse. This is the case, for instance, with the Strength, Weakness, Opportunity, Threat (SWOT) analysis that required the students a deep analysis of the motivations behind the

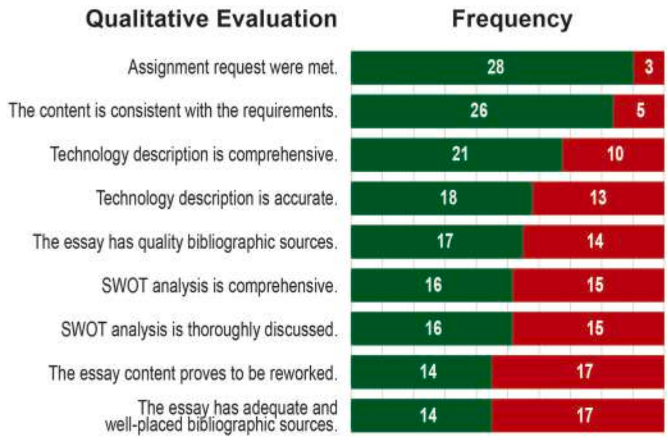

Fig. 4. Qualitative evaluation considering specific criteria.

suggestion from ChatGPT. Finally, the lowest value is reported on text reworking. ChatGPT tends to provide schematic answers and frequently employs bulleted lists. Also to different questions, but on the same broader topic, it tends to repeat itself. Several students relied on these answers without reworking them making parts of the essays of little merit.

3.4. Experiment part 2

3.4.1. Second round of the questionnaire administration

Once the students received the evaluation, they were invited to resubmit the questionnaire to evaluate their perception before and after the use of ChatGPT. Students were invited to complete the questionnaire during class hours. 28 out of 31 students completed the second round of the questionnaire.

3.4.2. AI detection

As mentioned before, an approach that some institutions are adopting is to consider the use of LLMs as a form of cheating that allows students to bypass certain stages of the learning process. Consequently, certain educational institutions explicitly prohibit the use of LLMs. This perspective regards LLMs not as study aids, but as study shortcuts, and it necessitates the ability to differentiate between texts written with and without the use of LLMs. Hence, we sought to address RQ2 "Can available LLM identification systems identify university essays produced with LLM?".

It is important to note that students rarely submit the text produced by an LLM as-is. Instead, they review and rework the text as it is tied to an evaluation. Therefore, the educator does not need to determine whether a text has been written by an LLM, but rather how much of it has been written by an LLM.

The research proceeded through the following steps:

1. Identification of commercial and non-commercial tools for written text identification by LLM (LLM detection tools - LLM-DT). This phase aimed to create a ready-to-use list of LLM-DTs.
2. Evaluation of LLM-DT tools in terms of limitations on the number of characters/words/tokens, type of response provided, and availability of API access. This phase aims to understand how to perform checks on the texts.
3. 1st run test (basic-test) for the selection of LLM-DTs. This phase involves selecting LLM-DTs to be used on the entire set of available documents and analyzing metrics for result comparison.
4. 2nd run test (full-test) for evaluating the effectiveness of LLM-DTs.

The following section will delve into a comprehensive explanation of these four elements.

1. The identification is conducted through a simple Google search, Bing search, GitHub search, and GitLab using research queries that included the terms "llm," "GPT," "ChatGPT," and "detect*" (detection, detector, etc.) OR "classif*" (classifier, classification, etc.). The search focuses on ready-to-use tools and, therefore, it is not carried out on scientific databases where researchers are studying new detection techniques with algorithms that are not yet ready for use. This phase provided the Table 1.

Each LLM-DT presents its own metrics, often with a measure of the probability percentage that the text is written by an AI or as a percentage of text likely to be authored by AI. However, many systems provide qualitative phrases such as "Your text is likely to be written entirely by a human." This variability in measures and vagueness in their definition leads to limited comparability among LLM-DTs (Tang et al., 2023). Other factors, such as costs and additional functionalities, have also been identified, such as the ability to perform checks through API usage rather than manual inspection. This analysis has resulted in the definition of certain rules for conducting AI detection in the text:

- If the text length is below the required threshold, no evaluation is required.
- If the text length exceeds the allowed limit, it is truncated into coherent subtexts based on sentence endings. The scores of the different subtexts are then aggregated into a text score calculated as the average of the subtext scores.
- If the evaluation is qualitative in nature, the LLM-DT has not been used for phase 2 unless APIs were available to enable straightforward usage.

3. The 1st run test is conducted to determine which LLM-DTs to exclude from the extensive analysis. The extensive analysis is highly time-consuming, requiring preprocessing of excessively long documents and, in many cases, significant manual effort to input texts in the LLM-DT dedicated websites, and retrieve scores from these, where useful APIs for process automation were not available. APIs availability (see Table 1) permit the authors to use a Python code to send automatically texts to LLM-DT and receive the AI detection score. The authors did not employ automatic web scraping techniques. To avoid lengthy yet futile activities, the LLM-DTs are initially evaluated on 5 essays. The essays are selected from those in which the authors informally declared extensive use of LLM. Since each essay consists of 5 sections, a total of 25 processed texts are used. The Detector05 has been excluded from the analysis due to its metrics being fundamentally different from those of the other detectors. The results of the 1st run test are presented in Table 2. Some LLM-DTs were excluded for the following reasons:

- LLM-DTs unable to recognize the contribution of LLM.
- LLM-DTs with non-comparable qualitative measures.

The 1st run also serves to define how to compare the measures, whose meanings differ. The comparison is not intended to be mathematically accurate but merely aims to determine whether LLM-DTs tend to overestimate or underestimate the use of LLM. LLM-DTs that provide the percentage of text produced by LLM and the probability that the text was generated by LLM were considered comparable parameters and reported on a 0–100 scale. Once again, it is emphasized that the comparison is not mathematically accurate but serves the purpose of determining whether LLM-DTs are more or less "biased," i.e., oriented towards identifying or denying the presence of AI. The accuracy of the tools was not measured because the students did not provide information about the quantity of AI text, but all declared LLM usage. The possibility of algorithmic biases may occur and were not controlled; the motivation is that the paper mimicked a generic teacher who wants to check whether LLM is used in writing an essay assigned to students. The teacher has no information on the algorithms used by LLM-DTs. These algorithms are secrets protected by the companies offering the detection

Table 2
LLM-DT analysis – 1st run results.

Essay	Essay01					Essay02					Essay03				
Detector	Section1	Section2	Section3	Section4	Section5	Section1	Section2	Section3	Section4	Section5	Section1	Section2	Section3	Section4	Section5
Detector01	0,50	0,50	0,00	0,00	0,50	0,50	0,00	0,00	0,00	0,50	0,00	0,00	0,50	0,00	0,00
Detector02	0,78	1,00	0,80	0,97	0,99	1,00	1,00	1,00	1,00	1,00	0,99	1,00	1,00	0,98	0,98
Detector03	0,88	0,58	0,79	0,77	0,45	0,00	0,04	0,00	0,00	0,00	0,22	0,20	0,00	0,04	0,04
Detector04	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,01
Detector06	0,00	0,00	0,16	0,52	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,07	0,07
Detector07	0,17	0,53	0,11	0,97	0,43	0,76	0,44	0,11	0,23	0,16	0,79	0,00	0,41	0,58	0,06
Detector08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Detector09	1,00	0,02	0,02	0,07	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Detector10	1,00	0,01	0,56	0,91	0,34	0,06	0,01	0,03	0,01	0,03	0,37	0,02	0,02	0,58	0,58
Detector11	0,71	0,20	0,80	1,00	0,42	0,01	0,00	0,08	0,00	0,01	0,82	0,00	0,00	0,00	0,00
Detector12	0,22	0,00	0,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,05	0,16
Detector13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Detector14	0,08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Essay	Essay05					Essay22				
Detector	Section1	Section2	Section3	Section4	Section5	Section1	Section2	Section3	Section4	Section5
Detector01	0,50	0,50	0,00	0,00	0,50	0,50	0,00	0,00	0,00	0,50
Detector02	0,36	0,41	0,25	0,38	0,86	0,38	0,88	0,69	0,47	0,89
Detector03	0,79	0,71	0,92	0,79	0,93	0,90	0,80	0,80	0,82	0,95
Detector04	0,83	1,00	0,75	0,00	1,00	1,00	1,00	0,00	1,00	0,00
Detector06	0,00	0,00	0,16	0,52	0,00	0,00	0,00	0,00	0,00	0,02
Detector07	0,34	0,06	0,59	0,05	0,78	0,42	0,36	0,23	0,36	0,32
Detector08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Detector09	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,98	0,00
Detector10	0,89	0,95	1,00	0,80	1,00	1,00	1,00	1,00	1,00	1,00
Detector11	0,63	0,20	1,00	0,79	0,81	0,40	0,49	0,35	0,26	0,81
Detector12	0,49	0,37	0,50	1,00	1,00	0,79	1,00	1,00	1,00	1,00
Detector13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Detector14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

service for a fee.

It is important to emphasize that the authors of the 5 essays stated that they utilized ChatGPT, but most of the content was original. This should have resulted in values around 0.5 of AI-generated content. Table 3 shows the general inability of the detectors to identify AI-generated content. Some detectors very rarely identified AI-generated text (Detector08, Detector09, Detector13, Detector14). Certain detectors were unable to differentiate, consistently presenting the same values (Detector01, Detector04). Some detectors identified a too little amount of AI-generated text (Detector06, Detector12). Detector03 was excluded from the second round due to its extreme evaluations and technical difficulties associated with its applicability to a large number of documents. Detector02, Detector07, Detector10, and Detector11 were used for the 2nd run on all 31 essays.

The results are available in Table 4. The dismay is significant: the vertical reading of the table reveals that no detector can provide reliable results. In fact, a good detector should exhibit values higher than a threshold suggesting the use of AI. Furthermore, all students reported using ChatGPT in writing the five sections of the essays. We lack information regarding the percentage of actual text generated by AI. Detector02 often deemed the use of AI as unlikely, while Detector07 identified only a few texts as AI-written. Detector10 was perhaps the best performer, although it generally overestimated the percentage of AI-generated text and, in some cases (Essay03), failed to identify the text. Detector11 appears unbalanced in its detection capability, with few intermediate values. The horizontal analysis of the table reveals that only a few essays have been identified as written by AI (Essay05, Essay13, Essay25). Overall, each essay received different evaluations depending on the detector. In view of these findings, it is prudent for educators to exercise caution in relying solely on a detector for identifying text produced with LLM. Therefore, the prohibition of students' use of LLM is of limited value and warrants reconsideration.

3.5. Questionnaire results

To investigate students' perceptions, the results of the questionnaire were analyzed. Students answered the same questionnaire before and after using ChatGPT. Answers were given in the 4-point Likert scale previously introduced, except for two questions (A_4, A_5 use "yes"/"no" answer). For the analysis of the results, a descriptive analysis is provided due to the limited data sample. This visual representation of the distribution and frequency of responses aids in identifying patterns and trends, allowing for preliminary understanding and as a foundation for future investigation with larger sample sizes. The questionnaire administered after using LLM led to the expansion of the Usefulness & Performance Expectancy construct. This expansion involved investigating certain contents that required firsthand experience. Thus, the construct "Usefulness & Performance Expectancy (Post Assignment Questions)" was formed with additional questions. In the analysis of questionnaire results, each item is represented by 2 bars: the left bar shows the responses before using LLM, and the right bar displays the post-use responses. The results are categorized into six constructs and presented accordingly for each item (question) within the construct.

3.5.1. Attitude

The attitude construct investigated the impression, feelings, and

Table 3
LLM-DT 1st run notes.

4. The 2nd run test involves the analysis of 31 essays. Each essay is examined based on the 5 paragraphs required by the professor: “General Definitions and Potential Industrial Applications,” “State-of-the-Art,” “Enabling Factors for Implementation,” “Key Limitations and Challenges,” and “Industrial Context Transformation.” The obtained measurements are separately analyzed for each individual essay and aggregated across all essays.

Id	Detector name	1st round (basic test)	1st round notes	2nd round
1	A.I. Text Classifier	Yes	Consistently presenting the same values	No
2	Content @ Scale	Yes	Unable to detect human-generated text	Yes
3	Copyleaks	Yes	Strongly umbalanced	No
4	Crossplag	Yes	Consistently presenting the same values	No
5	GLTR	No	Excluded based on metric criteria	No
6	GPT-2 Output Detector	Yes	Detection capability is excessively low	No
7	GPTKit	Yes	AI detected on all essays	Yes
8	GPTZero	Yes	Detection capability is excessively low	No
9	Hive Moderation	Yes	Detection capability is excessively low	No
10	Originality.AI	Yes	AI detected on all essays	Yes
11	Sapling.ai	Yes	AI detected on all essays	Yes
12	Smodin - AI Detection	Yes	Detection capability is excessively low	No
13	Writer - AI content detector	Yes	Unable to detect AI-generated text	No
14	ZeroGPT	Yes	Unable to detect AI-generated text	No

preconceptions students had about ChatGPT, before and after they utilize it. The questionnaire results are shown in Fig. 5.

Half of the students demonstrate a basic understanding of programming languages and consider themselves capable of programming and staying up-to-date on Artificial Intelligence topics (A_1, A_2, A_3). Similarly, half of them have previously utilized ChatGPT (A_4), albeit primarily for non-academic purposes (A_5). Before engaging with ChatGPT, the sample showed two groups of students: those who were aware of its strengths and limitations and those who were not. However, after interacting with the model, both groups became more familiar with these aspects (A_6, A_7). Initially, a percentage of students expressed a lack of readiness to use ChatGPT in academic or professional contexts, but their firsthand experience helped foster a sense of preparedness (A_8, A_9). There is a widespread belief that employing ChatGPT can boost individuals’ confidence in tackling university assignments or work-related activities (A_10), while the notion that daily interaction with ChatGPT will lead to increased comfort is less reinforced (A_11).

3.5.2. *Trust*

The trust construct investigates the level of trust the students have before and after they utilize ChatGPT. The questionnaire results are shown in Fig. 6.

Overall, ChatGPT usage does not induce fear or dread, especially after using it (T_1). Students are highly curious about the tool, especially before using it (T_2), and they disagree that ChatGPT may induce addiction or separation anxiety (T_3). They believe its answers are reliable (T_4), but after using it a few of them change their opinion about the comprehensiveness and the accuracy of the answers, finding them less satisfactory (T_5, T_6). Generally, the answers are considered comprehensible (T_7). Contents are recognized as not up-to-date (T_8). Most of the sample worries that ChatGPT poses a threat to creativity and originality, especially after experiencing it (T9).

3.5.3. *Social influence*

The Social Influence constructs investigate the social influence that the context generates on students that utilize ChatGPT. The questionnaire results are shown in Fig. 7.

Before using ChatGPT, a clear minority of students feel not influenced by whether the people next to them talk about the tool or use it. This influence decreases after they use it (S_1, S_2). The message spread by social networks, TV, or newspapers seems less influential (S_3). Generally, ChatGPT is perceived as a tool to stay updated (S_4).

3.5.4. *Fairness & Ethics*

The Fairness & Ethics construct investigates if students judge ChatGPT fair, and compliance with ethical or moral standards before and after the technology implementation. The questionnaire results are shown in Fig. 8.

Most students disagree that the use of ChatGPT can help students pass the exam by reducing the actual learning of the content, especially after using it (FE_1). Pre-use, many students felt that the use of ChatGPT cannot result in an evaluation consistent with the actual level of learning. The use of ChatGPT slightly changes this belief (FE_2). Students are divided on the issues of privacy, intellectual property, and copyright, which therefore appear not indifferent concerns (FE_3, FE_4). This is confirmed by the shared belief that the use of ChatGPT in the university context should be regulated by the university, faculty, or department (FE_5). Despite this, most students consider ethically correct to use it at the university, and the number increases after use (FE_6). Students are aware that the answers provided by ChatGPT will be subject to bias (e.g., gender/context/social factors/geographical origin bias) (FE_7). Before they used it, they perceived a ChatGPT answer as not distinguishable from a human’s answer, but the idea changes markedly after use (FE_8). The idea that ChatGPT can be used to disseminate misleading or false information, encouraging misinformation, especially in the post-use phase, is not widely shared (FE_9).

3.5.5. *Usefulness & performance expectancy*

The Usefulness & Performance Expectancy construct investigates the students’ perception of the usefulness and performance of ChatGPT. The questionnaire results are shown in Fig. 9.

The use of ChatGPT in education is widely perceived as helpful to simplify and speed up the writing of the works, the understanding, and the learning of the subjects (UPE_1, UPE_2, UPE_3, UPE_4, UPE_5, UPE_6, UPE_7). So, students agree that the use of ChatGPT will be valuable for further examinations or in a work context (UPE_8). Less conviction emerges about the changing teacher’s role, as some students do not believe that ChatGPT can generate hybrid teaching approaches in the future (UPE_9). Instead, they appreciate its open way of accessing content and ubiquity, believing that it can facilitate content integration (UPE_10), and speed it up (UPE_11). The majority agree that it can improve student motivation in learning, by allowing them to analyze content in a fun and stimulating environment (UPE_12).

3.5.6. *Usefulness & Performance Expectancy (Post Assignment Questions)*

The Usefulness & Performance Expectancy (Post Assignment Questions) construct investigates the students’ judgment on the usefulness of ChatGPT and on the experienced performance after they utilize the technology, i.e., the perceived learning usefulness. The questionnaire results are shown in Fig. 10.

After using ChatGPT, only a minor part of students stated that the writing of the paper was not simplified (UPE_post_1). The same is true for contents’ understanding and learning (UPE_post_2, UPE_post_3). More conviction emerges in the ChatGPT’s ability to speed up the writing (UPE_post_4). Also, almost all students state that understanding and learning have been speeded up (UPE_post_5, UPE_post_6).

3.5.7. *Effort & ease of use*

The Effort & Ease of Use construct investigate the students’ effort and

Table 4
LLM-DT analysis – 2nd round results.

Detector02

Essay	Section1	Section2	Section3	Section4	Section5
Essay01	0,41	1,00	0,32	0,19	0,13
Essay02	0,01	0,00	0,00	0,00	0,06
Essay03	0,04	0,15	0,00	0,00	0,11
Essay04	0,06	0,00	0,07	0,26	0,17
Essay05	0,36	0,41	0,25	0,38	0,86
Essay06	0,50	0,58	0,62	0,53	0,86
Essay07	0,01	0,00	0,00	0,10	0,20
Essay08	0,41	0,28	0,15	0,85	0,31
Essay09	0,85	0,86	0,83	0,29	1,00
Essay10	0,00	0,21	0,14	0,25	0,13
Essay11	0,12	0,08	0,25	0,19	0,04
Essay12	0,16	0,07	0,06	0,22	0,74
Essay13	0,25	0,00	0,87	0,85	0,90
Essay14	0,01	0,11	0,00	0,05	0,00
Essay15	0,00	0,04	0,30	0,49	0,16
Essay16	0,80	0,31	0,87	0,32	0,89
Essay17	0,06	0,21	0,25	0,20	0,31
Essay18	0,27	0,42	0,32	0,22	0,26
Essay19	0,01	0,01	0,15	0,31	0,89
Essay20	0,06	0,05	0,11	0,14	0,12
Essay21	0,18	0,18	0,00	0,21	0,11
Essay22	0,38	0,88	0,69	0,47	0,89
Essay23	0,02	0,00	0,18	0,07	0,09
Essay24	0,10	0,01	0,02	0,16	0,15
Essay25	1,00	0,21	0,22	0,52	1,00
Essay26	0,11	0,07	0,88	0,55	0,25
Essay27	0,09	0,13	0,14	0,25	0,23
Essay28	0,00	0,00	0,00	0,00	0,18
Essay29	0,85	0,65	0,23	0,90	0,02
Essay30	0,10	1,00	0,83	0,89	1,00
Essay31	0,17	0,02	0,20	0,48	0,86

Detector07

Essay	Section1	Section2	Section3	Section4	Section5
Essay01	0,18	0,00	0,15	0,00	0,07
Essay02	0,04	0,02	0,06	0,00	0,00
Essay03	0,01	0,00	0,00	0,00	0,00
Essay04	0,13	0,03	0,14	0,05	0,01
Essay05	0,34	0,06	0,59	0,05	0,78
Essay06	0,10	0,30	0,03	0,15	0,57
Essay07	0,02	0,00	0,11	0,01	0,07
Essay08	0,07	0,08	0,00	0,02	0,12
Essay09	0,57	0,11	0,09	0,12	0,12
Essay10	0,00	0,01	0,01	0,01	0,00
Essay11	0,07	0,12	0,03	0,04	0,01
Essay12	0,15	0,04	0,28	0,01	0,32
Essay13	0,24	0,15	0,07	0,89	0,81
Essay14	0,02	0,01	0,00	0,01	0,04
Essay15	0,14	0,08	0,20	0,10	0,09
Essay16	0,02	0,01	0,01	0,04	0,73
Essay17	0,11	0,10	0,01	0,13	0,34
Essay18	0,24	0,45	0,00	0,41	0,01
Essay19	0,01	0,01	0,01	0,29	0,85
Essay20	0,04	0,00	0,06	0,02	0,00
Essay21	0,10	0,05	0,01	0,03	0,01
Essay22	0,42	0,36	0,23	0,36	0,32
Essay23	0,23	0,00	0,00	0,03	0,03
Essay24	0,00	0,11	0,01	0,00	0,09
Essay25	0,40	0,41	0,44	0,99	0,13
Essay26	0,40	0,05	0,13	0,29	0,01
Essay27	0,01	0,03	0,02	0,05	0,00
Essay28	0,01	0,00	0,01	0,03	0,03
Essay29	0,27	0,58	0,33	0,35	0,00
Essay30	0,05	0,13	0,01	0,42	0,13
Essay31	0,04	0,02	0,01	0,17	0,41

ease of use expected and required to implement ChatGPT, comparing it to the effort needed to perform the same task on their own. The questionnaire results are shown in Fig. 11.

There, the sample often appeared divided. Before using it, most students believed that ChatGPT can provide immediate assistance, but the number decreases after using it (EE_1). The same applied to the belief that ChatGPT provides ready-to-use answers (EE_2). Nearly nobody thinks that using ChatGPT requires more cognitive effort than the normal performance process (EE_3). Also, most believe that thanks to ChatGPT, the writing task requires less time (EE_4).

3.6. Questionnaire results discussion

The results of the experiment indicate that students had a general understanding of ChatGPT and expressed readiness to use it, attributing their familiarity to word-of-mouth and media exposure. Initially, students perceived the ChatGPT as reliable, demonstrating curiosity rather than fear. However, their perception of reliability shifted after using it, leading to a more nuanced understanding. While they still considered ChatGPT reliable and comprehensible in terms of content, they found it lacking in comprehensiveness and readiness for immediate use. Consequently, further elaboration was required to refine the output generated by the tool, particularly in the context of preparing a high-quality report for an exam. Paradoxically, this need for further elaboration proved valuable in identifying both the strengths and limitations of integrating LLM into a university setting.

One specific advantage identified by students was ChatGPT ability to enhance performance and speed up the completion of assignments. Moreover, students noted that it facilitated their understanding of complex topics by providing comprehensive yet simplified explanations. Importantly, students expressed confidence that ChatGPT would not replace the role of teachers but rather enhance teaching practices in terms of enjoyment, reduction of repetitive tasks, and support in essay and report writing. However, students also emphasized the necessity for regulated use of LLM during exams, suggesting the importance of establishing guidelines and achieving consistency across different courses.

Furthermore, students recognized the practical relevance of using

Detector10

Essay	Section1	Section2	Section3	Section4	Section5
Essay01	1,00	0,60	1,00	0,41	0,99
Essay02	0,34	0,60	0,01	0,95	0,90
Essay03	0,03	0,03	0,04	0,05	0,05
Essay04	0,99	0,97	0,97	1,00	0,70
Essay05	0,89	0,95	1,00	0,80	1,00
Essay06	1,00	1,00	1,00	1,00	1,00
Essay07	0,93	1,00	0,59	0,59	0,46
Essay08	1,00	1,00	0,01	1,00	0,20
Essay09	0,99	0,55	0,53	0,89	1,00
Essay10	0,55	0,09	0,71	0,95	0,64
Essay11	0,80	0,00	0,98	0,00	0,14
Essay12	0,57	0,57	0,06	0,50	0,86
Essay13	1,00	0,74	1,00	1,00	1,00
Essay14	0,49	0,42	0,12	0,39	0,53
Essay15	0,72	0,49	1,00	0,99	0,99
Essay16	0,92	0,98	1,00	1,00	1,00
Essay17	0,52	0,89	0,73	0,51	0,80
Essay18	1,00	0,93	0,99	0,91	0,05
Essay19	0,83	0,02	0,07	0,99	1,00
Essay20	0,63	0,00	0,13	0,00	0,40
Essay21	0,05	0,98	0,95	0,99	0,67
Essay22	1,00	1,00	1,00	1,00	1,00
Essay23	0,49	0,29	0,64	0,02	0,05
Essay24	1,00	1,00	1,00	0,98	1,00
Essay25	1,00	1,00	1,00	1,00	1,00
Essay26	0,29	1,00	1,00	0,98	0,02
Essay27	0,08	0,00	0,62	0,75	0,03
Essay28	0,00	1,00	0,60	0,19	0,45
Essay29	1,00	1,00	0,94	1,00	0,00
Essay30	1,00	1,00	1,00	1,00	1,00
Essay31	0,46	0,97	1,00	0,98	1,00

Detector11

Essay	Section1	Section2	Section3	Section4	Section5
Essay01	0,52	0,33	0,65	0,18	0,42
Essay02	0,00	0,24	0,13	0,12	0,10
Essay03	0,00	0,00	0,11	0,00	0,00
Essay04	0,10	0,64	0,27	0,24	0,15
Essay05	0,63	0,20	1,00	0,79	0,81
Essay06	0,05	0,68	0,02	0,53	0,78
Essay07	0,01	0,26	0,94	0,79	0,00
Essay08	0,69	0,12	0,27	1,00	0,60
Essay09	0,44	0,95	0,22	0,24	0,19
Essay10	0,51	0,01	0,00	0,00	0,00
Essay11	0,74	0,01	0,84	0,41	0,00
Essay12	0,23	0,16	0,43	0,00	0,65
Essay13	0,75	0,46	0,91	1,00	0,99
Essay14	0,25	0,37	0,00	0,25	0,91
Essay15	0,17	0,00	1,00	0,91	0,13
Essay16	0,91	0,00	0,65	1,00	0,91
Essay17	0,25	0,25	0,34	0,56	0,28
Essay18	0,63	0,54	0,00	0,86	0,14
Essay19	0,15	0,06	0,20	0,77	0,99
Essay20	0,19	0,22	0,24	0,27	0,26
Essay21	0,19	0,28	0,26	0,20	0,31
Essay22	0,40	0,49	0,35	0,26	0,81
Essay23	0,09	0,18	1,00	0,00	0,32
Essay24	0,00	0,10	0,03	0,19	0,07
Essay25	0,94	0,81	0,97	1,00	0,93
Essay26	0,77	0,77	0,99	1,00	0,00
Essay27	0,24	0,18	0,00	0,23	0,00
Essay28	0,07	0,26	0,18	0,22	0,20
Essay29	0,62	0,99	0,29	1,00	0,00
Essay30	0,00	0,64	0,91	1,00	0,64
Essay31	0,00	0,39	0,45	0,74	0,99

such tools in a university context, as they believed it would better prepare them for the future working world, where AI is anticipated to play a prominent role. However, when it comes to critical issues such as intellectual property, copyright, privacy, and bias, students displayed varying levels of preparedness, offering inconsistent and incomplete responses. Although the students demonstrated sensitivity toward these topics, their lack of awareness and interest raised concerns. Therefore, educators should consider this when introducing innovative technologies to students.

Specifically, considering RQ1, the use of LLM did not significantly affect students' performance. However many students did not adequately rework and discuss the text generated by ChatGPT. These results partially contradict the findings of (Yilmaz & Karaoglan Yilmaz, 2023), which reported clear performance \improvements on simple programming tasks. The differing level of complexity may account for this divergence in evidence.

The study evaluated various AI detection systems in response to RQ2, which revealed insights into the effectiveness and usability of different approaches. Unfortunately, the results were negative, as none of the tested detectors could be considered reliable in identifying text generated by AI. It is confirmed that employing large-scale language models (LLMs) poses a challenge to the detectors (Kumarage et al., 2023), and that there are limitations to these tools (Uzun, 2023).

Considering RQ3, the students perceived LLM as reliable, but their usage experience led to a deeper understanding of its limitations. The centrality of teachers emerged, thus the paper pertains to the research strand aimed at investigating the potential for overcoming traditional teaching and learning methodologies (Chigbu et al., 2023; Groff, 2013; Haleem et al., 2022; Kinshuk et al., 2016). This strand assumes that as the research and education landscapes evolve, adopting innovative tools and learning approaches is crucial for generating a flexible and effective environment.

An integrated combination of AI and human support is recommended for researchers, educators, and students (Alqahtani et al., 2023; Colabianchi, Tedeschi, & Costantino, 2023). As the guidelines given by Kim et al. (2023) stated, ChatGPT should not be considered an author in scientific manuscripts, retaining it capable of producing the entire manuscript; when the authors use it, they should have at least a basic

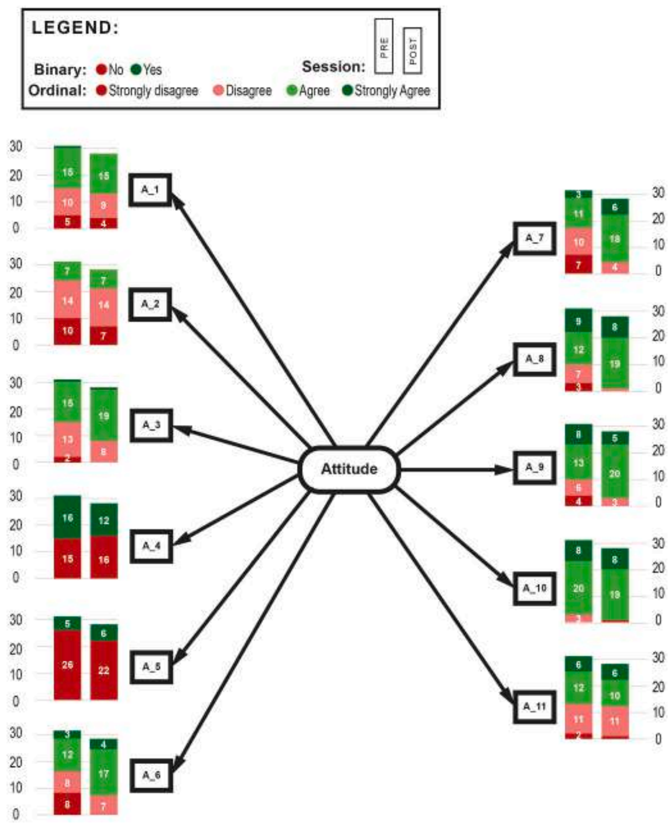

Fig. 5. Questionnaire results for the 11 items of the Attitude construct.

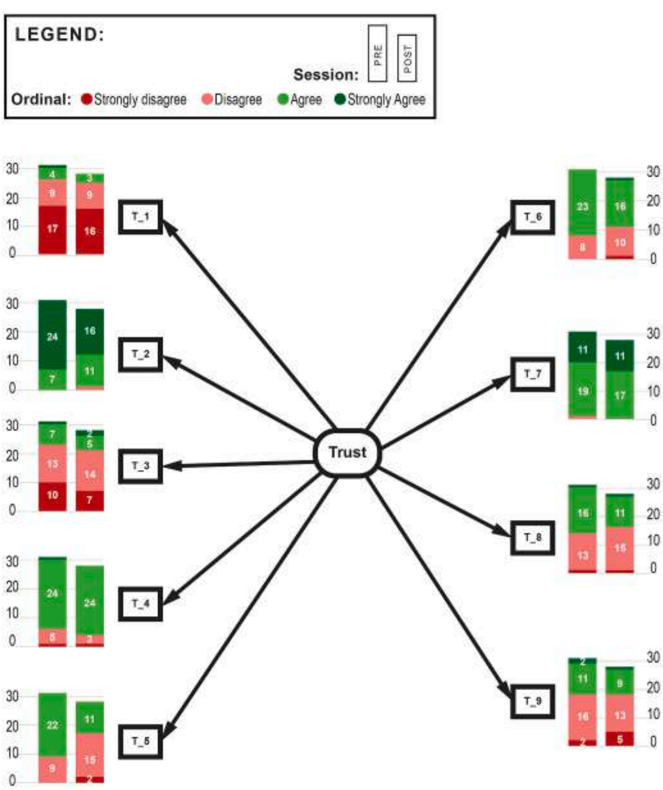

Fig. 6. Questionnaire results for the 9 items of the Trust construct.

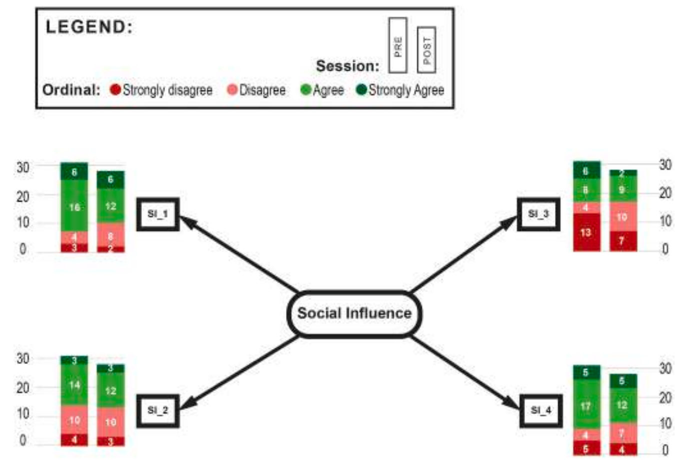

Fig. 7. Questionnaire results for the 4 items of the Social Influence construct.

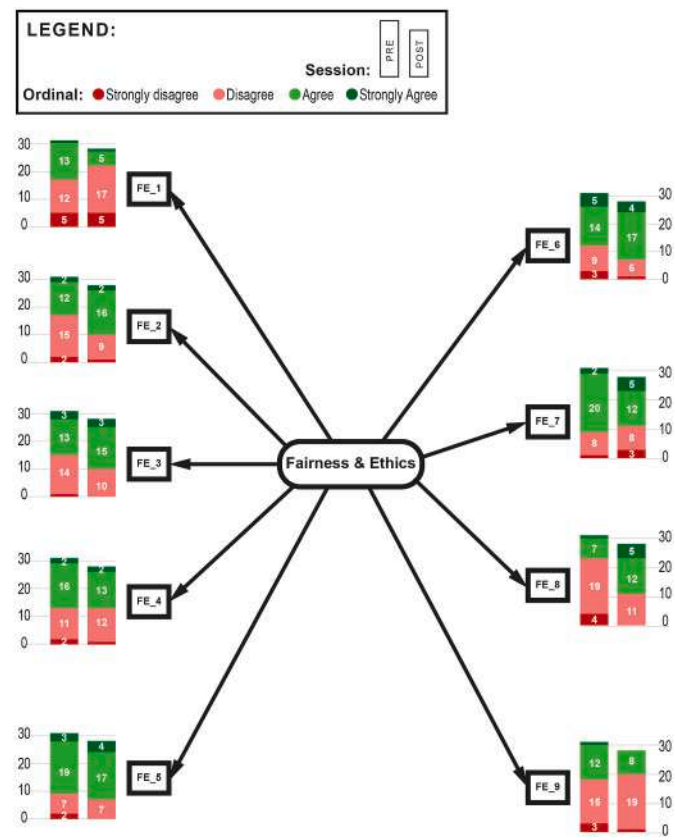

Fig. 8. Questionnaire results for the 9 items of the Fairness & Ethics construct.

understanding of what ChatGPT is, and they have to verify, edit and refine what ChatGPT generates. Also, it is necessary to address the challenges associated with these technologies, such as ethical concerns and algorithmic biases, for maximizing their potential to improve education and research outcomes (Alqahtani et al., 2023). This requires the implementation of mindful training, teaching, and learning strategies, and the setting of specific goals to be pursued to improve teaching. In terms of student perceptions, as also evidenced by Tsai et al. (2023), students express dissenting opinions on the usefulness of the tool, and common difficulties and misconceptions emerge. An extended sample of students perceived their ChatGPT's assisted performance to be as successful as that in conventional learning tasks (Shoufan, 2023), although the assisted task presented more challenges to them, as also occurred in

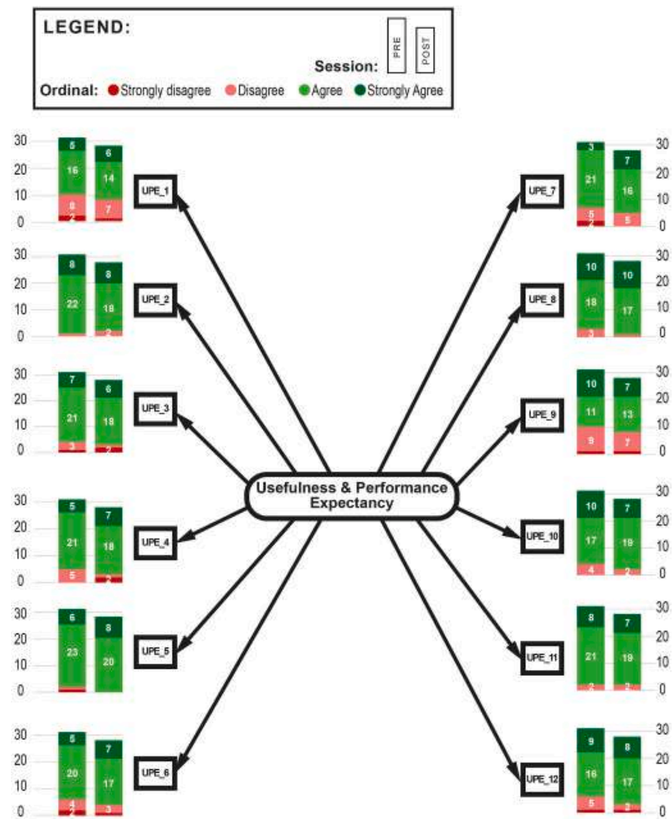

Fig. 9. Questionnaire results for the 12 items of the Usefulness & Performance Expectancy construct.

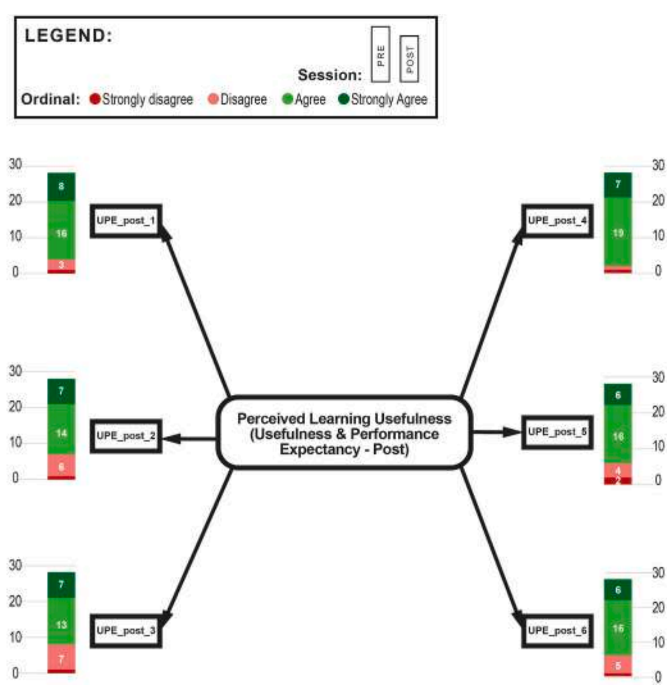

Fig. 10. Questionnaire results for the 6 items of the Usefulness & Performance Expectancy Post-use constructs.

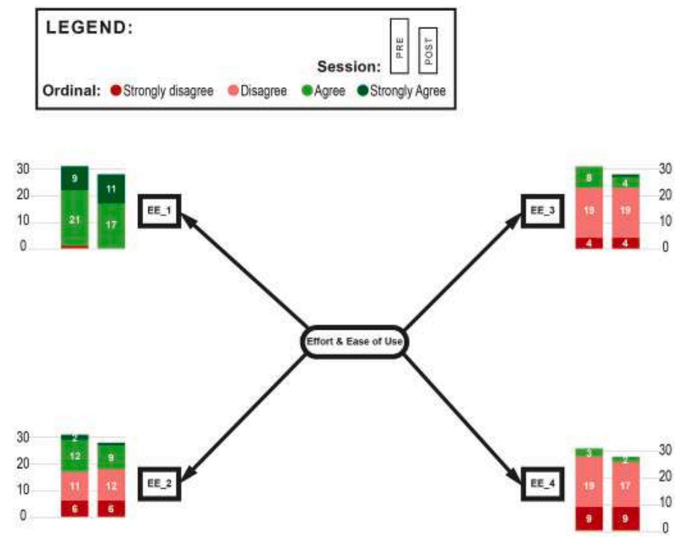

Fig. 11. Questionnaire results for the 4 items of the Effort & Ease of Use construct.

the research of Guo et al. (2023). The positive results, which in our study and in others, e.g. Li et al. (2023), happened in the task of writing an assignment, lead researchers to reflect on the potential of effectively integrating LLM technologies into education. Such potential can be better explore also considering the results of previous research and experiments in education, e.g., in medicine, pharmacy, and chemical engineering. In previous works, researchers investigate the state of acceptability of LLMs and ChatGPT and they offered a proposal for guidelines on utilization in the field (Kim et al., 2023). Moreover, the possibility of exploiting ChatGPT to generate reflective responses, to be combined with student-written reflection (Li et al., 2023), and for improving problem-solving efficiency (Tsai et al., 2023) was investigated. A LLM solution was developed to investigate the effects of chatbot-assisted in-class debates on students' argumentation skills and task motivation, by the generation of ideas for supporting students' positions and predicting opposing viewpoints (Guo et al., 2023). Additional innovative solutions were developed to complement traditional teaching and learning methods, to solve specific problems or cope with disabilities. Then, academics investigated the usability of the VR apps in experimental and development work (Radianti et al., 2020), the use of supplementary lecture recording to support students with specific learning difficulties (Nightingale et al., 2019), and the integration of multimedia materials to support the learning processes and outcomes in students with dyslexia, namely when they aim to learn factual knowledge (Knoop-van Campen et al., 2020).

4. Discussion on research findings

This research study explored the application of LLM in the education of engineers and examined students' acceptance, perceptions, and the impact of these technologies on the quality of education. The questionnaire used in the study was constructed based on the Technology Acceptance Model, the Unified Theory of Acceptance and Use of Technology, and its more updated versions (Davis, 1989; Venkatesh et al., 2003), and incorporated a BOCR analysis to identify relevant constructs and items.

The findings of the study, in response to Research Question 1, indicated that students produced good essays with well-distributed grades, suggesting that the use of LLM did not significantly affect their performance. In addition, students demonstrated their understanding of the topics during oral presentations, providing further evidence of the benefits of LLM assistance and the extent to which it did not undermine their preparation. However, as discussed in the previous section many students did not sufficiently rework and discuss the text produced by ChatGPT.

These insights led to defining recommendations for teachers and students related to the introduction of LLMs in higher education. First, LLMs tend to produce content that is accurate but shallow and not of high quality. Producing large volumes of text convinces people that they have obtained valuable content. Second, LLMs tend to use a lot of lists and bulleted lists. Often these are repeated with small differences. If the student does not rework the text by unifying it, they might get contradictory paragraphs. In addition, its knowledge base is based on both encyclopedic material and web news. Such knowledge generates on timely topics such as simulation or DTs, confusing or contradictory answers.

Particularly on topics such as DTs and simulation although their differences were explained during previous lectures, some students when ChatGPT presented the two technologies as equivalent did not question the answer they received. Finally, ChatGPT do not take into account bibliographic sources and is considered a "closed model." Indeed, it is not possible to trace the data on which ChatGPT was trained or on which it generated the answer. If it was trained on incorrect or misleading data, it will pull them out. Therefore, it is recommended to supplement and verify what the tool produced with peer-reviewed scientific papers. Finally, a positive observation also emerged during the experiment. Some students used the tool to get better explanations of complex concepts found in some scientific publications. Specifically, they provided such topics as prompts to ChatGPT asking them to paraphrase them in more straightforward words. Once better understood they were integrated into the essay. This approach is indeed a strategy to suggest.

To answer Research Question 2, the study tested different AI detection systems, providing insights into the effectiveness and usability of different approaches, with a negative outcome. None of the tested detectors can be deemed reliable in identifying text generated by AI, thus it is advisable for teachers not to depend solely on a detector for identifying text generated using LLM. Consequently, enforcing restrictions on students' utilization of LLM holds little significance and should be reevaluated. The educational sector should invest on the development of students' critical thinking, problem analysis, and proficiency in seeking effective solutions (Yu, 2023), considering every available technological solution. To attain this objective, educational institutions should concentrate on fostering students' competencies to effectively harness technology for societal advancement, adeptly utilize data and analytical methodologies to make informed decisions, and critically assess and evaluate artificial intelligence with proficiency.

Finally, concerning Research Question 3, the research shed light on the evolving perspectives of students regarding the integration of LLM into their academic experiences. Although students initially perceived the tool as reliable, their usage experience led to a deeper understanding of its limitations. Nevertheless, students recognized the potential benefits of LLM in improving task performance, enhancing understanding, and supporting teachers. They stressed the importance of the teacher, who with his or her emotional skills and knowledge remains the expert in the loop of education and yet can be supported by AI in terms of innovation, fun, and interaction. Moreover, they highlighted the need for regulated use of LLM during exams, consistent guidelines across courses, and adequate preparation for an AI-driven future. The study also underscored the necessity for educators to address concerns related to intellectual property, copyright, privacy, and bias, as students displayed varying levels of knowledge and engagement in these areas.

The outcomes presented confirm some findings disclosed in the literature. The effectiveness of implementing LLM technologies for educational purposes depends on how they are implemented by teachers and perceived by students. In general, LLM technologies should not be an alternative to human work. Students are advised to use a combination of AI and human support. Essays should not consider ChatGPT as an author. Instead, authors should have a basic understanding of ChatGPT and verify, edit, and refine what it generates. To maximize the potential of technologies in improving education and research outcomes, it is essential to address the challenges associated with ethical concerns and algorithmic biases. This requires the implementation of mindful learning strategies, and the setting of specific goals to be pursued to improve teaching and avoid common difficulties and misconceptions in the LLM usage. The results of our study suggest that LLM technologies can be effectively integrated into education.

We, the researchers, believe that ChatGPT can be a valuable support tool for students to speed up comprehension and learning of contents and to cope with difficulties in learning, provided that it is properly cast in context and integrated with other innovative solutions proposed in the literature. However, effective implementation of ChatGPT requires certain prerequisites, including training of professors and students, the ability to critically analyze and verify the outputs of the tool, and supplementing them with additional content. Furthermore, it is essential to set narrow objectives that are different or multiple, such as using ChatGPT to speed up or simplify the delivery of an assignment or cope with disabilities. In this way, students can get reliable support, and they can acquire vital Industry 4.0 skills for contemporary industrial practices, and relevant issues nowadays (Tsai et al., 2023). Also, effective classroom orchestration systems can establish, wheatear careful architectures for educational technology systems, and focused teaching and learning strategies arise (Feng et al., 2023). Although this study has its limitations, such as the limited sample size representing only two engineering courses, the methodology, constructs, and items used are scalable to other courses for future testing. Additionally, the study proposed a path model through the questionnaire, but its validation through a complete path analysis was not possible due to the limited sample size. Further research is needed to validate and extend the findings of this study. Finally, this research provides valuable insights into the integration of LLMs in engineering education, highlighting their benefits, limitations, and considerations for effective implementation. The findings provide a foundation for further exploration and can guide the development of guidelines and strategies to harness the full potential of LLMs in educational settings.

Further examination could be conducted by considering whether LLMs help to reach the expected learning outcomes of engineering courses. If we go back to the starting examples of calculators and CAD software, we accept their usage because the objective of the learning process is not the capability to math or draw. As part of the Bologna Process, an initiative aimed at reforming and harmonizing university education throughout Europe, a key objective was the enhancement of higher education by establishing clear expected learning outcomes for every course/module. Therefore, it is essential to evaluate whether LLMs can help students achieve these expected learning outcomes (Kennedy et al., 2007). Do LLMs contribute to the achievement of these expected learning outcomes? Returning to the initial examples of calculators and CAD software, their acceptance stems from the underlying concept that the learning process aims not at the ability to perform mathematical calculations or draw designs. Generally, the expected learning outcomes in engineering university courses consider problem solving objectives (e.g. "Develop the ability to analyze complex engineering problems and devise effective solutions using [the specific subject]"), technical knowledge objectives (e.g. "Acquire a solid foundation in the fundamental principles and theories of [the specific subject]"), design skills objectives (e.g. "Learn how to design and evaluate engineering systems, components, or processes using [the specific subject]"), engineering analysis objectives (e.g. "Develop skills in analyzing and interpreting data, applying mathematical models, and using appropriate engineering tools and software"). It is an infrequent occurrence for the projected learning outcomes to encompass objectives such as "mastering the capacity to discern, evaluate, and select pertinent information sources in formulating a personalized viewpoint on [the specific subject]," or "acquiring the proficiency to write a well-structured essay discussing a specific topic within [the specific subject]". These latter examples are instances of skills that risk being replaced by LLMs. Therefore, it is essential for the education community to develop novel approaches to assess students' proficiency in utilizing the content of a university course, encompassing their ability to effectively analyze and solve complex problems that surpass mere information retrieval, content selection, integration, and accurate paraphrasing.

5. Conclusions and future research

This research has provided valuable insights into the application of Large Language Models (LLMs), with a specific focus on ChatGPT, in the context of engineering higher education. We addressed three key research questions: the ability of engineering students to produce quality essays with LLM assistance, the effectiveness of LLM identification systems, and students' perceptions of LLMs in terms of usefulness and acceptance in the learning process. Our findings indicate that engineering students can indeed generate high-quality essays with the help of LLMs, although recommendations were made to ensure deeper engagement and critical review of the generated content. We also discovered that existing LLM identification systems currently lack the reliability to detect essays produced with LLMs, suggesting the need for a reevaluation of restrictions on LLM usage in educational settings. Furthermore, our study illuminated the evolving perspectives of students regarding the integration of LLMs into their academic experiences. While students initially perceived LLMs as reliable tools, their usage led to a deeper understanding of both their benefits and limitations. They emphasized the crucial role of teachers in the educational process, highlighting the need for guidelines, preparation for an AI-driven future, and addressing ethical concerns and biases. In summary, this research underscores the potential of LLM technologies, like ChatGPT, as valuable support tools for students in engineering education. However, effective implementation requires careful consideration, training, and critical analysis of the tool's outputs. It is essential to maintain a balance between AI assistance and human involvement, emphasizing the development of critical thinking and problem-solving skills. Our work contributes to the ongoing discourse on the integration of LLMs in education and lays the foundation for further exploration and the development of guidelines to harness their full potential in educational settings. As we move forward, it is imperative that we continue to evaluate how LLMs align with expected learning outcomes and adapt our educational strategies accordingly, ensuring that they enhance rather than replace essential skills and competencies in our students.

Statements on open data and ethics

All participants provided informed consent before participating. The data were collected anonymously, and the study was designed to minimize any potential risks or discomfort for the participants. The study was conducted following the ethical guidelines of the University of Rome Sapienza's ethics committee and it did not need the ethical approval from the research review committee in the authors' affiliations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.caeai.2023.100172.

Appendix

Complete final version of the questionnaire.

Construct	Items
Attitude	A_1 I have a basic knowledge of programming languages.
	A_2 I am able to code programming.
	A_3 I am updated on news on Artificial Intelligence themes.
	A_4 I have already used ChatGPT.
	A_5 I have already used ChatGPT for a university assignment.
	A_6 I know the strengths of ChatGPT.
	A_7 I know the limitations of ChatGPT.
	A_8 I consider myself ready to use ChatGPT in a university context.
	A_9 I consider myself ready to use ChatGPT in a work context.
	A_10 Using ChatGPT can make people feel more confident in carrying out university tasks or work activities.
	A_11 Daily interaction with ChatGPT will, in the hypothetical future, make people feel comfortable.
Trust	T_1 The use of ChatGPT may induce fear or dread.
	T_2 ChatGPT is a tool that can induce curiosity.
	T_3 The use of ChatGPT may induce addiction or a separation anxiety.
	T_4 ChatGPT's answers are reliable (truthful).
	T_5 ChatGPT answers are exhaustive.
	T_6 ChatGPT answers are accurate (detailed).

(*continued on next page*)

(*continued*)

Construct	Items
	T_7 ChatGPT answers are comprehensible.
	T_8 ChatGPT's answers are topical/up-to-date.
	T_9 ChatGPT poses a threat to creativity and originality.
Social Influence	SI_1 I plan to use ChatGPT because people around me have mentioned it.
	SI_2 I plan to use ChatGPT because people around me use it.
	SI_3 I plan to use ChatGPT because I heard about it on social networks/TV/newspapers.
	SI_4 I plan to use ChatGPT to stay updated.
Fairness & Ethics	FE_1 The use of ChatGPT can help students pass the examination by reducing the actual learning of the content.
	FE_2 The use of ChatGPT can result in an overall evaluation consistent with the actual level of learning.
	FE_3 The privacy of data is influential in the use of ChatGPT.
	FE_4 Intellectual property and copyright issues are influential in the use of ChatGPT.
	FE_5 The use of ChatGPT in the university context should be regulated by the university/faculty/department.
	FE_6 The use of ChatGPT for writing a university assignment is ethically correct.
	FE_7 Answers provided by ChatGPT will be subject to bias (e.g. gender/context/social factors/geographical origin bias).
	FE_8 A ChatGPT answer is distinguishable from a human being's answer.
	FE_9 ChatGPT can be used to disseminate misleading or false information, encouraging misinformation.
Usefulness & Performance Expectancy	UPE_1 The use of ChatGPT will establish itself in education (compulsory schooling, universities, training courses etc.).
	UPE_2 Using the results provided by ChatGPT will simplify the conduct of reports/essays/written work.
	UPE_3 Using the results provided by ChatGPT will simplify understanding of the subject matter.
	UPE_4 Using the results provided by ChatGPT will simplify learning of the subject matter.
	UPE_5 Using the results provided by ChatGPT will speed up the conduct of reports/essays/written work.
	UPE_6 Using the results provided by ChatGPT will speed up understanding of the subject matter.
	UPE_7 Using the results provided by ChatGPT will speed up learning of the subject matter.
	UPE_8 The use of ChatGPT will be useful for further examinations or in a work context.
	UPE_9 ChatGPT may, in the future, generate hybrid teaching approaches, working alongside teachers in the teaching role.
	UPE_10 ChatGPT will facilitate the integration of information into teaching, due to the open way of accessing content.
	UPE_11 ChatGPT will speed the integration of information into teaching, due to the ability to access content at any time and from any place.
	UPE_12 ChatGPT will motivate students in learning by allowing them to analyze content in a fun and stimulating environment.
Usefulness (Post Assignment Questions)	UPOST_1 Using the results provided by ChatGPT simplified the conduct of reports/essays/written work.
	UPOST_2 Using the results provided by ChatGPT has simplified understanding of the subject.
	UPOST_3 Using the results provided by ChatGPT has simplified learning of the subject.
	UPOST_4 Using the results provided by ChatGPT has made it faster to write reports/essays/written work.
	UPOST_5 Using the results provided by ChatGPT has sped up the understanding of the subject.
	UPOST_6 Using the results provided by ChatGPT has sped up the learning of the subject.
Effort & Ease of Use	EEU_1 ChatGPT can provide immediate assistance in researching and formulating content, representing instant support to students/workers.
	EEU_2 ChatGPT answers are ready-to-use (directly useable without processing).
	EEU_3 Using ChatGPT for the performance of written papers requires more cognitive effort than the normal performance process.
	EEU_4 Using ChatGPT to carry out written papers requires more time than the normal unfolding process.

Assignments assessment.

Assignment	Essay Evaluation (scale 0–30)	Oral Presentazion Evaluation (scale 0–30)	Assignment Evaluation (scale 0–30)	Assignment requests were met.	The content is consistent with the requirements.	Technology description is comprehensive.	Technology description is accurate.	The essay has quality bibliographic sources.	SWOT analysis is comprehensive.	SWOT analysis is thoroughly discussed.	The essay content proves to be reworked.	The essay has adequate and well-placed bibliographic sources.
Assignment 1	23	30	27	0	0	1	1	0	0	0	0	0
Assignment 2	26	30	28	1	0	0	0	1	1	0	0	1
Assignment 3	29	24	27	1	1	1	1	0	1	1	1	0
Assignment 4	26	26	26	1	1	0	0	0	1	1	0	0
Assignment 5	26	30	28	1	1	1	1	1	0	0	0	0
Assignment 6	27	25	26	1	1	1	0	0	1	1	1	0
Assignment 7	30	26	28	1	1	0	1	1	1	1	1	1
Assignment 8	24	26	25	1	1	0	1	0	0	0	1	0
Assignment 9	30	24	27	1	1	1	1	1	1	1	1	1
Assignment 10	25	24	25	1	1	1	0	0	0	0	0	0
Assignment 11	26	25	26	1	1	1	1	0	0	0	0	0
Assignment 12	26	25	26	1	1	0	0	0	0	1	0	0
Assignment 13	30	25	28	1	1	0	1	1	0	1	1	1
Assignment 14	27	26	27	1	1	1	1	1	0	0	1	1
Assignment 15	30	30	30	1	1	1	1	0	1	1	1	0
Assignment 16	29	25	27	1	1	0	0	1	1	1	1	1
Assignment 17	29	30	30	1	1	0	0	1	1	1	1	1
Assignment 18	27	30	29	1	1	1	1	1	0	0	1	1
Assignment 19	30	25	28	1	1	1	1	1	1	0	0	1
Assignment 20	26	30	28	1	1	1	0	1	1	1	0	0
Assignment 21	29	25	27	1	1	1	1	1	0	0	1	1
Assignment 22	26	30	28	1	1	1	0	0	1	1	1	0
Assignment 23	27	30	29	1	1	1	0	1	0	0	0	0
Assignment 24	17	30	24	0	0	0	0	0	0	0	0	0
Assignment 25	25	24	25	1	1	1	1	0	1	1	0	0
Assignment 26	26	25	26	1	1	1	1	1	0	0	0	1
Assignment 27	26	25	26	1	1	0	1	1	0	0	0	1
Assignment 28	27	26	27	1	1	1	0	1	1	1	0	1
Assignment 29	20	30	25	0	0	1	1	0	0	0	0	0
Assignment 30	23	30	27	1	0	1	0	0	1	1	0	0
Assignment 31	30	30	30	1	1	1	1	1	1	1	1	1

References

Afonso, C. M., Roldán, J. L., Sánchez-Franco, M., & O, M. De (2012). The moderator role of gender in the unified theory of acceptance and use of technology (UTAUT): A study on users of electronic document management systems. In *7th international conference on partial least squares and related methods* (pp. 1–8).

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alowais, S. A., Alshaya, O. A., Rahman, I., Al, M. S., & Albekairy, A. M. (2023). Research in Social and Administrative Pharmacy the emergent role of artificial intelligence , natural learning processing , and large language models in higher education and research. *Research in Social and Administrative Pharmacy*. https://doi.org/10.1016/j.sapharm.2023.05.016. April.

Ayaz, A., & Yanartaş, M. (2020). An analysis on the unified theory of acceptance and use of technology theory (UTAUT): Acceptance of electronic document management system (EDMS). *Computers in Human Behavior Reports, 2*(March). https://doi.org/10.1016/j.chbr.2020.100032

Aziz, F., Rami, A. M., Rasdi, R. M., & Aina, N. (2021). The integration of work ethics and technology acceptance towards enhancing online learning environment among lecturers. *Turkish Journal of Computer and Mathematics Education, 12*(9), 2979–2982.

Chigbu, B. I., Ngwevu, V., & Jojo, A. (2023). The effectiveness of innovative pedagogy in the industry 4.0: Educational ecosystem perspective. *Social Sciences & Humanities Open, 7*(1), Article 100419. https://doi.org/10.1016/j.ssaho.2023.100419

Choi, B. C., Pak, A. W., & Cdc, for (2005). *A catalog of biases in questionnaires*.

Colabianchi, S., Bernabei, M., & Costantino, F. (2022). Chatbot for training and assisting operators in inspecting containers in seaports. *Transportation Research Procedia, 64*, 6–13. https://doi.org/10.1016/j.TRPRO.2022.09.002

Colabianchi, S., Tedeschi, A., & Costantino, F. (2023). Human-technology integration with industrial conversational agents: A conceptual architecture and a taxonomy for manufacturing. *Journal of Industrial Information Integration, 35*, 100510. https://doi.org/10.1016/j.jii.2023.100510

Cribben, I., & Zeinali, Y. (2023). The benefits and limitations of ChatGPT in business education and research: A focus on management science, operations management and data analytics. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4404276

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems, 13*(3), 319–339. https://doi.org/10.2307/249008

Dizon, G., Tang, D., & Yamamoto, Y. (2022). A case study of using Alexa for out-of-class, self-directed Japanese language learning. *Computers & Education: Artificial Intelligence, 3*(March), Article 100088. https://doi.org/10.1016/j.caeai.2022.100088

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, Article 102642. https://doi.org/10.1016/J.IJINFOMGT.2023.102642

Feng, S., Zhang, L., Wang, S., & Cai, Z. (2023). Effectiveness of the functions of classroom orchestration systems: A systematic review and meta-analysis. *Computers & Education.* , Article 104864. https://doi.org/10.1016/j.compedu.2023.104864

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States medical licensing examination? The implications of Large Language Models for medical education and knowledge assessment. *JMIR Medical Education, 9*. https://doi.org/10.2196/45312

Glette, M. K., & Wiig, S. (2022). The headaches of case study research: A discussion of emerging challenges and possible ways out of the pain. *Qualitative Report, 27*(5), 1377–1392. https://doi.org/10.46743/2160-3715/2022.5246

Groff, J. (2013). Technology-rich innovative learning environments. *Oecd.Org*, 1–30.

Guo, K., Zhong, Y., Li, D., Kai, S., & Chu, W. (2023). Effects of chatbot-assisted in-class debates on students ' argumentation skills and task motivation. *Computers & Education, 203*(March), Article 104862. https://doi.org/10.1016/j.compedu.2023.104862

Haleem, A., Javaid, M., Qadri, M. A., & Suman, R. (2022). Understanding the role of digital technologies in education: A review. *Sustainable Operations and Computers, 3* (May), 275–285. https://doi.org/10.1016/j.susoc.2022.05.004

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 1–20. https://doi.org/10.1007/S10639-023-11834-1/TABLES/5

Ji, H., Han, I., & Ko, Y. (2022). *A systematic review of conversational AI in language education: Focusing on the Collaboration with Human Teachers*. https://doi.org/10.1080/15391523.2022.2142873

Johnson, A. (2023). *ChatGPT in schools: Here's where it's banned—and how it could potentially help students*. Forbes. https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/.

Kabanda, S., & Brown, I. (2017). A structuration analysis of Small and Medium Enterprise (SME) adoption of E-Commerce: The case of Tanzania. *Telematics and Informatics, 34*(4), 118–132. https://doi.org/10.1016/j.tele.2017.01.002

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, 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*, Article 102274. https://doi.org/10.1016/J.LINDIF.2023.102274

Katz, D. M., Bommarito, M. J., Gao, S., & David Arredondo, P. (2023). *GPT-4 Passes the Bar Exam. March 15*. https://doi.org/10.2139/ssrn.4389233

Kennedy, D., Hyland, A., & Ryan, N. (2007). Writing and using learning outcomes: A practical guide. *Behaviour & Information Technology*, 1–30. *August* https://cora.ucc.ie/handle/10468/1613.

Keržič, D., Tomaževič, N., Aristovnik, A., & Umek, L. (2019). Exploring critical factors of the perceived usefulness of blended learning for higher education students. *PLoS One, 14*(11), 1–18. https://doi.org/10.1371/journal.pone.0223767

Kim, J. K., Chua, M., & Rickard, M. (2023). ChatGPT and large language model (LLM) chatbots : The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology, xxxx*. https://doi.org/10.1016/j.jpurol.2023.05.018

Kinshuk, Chen, N. S., Cheng, I. L., & Chew, S. W. (2016). Evolution is not enough: Revolutionizing current learning environments to Smart learning environments. *International Journal of Artificial Intelligence in Education, 26*(2), 561–581. https://doi.org/10.1007/s40593-016-0108-x

Knoop-van Campen, C. A. N., Segers, E., & Verhoeven, L. (2020). Effects of audio support on multimedia learning processes and outcomes in students with dyslexia. *Computers & Education, 150*(February), Article 103858. https://doi.org/10.1016/j.compedu.2020.103858

Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., & Liu, H. (2023). *Stylometric detection of AI-generated text in twitter timelines*. https://arxiv.org/abs/2303.03697v1.

Lavarda, R. B., & Bellucci, C. F. (2022). Case study as a suitable method to research strategy as practice perspective. *Qualitative Report, 27*(2), 539–555. https://doi.org/10.46743/2160-3715/2022.4296

Lee, A. H., Kang, H.-Y., & Chang, C.-C. (2011). An integrated interpretive structural modeling-fuzzy analytic network process-benefits, opportunities, costs and risks model for selecting technologies. *International Journal of Information Technology and Decision Making, 10*(5). https://doi.org/10.1142/S0219622011004592

Lesage, J., Brennan, R., Eaton, S. E., Moya, B., McDermott, B., Wiens, J., & Herrero, K. (2023). Exploring natural language processing in mechanical engineering education: Implications for academic integrity. *International Journal of Mechanical Engineering Education*. https://doi.org/10.1177/03064190231166665/ASSET/IMAGES/LARGE/10.1177_03064190231166665-FIG6.JPEG

Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large language models write reflectively. *Computers & Education: Artificial Intelligence, 4*(April). https://doi.org/10.1016/j.caeai.2023.100140

Mageira, K., Pittou, D., Papasalouros, A., Kotis, K., Zangogianni, P., & Daradoumis, A. (2022). Educational AI chatbots for content and language integrated learning. *Applied Sciences, 12*(7). https://doi.org/10.3390/app12073239

Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. *SSRN Electronic Journal*, 1–19. https://doi.org/10.2139/ssrn.4354422

Mosweu, O., Bwalya, K. J., & Mutshewa, A. (2017). A probe into the factors for adoption and usage of electronic document and records management systems in the Botswana context. *Information Development, 33*(1), 97–110. https://doi.org/10.1177/0266666916640593

Nightingale, K. P., Anderson, V., Onens, S., Fazil, Q., & Davies, H. (2019). Developing the inclusive curriculum: Is supplementary lecture recording an effective approach in supporting students with Specific Learning Difficulties (SpLDs)? *Computers & Education, 130*, 13–25. https://doi.org/10.1016/j.compedu.2018.11.006. June 2018.

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers & Education: Artificial Intelligence, 2*, Article 100033. https://doi.org/10.1016/j.caeai.2021.100033

OpenAI. (2022). *Introducing ChatGPT*.

OpenAI. (2023). GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf.

Osmani, M., El-Haddedeh, R., & Hindi, N. (2021). Blockchain for next generation services in banking and finance: Cost, benefit, risk and opportunity analysis. *Journal of Enterprise Information Management, 34*(3). https://doi.org/10.1108/JEIM-02-2020-0044

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education, 147*(December 2019), Article 103778. https://doi.org/10.1016/j.compedu.2019.103778

Rashid, Y., Rashid, A., Warraich, M. A., Sabir, S. S., & Waseem, A. (2019). Case study method: A step-by-step guide for business researchers. *International Journal of Qualitative Methods, 18*. https://doi.org/10.1177/1609406919862424/ASSET/IMAGES/LARGE/10.1177_1609406919862424-FIG1.JPEG

Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using Large Language Models. In *ICER 2022 - proceedings of the 2022 ACM conference on international computing education research* (Vol. 1, pp. 27–43). https://doi.org/10.1145/3501385.3543957

Sezer, B., & Yilmaz, R. (2019). Learning management system acceptance scale (LMSAS): A validity and reliability study. *Australasian Journal of Educational Technology, 35*(3), 15–30. https://doi.org/10.14742/ajet.3959

Shoufan, A. (2023). Exploring students' perceptions of ChatGPT: Thematic analysis and follow-up survey. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3268224

Tabatabaee, S., Mahdiyar, A., Durdyev, S., Mohandes, S. R., & Ismail, S. (2019). An assessment model of benefits, opportunities, costs, and risks of green roof installation: A multi criteria decision making approach. *Journal of Cleaner Production, 238*, Article 117956. https://doi.org/10.1016/j.jclepro.2019.117956

Tang, R., Chuang, Y.-N., & Hu, X. (2023). ArXiv Preprint ArXiv:2303.07205. *The science of detecting LLM-generated texts; the science of detecting LLM-generated texts*.

Tsai, M. L., Ong, C. W., & Chen, C. L. (2023). Exploring the use of large language models (LLMs) in chemical engineering education: Building core course problem models with Chat-GPT. *Education for Chemical Engineers, 44*, 71–95. https://doi.org/10.1016/J.ECE.2023.05.001

Tuyet Mai, N. T., Yoshi, T., & Tuan, N. P. (2013). Technology acceptance model and the paths to online customer loyalty in an emerging market. *Tržište, 25*(2), 231–248.

Uzun, L. (2023). ChatGPT and academic integrity concerns: Detecting artificial intelligence generated content. *Language Education and Technology, 3*(1), 45–54. http://www.langedutech.com/letjournal/index.php/let/article/view/49.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly: Management Information Systems, 27*(3), 425–478.

Wang, J., Hwang, G. H., & Chang, C. Y. (2021). Directions of the 100 most cited chatbot-related human behavior research: A review of academic publications. *Computers & Education: Artificial Intelligence, 2*, Article 100023. https://doi.org/10.1016/j.caeai.2021.100023

Wang, Y. S., & Shih, Y. W. (2009). Why do people use information kiosks? A validation of the unified theory of acceptance and use of technology. *Government Information Quarterly, 26*(1), 158–165. https://doi.org/10.1016/j.giq.2008.07.001

Warmbrod, J. R. (2014). Reporting and interpreting scores derived from likert-type scales. *Journal of Agricultural Education, 55*(5), 30–47. https://doi.org/10.5032/jae.2014.05030

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of Large Language Models*. https://arxiv.org/abs/2206.07682v2.

Wijnmalen, D. J. D. (2007). Analysis of benefits, opportunities, costs, and risks (BOCR) with the AHP-ANP: A critical validation. *Mathematical and Computer Modelling, 46* (7–8), 892–905. https://doi.org/10.1016/j.mcm.2007.03.020

Wu, K., Zhao, Y., Zhu, Q., Tan, X., & Zheng, H. (2011). A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. *International Journal of Information Management, 31*(6), 572–581. https://doi.org/10.1016/j.ijinfomgt.2011.03.004

Yilmaz, R., & Karaoglan Yilmaz, F. G. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers & Education: Artificial Intelligence, 4*, Article 100147. https://doi.org/10.1016/J.CAEAI.2023.100147

Yin, J., Goh, T. T., Yang, B., & Xiaobin, Y. (2021). Conversation technology with micro-learning: The impact of chatbot-based learning on students' learning motivation and performance. *Journal of Educational Computing Research, 59*(1), 154–177. https://doi.org/10.1177/0735633120952067

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology, 14*. https://doi.org/10.3389/FPSYG.2023.1181712

Zakeri, S. M. H., Tabatabaee, S., Ismail, S., Mahdiyar, A., & Wahab, M. H. (2023). Developing an MCDM model for the benefits, opportunities, costs and risks of BIM adoption. *Sustainability, 15*(5), 1–19. https://doi.org/10.3390/su15054035