*Article*

# Bridging LLMs, Education, and Sustainability: Guiding Students in Local Community Initiatives

Nebojša Jurišević [1,*], Novak Nikolić [1], Artur Nemś [2], Dušan Gordić [1], Nikola Rakić [1], Davor Končalović [1] and Dénes Kocsis [3]

1    Faculty of Engineering, University of Kragujevac, 34000 Kragujevac, Serbia
2    Faculty of Mechanical and Power Engineering, Wrocław University of Science and Technology, 50-370 Wroclaw, Poland; artur.nems@pwr.edu.pl
3    Faculty of Engineering, University of Debrecen, 4032 Debrecen, Hungary; kocsis.denes@eng.unideb.hu
*    Correspondence: jurisevic@kg.ac.rs

## Abstract

The introduction of large language models (LLMs) has significantly influenced learning and learning assessments, dividing the academic community with arguments for and against their implementation. This study investigates how LLMs can be effectively incorporated into student assignments on sustainable development in local communities. In that regard, the study pairs traditional, community-oriented tasks with emerging frameworks for structured LLM use, emphasizing that output quality depends on prompt quality. Accordingly, several prompting frameworks were outlined, and the suitability of ChatGPT and Gemini for specific assignment tasks was assessed. The effectiveness of the approach was evaluated with a survey of two student groups: one using supervised LLM support (23 students) and another using LLMs independently (17 students). Compared to the unsupervised group, the supervised group reported that the frameworks enhanced project preparedness, fostered critical thinking, and reduced reliance on mentors. The supervising mentor noted a slightly lower workload than in earlier projects, while the mentor of the unsupervised group reported higher effort in guiding and refining outcomes. Overall, the findings suggest that guided LLM integration has the potential to improve learning, deepen critical engagement, foster independence, and reduce mentor workload when compared to those who do not provide structured guidance in LLM use.

**Keywords:** community-based learning; problem-based learning; education; LLMs; ChatGPT; Gemini; SDGs

## 1. Introduction

The theoretical landscape surrounding the contribution of digital transformation to sustainability issues is lacking in the literature [1]. This can be considered an issue, as societies are increasingly expected to rely on novel technologies in address modern challenges. In this context, the emergence of large language models (LLMs) introduces an additional layer of complexity to the ongoing discourse on digital transformation and sustainability. However, LLMs in sustainability education are still a contentious issue, dividing the academic community with arguments for and against their implementation [2]. On the negative side, the unsupervised and untrained use of LLMs raises ethical and privacy concerns [3], specifically in written assessments [4]. Furthermore, their threat to the cognitive development of critical thinking and the proclivity for artificial hallucinations [5]

undermine accuracy and dependability. If students are untrained in prompting frameworks (PFs), they may not receive appropriate responses from LLM chatbots, and there is a risk of skill degradation, particularly in the ability to construct well-crafted arguments [6]. Despite advances in emotional understanding, LLMs cannot provide the necessary level of empathy [7], one of the most important aspects of teacher–student interaction. Aside from that, professors and lecturers are still relatively unaware of the technology's capabilities, let alone how to use it properly or detect its use. The professional community assumed that two years after the LLM was introduced, students would learn how to use the technology; three years later, professors would become aware of when students are using it, and five years afterward, universities would decide how to respond [8] (Figure 1).
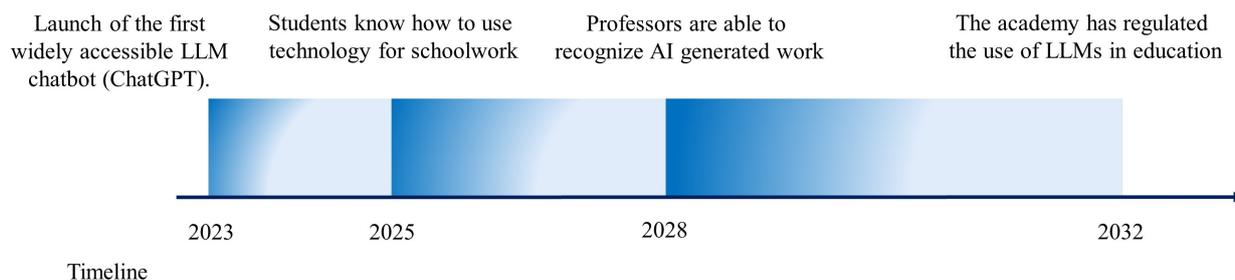


**Figure 1.** Projected timeline of LLM adoption and university response (derived from [8]).

This was a trigger for many to start adapting educational programs [9–13], emphasizing the need to teach students and educators how to effectively utilize the technology so they can program machines and not vice versa. On a positive side, a number of review studies show that LLMs can help with decision-making in medicine [14], law [15], computer coding [16], and other educational domains [17]. However, only a small number of papers have looked into the usability of LLMs in the context of sustainability (Section 2 of this study), and, to date, no research has been found that deals with bridging sustainability education with LLMs.

To fill the existing gap and introduce domain-specific novelty, this study investigates whether guided use of LLMs can improve students' critical thinking, work quality, and independence in projects related to the sustainable development of local communities. To quantify the findings, the study used descriptive and analytical approaches. The quantitative approach aimed to allow for comparisons of students' perceptions across groups and contexts.

The research also examines whether supervised LLM guidance reduces mentors' workload when compared to mentoring projects with unsupervised LLM use, through a qualitative approach. In addition to these findings, the study contributes to the academic community by proposing a project framework that enables professors to effectively leverage emerging technologies in sustainability projects within local communities, thereby enhancing university curricula and benefiting societies.

The remaining sections of this paper are organized as follows. Section 2 summarizes previous research on leveraging LLMs for sustainable development and highlights the research gap. Section 3 describes the materials and methods used, including (Section 3.1) the selection of LLMs for student projects, (Section 3.2) the selection of prompting frameworks and effective communication strategies, (Section 3.3) the project assignment on local community sustainability, and (Section 3.4) evaluation design. Section 4 presents the results and discussion, including (Section 4.1) assignment with LLM prompting frameworks, (Section 4.2) descriptive evaluation of supervised versus unsupervised LLM use, and (Section 4.3) inferential evaluation using logistic regression and post hoc power analysis.

Section 5 discusses the study's challenges and future research directions, while Section 6 concludes by discussing the implications for "digitainability".

## 2. Literature Review: Leveraging LLMs for Sustainable Development

The scientific and professional community has started looking into the advantages and disadvantages of novel technologies to assess LLMs' usability in various sustainability sectors. Thurzo et al. [18] claim that ChatGPT (gpt-3.5) can boost operational effectiveness by facilitating timely decisions. Rathore [19] investigated the potential use of LLMs in the textile sector for sustainable development, showing that technology can reduce waste production, enhance product quality, and support sustainability objectives. Similar findings were reached by Alves et al. [20], who confirmed that AI chatbots can support natural resource management decision-making. Parović [21] argues that LLMs can positively impact a just energy transition. In a basic building project, Prieto et al. [22] showed that GPT is capable of producing a logical construction schedule. The authors claim that the platform successfully completed the work scope. According to other studies, generative AI can help supply chains run more efficiently and enable intelligent traffic management systems [23]. The effects of climate change on particular populations can also be determined using AI-driven data [24]. For instance, Jungwirth and Haluza [25] point out that LLMs may help tackle social megatrends. Recent research has also looked into LLMs in domestic assistants and LCA workflows, knowledge-building discourse, e-commerce product claims, classroom programming support, and sustainability reporting (Table 1) [26–34].

**Table 1.** Overview of LLM's usability in sustainable development studies.

| Ref. | Country | LLM | Aim | Findings |
|------|---------|-----|-----|----------|
| [26] | Italy, Belgium | ChatGPT, BingAI, Bard, LLaMA | to compare the generative capabilities of four large language models in ecological sustainability | ChatGPT is a promising LLM for embedding into home assistants to foster sustainable behaviors in domestic environments. |
| [27] | USA, S. Arabia | Gemini, GPT-4 | to explore the opportunities, challenges, and risks associated with using LLMs in various life cycle assessment tasks | LLMs can enhance efficiency and accessibility but pose challenges such as hallucinations, data reliability issues, and the need for responsible oversight |
| [28] | Italy | GPT-3.5, GPT-4, Claude 1, Claude 2, Bard, LLaMA, LLaMA 2 | to propose a pedagogical and technological framework for using LLMs to address sustainability issues in assessment and feedback in higher education | LLMs enhance assessment in higher education by automating evaluation, providing personalized feedback, reducing workload, improving objectivity, and enabling continuous learning. |
| [29] | Greece | GPT-3.5, GPT-4 | to automate the identification of sustainable product attributes from product descriptions, titles, and product specifications. | GPT effectively identifies sustainable product features, supports eco-friendly purchasing, but requires improved response formatting and consistency |
| [30] | Singapore | ChatGPT /GPT-3.5 | to explore how GPT can be utilized in designs that support sustainable student discourse and knowledge creation. | Generative AI enhances student discourse but requires oversight due to potential inaccuracies, biases, and ethical concerns in educational use |

**Table 1.** *Cont.*

| Ref. | Country | LLM | Aim | Findings |
|------|---------|-----|-----|----------|
| [31] | Brazil | GPT-3.5 | to examine how the integration of ChatGPT in coding and programming courses impacts student perceptions of educational support, sustainability, and individual learning experiences. | The study finds that over 90% of students value ChatGPT in programming courses, 70% emphasize teacher interaction, and 29.7% worry about generative AI overuse in assessments |
| [32] | Qatar, Canada | ChatGPT /GPT-4 | to examine the integration of generative AI tools in education in the context of education for sustainable development (ESD). | Generative AI enhances personalized learning and collaboration in ESD but raises ethical concerns, digital divide issues, and risks of overreliance |
| [33] | S. Africa, N. Zealand, UK | ChatGPT | to analyze the benefits and risks of using generative AI tools for sustainability reporting. | Generative AI improves sustainability reporting efficiency but risks greenwashing, reliability issues, and transparency concerns, requiring governance, assurance, and ethical oversight |
| [34] | UAE | ChatGPT | to discuss the potential impact of generative AI on education and the need for a sustainable approach to its implementation. | Generative AI enhances personalized learning, tutoring, and assessment but raises concerns about privacy, bias, academic integrity, and teacher–student dynamics. |

In summary, this research suggests that LLMs can improve efficiency (e.g., faster collection and interpretation of life cycle inventory) [27], personalize learning and feedback [28,31,32], and provide easier access to high-quality guidance [32,34]. However, they also recognize persistent threats of hallucinations and bias [27,33,34], weak source verifiability [27,29], greenwashing in narratives [33], issues over privacy and fairness [32,34], and student overreliance [31,34]. Holmes et al. [35] see these prior opportunities as a threat to humanity, as AI chatbots may not always reflect the values of society as a whole. Along these lines, Hartmann et al. [36] provided converging evidence on ChatGPT's pro-environmental, left-libertarian orientation, while Subaveerapandiyan et al. [37] indicate that LLMs should aid decisions rather than generate ideas. Nevertheless, the emerging opinion is to adopt rather than ban LLMs on strict terms: retrieval-augmented prompting and transparent citation [27], human-in-the-loop review and board/faculty AI literacy [27,33,34], integrity by design evaluation and oversight [28,31,34], robust data governance and independent assurance for reporting [33], and continuous capacity building in accordance with education for sustainable development goals [32]. Henceforth, students and the general public should be properly trained to utilize novel technology, maximizing its benefits while avoiding misuse.

## 3. Materials and Methods

To enhance clarity, the Materials and Methods Section is organized into the following subsections: Section 3.1. *LLM Chatbots for Sustainable Development Projects*, which provides a brief comparison of GPT and Gemini, i.e., a summary of existing assessments that informed the selection of LLMs for the student tasks; Section 3.2. *LLM Frameworks—Effective Communication Strategies for Interacting with LLM Chatbots*, describing the methodology employed in choosing the most effective LLM prompting frameworks; Section 3.3. *The Project Assignment—Students Engaging with Local Community Sustainability*, presenting the

organization of the student project this research refers to; and Section 3.4. *Evaluation of LLM Effectiveness in Student Sustainability Projects*, explaining the procedure used to assess how guided LLM use influenced students' critical thinking, work quality, and independence.

### 3.1. LLM Chatbots for Sustainable Development Projects

To identify the LLM that would provide the most appropriate answer to students' inquiries, this study relied on existing assessments in the literature. Specifically, the study adopts the methodological description and comparative framework used by Wangsa et al. [38] to compare two major AI research organizations, OpenAI (ChatGPT) and Google DeepMind (Gemini). A summary of the key strengths and limitations of the two LLMs is presented in Table 2. Based on their comparative advantages, this study recommends one chatbot over the other in dealing with specific project tasks, prioritizing Gemini for tasks requiring up-to-date online searches and ChatGPT for tasks focusing on academic writing.

**Table 2.** Comparative analysis overview of chatbot models (as refined from [38]).

| Criteria | ChatGPT | Gemini |
|---|---|---|
| Strengths and Opportunities | Multiple choice questions, academic text, advanced in academic and practical tasks | Non-original problems online, interactive feedback, real-time internet access |
| Weaknesses and Challenges | Inaccurate information, consistency issues, academic concerns, lack of contextual awareness, safety concerns | Generates different answers, consistency issues, original scientific contribution |
| Performance Metrics [1] | High | Medium |
| Scope of Knowledge [2] | High | High (Real-time) |
| Use Cases Excelling In | Academic research, trivia, education | Education, real-time queries |
| Use Cases Limited By | Scientific contributions | Detailed scientific writing, complex reasoning tasks |
| Knowledge Update Frequency | High (Quarterly updates) | High (Real-time) |

[1] Performance Metrics reflect a subjective synthesis based on the performance and behaviors of the models as reported in the literature [38]. [2] The Scope of Knowledge is an interpretive assessment of the range of topics and the depth of discourse that each model demonstrates across various domains.

### 3.2. LLM Frameworks—Effective Communication Strategies for Interacting with LLM Chatbots

To fully harness the potential of LLMs, one must properly formulate a question and provide relevant details for accurate and meaningful responses. While individual prompting techniques (e.g., zero-shot (zero-shot learning aims to solve unseen tasks without labeled training examples [39]), role prompting, chain-of-thoughts (chain-of-thought prompting enables complex reasoning capabilities through intermediate reasoning steps [40]), etc.) help refine certain aspects of interactions, only a well-designed PFs (combining multiple prompting techniques) ensures a systematic and scalable approach. Having this in mind, the five authors of this study, serving as the project mentors, reviewed PFs established in the professional community [41], leveraging knowledge transfer to identify those most suitable for the project tasks. Among the nine PFs analyzed [41], the main differences concerned the order of instructions and the level of detail requested in LLM responses. Accordingly, the mentors selected PFs whose instruction order aligned with task expectations, differing primarily in the degree of detail the LLM was expected to provide. The mentors then assessed the quality of LLM responses produced under each PF on the project tasks. PF selection (Table 3) and task-to-PF matching (Table 4) were based on interrater reliability.

(Inter-rater reliability is the extent to which two or more raters (or observers, coders, examiners) agree. It addresses the issue of consistency of the implementation of a rating system [42]). Each of the five mentors (a) rated the suitability of each PF's structure [41] on a 1–9 scale (9 = highest) and (b) evaluated the quality of LLM responses to example questions for PFs provided by the coordinating mentor on a 1–2 scale. For the former, the study selected the two best-ranked PFs (Table 3), while for the latter, the study selected the PF with the better rank corresponding to the project task.

**Table 3.** Selected PFs, their meaning, and the activities they refer to.

| PFs | Meaning | Activity |
| --- | --- | --- |
| **R.A.C.E.** | Role | Specify the role of the LLM chatbot. |
| | Action | Detail what action is needed. |
| | Context | Provide relevant details of the situation. |
| | Expectation | Describe the expected outcome. |
| **R.O.S.E.S.** | Role | Specify the role of the LLM chatbot. |
| | Objective | State the goal or aim. |
| | Scenario | Describe the situation. |
| | Expected Solution | Define the desired outcome. |
| | Steps | Ask for the actions needed to reach the solution. |

**Table 4.** Project assignment with recommended LLMs and frameworks for task completion. Project topic: contributing the sustainable development of local communities—example of (challenge/problem) in (location community).

| Task | LLM | PF |
| --- | --- | --- |
| **#1 INTRODUCTION** | | |
| • Describe the activity or action that leads to the sustainability challenge the team aims to overcome/improve. | Original work | Original work |
| • The introduction should include general information on social activities that hinder the sustainable development of society. | ChatGPT | R.A.C.E. |
| • Study and cite available literature dealing with similar issues. | Gemini | R.A.C.E. |
| **#2 RESEARCH OBJECT** | | |
| • Present a specific local community issue that hinders sustainable development. | Original work | Original work |
| • Describe the geographical location. | Original work | Original work |
| • Take photographs of the problem or its consequences. | Original work | Original work |
| • Explain the impact of the local issue on the environment, public health, and overall quality of life. | Gemini | R.O.S.E.S. |
| • Assess the financial implications: Who bears the cost? How much? Study and cite available literature dealing with similar issues | Gemini | R.A.C.E. |
| **#3 LONG-TERM EFFECTS** | | |
| • What consequences would your neighbourhood face if the problem you are dealing with is not addressed? Examine the impact trend in 5, 10, and 20 years (consider CO2 emissions, impact on health, spatial increase, increase in mass, etc). | ChatGPT | R.A.C.E./ Original work |

**Table 4.** *Cont.*

| Task | LLM | PF |
|---|---|---|
| **#4 IMPACT OF THE PROBLEM ON SDGs** | | |
| • How will addressing the local problem/challenge from task #1 impact the SDGs? | ChatGPT | R.O.S.E.S. |
| **#5 EXAMPLES OF GOOD PRACTICES** | | |
| • Identify two locations where a similar problem has been resolved or does not exist. If resolved, explain how. If it does not exist, discuss why. Study and cite available literature dealing with similar issues. | Gemini | R.O.S.E.S. |
| **#6 PROPOSED SOLUTION** | | |
| • How can the problem be solved? | ChatGPT | R.A.C.E. |
| • Is it possible to apply existing solutions from task #5? | ChatGPT | R.O.S.E.S. |
| • Often, existing solutions need further adaptation—explain why. | ChatGPT | R.A.C.E. |
| • If local community organizations do not function as expected, analyze why and suggest ways to encourage change. | Gemini | R.O.S.E.S. |
| • Are such problems an obstacle or an opportunity for private entrepreneurship and new workplaces? | ChatGPT | R.A.C.E. |
| **#7 ASSESSMENT OF IMPROVEMENT** | | |
| • How does the team expect their solution to contribute to the sustainable development of society? | ChatGPT | R.O.S.E.S. |
| • Will it reduce pollutant emissions and resource consumption? | ChatGPT | R.A.C.E. |
| • Will it save time, energy, and money? | ChatGPT | R.A.C.E. |
| • Will it create new jobs? | ChatGPT | R.A.C.E. |
| • Will it improve public health and overall quality of life? | Gemini | R.A.C.E. |

### 3.3. The Project Assignment—Students Engaging with Local Community Sustainability

This research was conducted within a course on sustainable development in first-year engineering studies. In addition to lectures, the course consists of a project requirement that allows students to learn by working on sustainable development in their communities. The project was designed to be a group project of three to four students led by a mentor. The aim was for students to identify environmental or energy challenges in their local community and propose solutions for their improvement. As recommended by the European School Education Platform [43], the project tasks were written to follow the principle of the 4 Cs: Critical thinking, Creativity, Communication, and Collaboration. In brief, the project required students to (1) define the problem/challenge (quantify its impact on society or the environment), (2) analyze trends (explain the need for action), (3) determine the impact of the problem on SDGs (link local initiatives to global challenges), (4) learn from successful solutions, (5) tailor existing solutions to the local challenge, (6) outline problem-solving steps, and (7) cCalculate energy and $CO_2$ savings. As noted by project mentors, the bottleneck in task completion was students' inexperience, as not all the students knew who the stakeholders could be, what consequences the problem/challenge causes, where to start the action, or how. In addition, this was exacerbated by a lack of structured, step-by-step

reasoning. To address these issues, this study proposes the integration of LLM chatbots and proper LLM frameworks into the workflow.

### 3.4. Evaluation of LLM Effectiveness in Student Sustainability Projects

To evaluate the proposed framework, a pilot study was conducted with two independent student groups. One group (collectively referred to as Mentored Group A (MGA)) was guided through the assignment using the new hybrid framework with supervised application of LLMs, while the other (Mentored Group B (MGB)) completed the same tasks under a traditional mentoring approach without structured LLM guidance. Both cohorts (MGA and MGB) consisted of students from the same year, enrolled in the same mandatory course. Mentors worked together to align project requirements, reducing variability in expectations. At the end of the project, both groups completed a survey questionnaire (Supplementary Material S1) designed to capture their perceptions of usability, learning outcomes, and critical thinking development. The pilot study design thus enabled a controlled comparison between test and control conditions. The comparison involved descriptive and inferential approaches. The former (Section 4.2) used box-and-whisker plots, showing quartiles, medians, and average values of survey ratings. The latter (Section 4.3) involved univariate and multivariate logistic regressions to assess between-group differences (test group/control group), reporting $p$-values, odds ratios, and confidence intervals to determine statistical significance. (An odds ratio (OR) is a statistic that gauges how closely two events, A and B, are related. If OR is equal to 1, then the two occurrences are considered independent. If the OR is greater than 1, A and B are associated (correlated) in that having B increases the odds of having A compared to not having B, and the opposite is true. If the OR is less than 1, then A and B are negatively associated, meaning having one event lowers the likelihood of having the other [44]). To determine whether the research with this sample size has sufficient statistical power, the study included a post hoc analysis using G*Power (version number 3.1.9.7) [41]. The analysis is based on the observed effect size (Cohen's d) and the actual sample size under the conditions of a two-tailed paired $t$-test with a significance level ($\alpha$) set at 0.05.

## 4. Results and Discussion

To enhance clarity, the Results and Discussion Section is organized into the following subsections: Section 4.1. *The Project Assignment with LLM Prompting Frameworks*, presenting the assignment proposed in this study; Section 4.2. *Descriptive Evaluation of Student Responses: A Pilot Study of Supervised and Unsupervised LLM Use*, and Section 4.3. *Inferential Evaluation of Student Responses: A Pilot Study of Supervised and Unsupervised LLM Use*, comparing the experiences of students and lecturers involved in the assignment with supervised LLM use with those who completed the same assignment tasks with unsupervised LLM use.

### 4.1. The Project Assignment with LLM Prompting Frameworks

By using the existing project coursework (Section 3.2), frameworks introduced by the professional community (Table 3), and insights into the strengths and weaknesses of the two prevalent LLMs (Table 2), this study proposes a project assignment focused on sustainable community development (Table 4). When recommending specific LLMs, the mentors followed benchmarks reported in [38], giving preference to Gemini for tasks requiring up-to-date online searches, while ChatGPT was adopted as the better option for academic writing. Based on the mentors' interrater assessments, two selected prompts (Table 3) follow a structured order: role first, action or objective second, context or scenario third,

and expectation/expected solution fourth, with the only distinction being the requirement of "steps" in the more detailed PF.

Students were encouraged to respond to all project tasks by completing or sharing the bullet subtasks. Table 4 contains PFs that students could use in responding to the subtasks, while some of the subtasks require no other work than the original team expressions. Supplementary Material S2 provides more details on organizing a potential PF. Following the chatbot assistance, team members should refine the chatbot responses and organize them into (1) an appropriate essay chapter and (2) bullet points for a PowerPoint presentation. Teams should be prepared to defend their ideas during a multi-team meeting moderated by a mentor. In this session, opposing team members will take on the role of the *advocatus diaboli*, challenging the presented ideas with counterarguments. These briefings should not influence the negative points awarded to the team defending, but they may provide extra points to the team presenting counterarguments. The mentor will determine and discuss the essay structure with all teams, ensuring that arguments are heard and debated multiple times. Each task should be reviewed in biweekly briefings before the final oral defense, where the entire project will be presented in front of teams led by different mentors. The number of teams in a defense session should equal the number of mentors leading the teams (and moderating the session), with one team from each mentor. During the oral defense, the defending team will face off against other groups led by different mentors, students with whom they have not previously interacted. This setup increases the likelihood of encountering new counterarguments, which requires them to think critically and respond effectively in real time. As a result, this should help to develop critical thinking skills and may indicate that defending an idea and critically observing all of its consequences in advance is more important than just proposing one.

*4.2. Descriptive Evaluation of Student Responses: A Pilot Study of Supervised and Unsupervised LLM Use*

The results presented in this section are based on the survey (Supplementary Material S1). For comparison, the survey considered two independent groups: MGA (six teams, 23 students) serving as a test group, and MGB (five teams, 17 students), serving as the control group. An analysis of the survey responses revealed that the majority of students from both groups had used LLMs on previous and during this assignment (85%, MGA; 94%, MGB); however, none were familiar with the prompting frameworks.

Regarding the other responses (Figure 2), the questionnaire revealed distinct variations between students who were supervised using LLMs and those who were not. According to the mean scores (M) of Likert-scale ratings, the MGA perceived that their project framework prepared them better for future tasks using LLMs (M = 3.71) than MGB (M = 3.07). Moreover, MGA provided 19% stronger affirming responses (fours and fives), compared to MGB (40%), shifting the median answer to 4 for MGA and 3 for MGB. The highest discrepancy between the two group responses was noticed when answering to what extent LLMs encouraged critical thinking (MGA (M = 3.53), MGB (M = 2.73)). This was to be expected, given that MGA had a mentor who was prepared to critically discuss the use of LLMs in the assignment, as opposed to MGB's traditional mentoring approach, which focused on the students' work per se. Following the disparity in the mentoring approaches, students from the MGA group were less confused by the answers received from the LLMs (M = 2.76) than those from the MGB (M = 3.13). Moreover, the response deviation for MGA was 35% smaller than for MGB, indicating that the first group's responses were more consistent, with more homogeneous opinions. This could be the result of critical mentoring approaches in MGA, especially given that MGA students consider themselves better prepared for future assignments than the MGB students (response to question 3, Supplementary Material S1). When assessing the time saved by using LLMs to perform the

tasks, the responses were reasonably similar (MGA, M = 3.38; MGB, M = 3.20), as was the evaluation of the importance of the mentor in performing the task (MGA, M = 4.06; MGB, M = 4.13). When assessing the potential for LLM chatbots to replace mentors in the future, there was a difference in the two groups' responses, with MGB's median answer being 1 and MGA's being 2. This aligns with the response outcomes of the second question of the survey (Figure 2), particularly keeping in mind that in MGA, 29% of responses were three or above, while in MGB, only 7% were. This suggests that students trained with effective communication frameworks for LLMs perceive themselves as more capable of conducting independent research than their counterparts who lack such structured guidance.
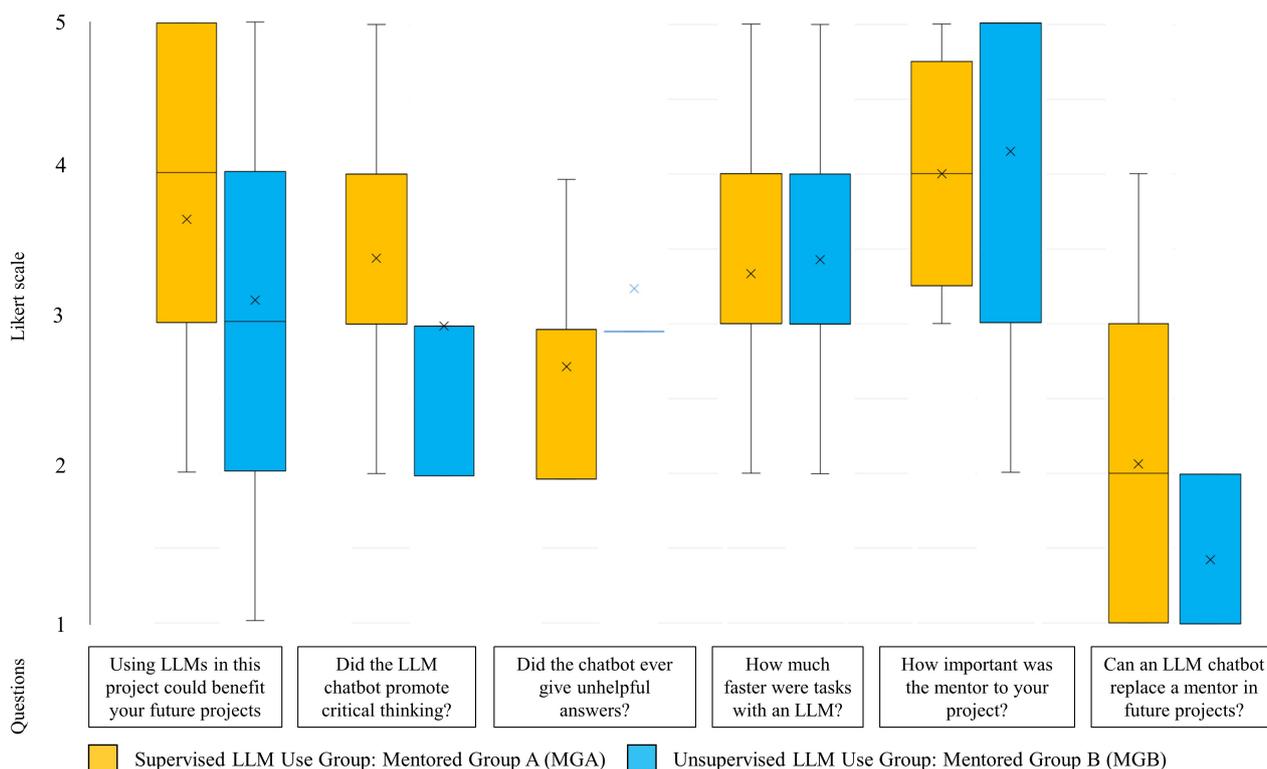


**Figure 2.** Likert scale distribution of student responses to the questionnaire (Supplementary Material S1).

In terms of overall student work grade, both of the mentored groups provided similar results, with an average score (on a scale of 0 to 30, with 30 being the highest grade) of 25 for the MGA and 24.5 for the MGB. Members of the grading commission evaluated MGA teams' project work as organized and concise. The students were well prepared for the defense, and the use of LLMs was more noticeable in MGB's work than in MGA's work. The MGA mentor reported no bottlenecks, as seen in previous generations. The main challenge this time was explaining the frameworks (Table 3) and designing the appropriate prompts (Supplementary Material S2). On the other hand, the MGB mentor reported obvious use of LLMs, as the initial work was too broad and not concise. In addition to the usual bottlenecks, supervision of MGB required further efforts to make presentations and papers more concise. To sum up, the MGA mentor described the introductory lectures as the most challenging, describing the efforts on the assignment as somewhat easier than before. By contrast, the MGB mentor noted increased efforts to shape the work properly, describing it as somewhat more challenging than with previous generations.

*4.3. Inferential Evaluation of Student Responses: A Pilot Study of Supervised and Unsupervised LLM Use*

Based on the post hoc power analysis using G*Power [39] to evaluate the sample size (n = 32) and observed effect size (Cohen's d) for a two-tailed paired *t*-test at $\alpha = 0.05$, the computation revealed a statistical power of 0.8213, indicating that the research had an 82.13% probability of correctly detecting the large effect, which exceeds the typically accepted threshold of 80% [45].

Furthermore, to validate the findings from Section 4.2, the study performed univariate and multivariate analyses in R software (version number 4.5.2), comparing student responses to questionnaire items (Qs) between the two independent groups (Table 5). The analysis indicated that LLM use matters to participants' experience. The strongest finding was for perceived critical-thinking support: agreeing that the LLM promoted critical thinking (Q2) had an independent association with lower odds (OR) for the outcome, indicating a significant and statistically robust effect even after adjustment. Two other findings were encouraging as well. Reporting faster completion of tasks with an LLM (Q4) had higher odds of the outcome in univariate analysis and, upon adjustment, was directional as well, potentially indicating a real effect to be confirmed. Reporting the hope that an LLM could replace a mentor (Q6) was significant univariately ($p = 0.048$), but, upon adjustment, barely missed significance ($p = 0.054$). The other measures (Q1, Q3, and Q5) were insignificant (all $p > 0.05$), and several of the confidence intervals (CI) were very wide, indicating that at this sample size, precision was poor.

**Table 5.** Results of univariate and multivariate logistic regression analyses for questionnaire items.

| Variable | Questions | Univariate Analysis | | | Multivariate Analysis | | |
|---|---|---|---|---|---|---|---|
| | | **OR** | **95%CI** | ***p*-Value** | **OR** | **95%CI** | ***p*-Value** |
| Q1 | Using LLMs in this project could benefit your future projects? | 0.670 | 0.199–2.251 | 0.517 | | | |
| Q2 | Did the LLM chatbot promote critical thinking? | 0.137 | 0.022–0.862 | 0.034 | 0.224 | 0.058–0.860 | 0.029 |
| Q3 | Did the chatbot ever give unhelpful answers? | 3.657 | 0.627–21.548 | 0.149 | | | |
| Q4 | How much faster were tasks with an LLM? | 14.440 | 1.180–176.630 | 0.037 | 3.876 | 0.881–17.057 | 0.073 |
| Q5 | How important was the mentor to your project? | 0.301 | 0.056–1.616 | 0.161 | | | |
| Q6 | Can an LLM chatbot replace a mentor in future projects? | 0.155 | 0.025–0.984 | 0.048 | 0.278 | 0.072–1.020 | 0.054 |

Overall, these pilot data reveal LLMs' promise for supporting critical thinking and efficiency, justify a larger, properly powered investigation, and suggest critical thinking promotion, task effectiveness, and mentor importance as the highest research priorities in the future.

## 5. Challenges and Future Research Directions

This study was conducted entirely online during a national faculty blockade (2024/2025) caused by a student-wide protest in Serbia [46], which may have influenced student engagement and collaboration compared to traditional in-person settings. This

context may have influenced student participation, communication, and motivation by reducing the level of direct interaction while increasing solidarity and shared purpose. The study is a limited pilot that serves as a proof of concept for incorporating LLM-supported, project-based work on local sustainability challenges. As a limited pilot, this study did not include qualitative feedback from students and instead relied on quantitative survey metrics to ensure comparability.

Future research, building on the findings presented here, will be carried out as part of a bilateral project between two Central European universities, involving larger and more diverse cohorts of students and instructors to validate and generalize the findings across multicultural contexts. In addition to this existing research, this should result in better settings for examining small effects. The research will examine specific questions about the effects of supervised and unsupervised LLM use on student learning outcomes across both cultural and disciplinary (engineering and environmental protection) contexts. The design will be strengthened by using team- or section-level randomization to compare supervised versus unsupervised LLM use, pre/post measures of learning and AI literacy, and blinded, rubric-based grading with inter-rater reliability tests.

## 6. Conclusions

This study proposed and tested a group project assignment using LLM chatbots to address sustainability issues in local communities. Faculty members involved in the course had identified key difficulties in previous assignments, recognizing that effective use of LLMs could foster improved problem-solving approaches and allow for faster responses. In this regard, the approach focuses on teaching students how to communicate with LLMs through appropriate prompting frameworks. Simultaneously, to mitigate potential drawbacks such as decreased critical thinking and emotional depth, this framework proposes biweekly review sessions to track student progress and ensure meaningful engagement. Mentors who oversee the groups should also facilitate debates between different teams, encouraging critical thinking and students' ability to defend the presented points of view.

The pilot involved eleven student teams (3 to 4 students each). Six teams were guided by a mentor trained for this purpose, and their experiences were compared at a high level to a cohort using LLMs without supervision. The supervised cohort avoided the recurring design bottlenecks seen in previous generations, but it required initial instruction in prompt engineering. Students reported that the frameworks would be useful in future projects, that LLMs facilitated critical thinking, and that the mentor's role was less important than that of a cohort involved in traditional processes. The supervising mentor described his overall effort as being slightly lower than in previous generations, whereas the mentor of the unsupervised cohort reported greater effort to shape and refine the project presentation.

Taken together, these findings suggest that strategically incorporating LLM chatbots into a project framework can help students become more self-sufficient and better prepared to face emerging societal challenges. Furthermore, mentoring students who have been trained to use LLMs properly would take less effort than mentoring those who use the technology incorrectly. More broadly, the work is consistent with the concept of "digitainability"—the intersection of digital technology and sustainability—as a catalyst for accelerating progress toward the SDGs.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su172210148/s1, Supplementary Materials S1: Student Feedback on Chatbot Support in Project Development; Supplementary Materials S2: Table S1: Potential PF for the project assignment tasks presented in Table 4. Project topic: Contributing sustainable development of local communities—example of (challenge/problem) in (location community).

## Abbreviations/Nomenclature

The following abbreviations/discipline nomenclature are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| CI | Confidence Interval |
| LLM | Large Language Models |
| OR | Odds Ratio |
| PF | Prompting Framework |
| SDGs | Sustainable Development Goals |
| R.A.C.E. | Role, Action, Context, Expectation |
| R.O.S.E.S. | Role, Objective, Scenario, Expected Solution, Steps |

## References

1. Shaya, N.; AbuKhait, R.; Madani, R.; Ahmed, V. Conceptualizing Blended Learning Models as a Sustainable and Inclusive Educational Approach: An Organizational Dynamics Perspective. *Int. J. Sustain. High. Educ.* **2025**, *26*, 90–111. [CrossRef]
2. Meyer, J.G.; Urbanowicz, R.J.; Martin, P.C.N.; O'Connor, K.; Li, R.; Peng, P.C.; Bright, T.J.; Tatonetti, N.; Won, K.J.; Gonzalez-Hernandez, G.; et al. ChatGPT and Large Language Models in Academia: Opportunities and Challenges. *BioData Min.* **2023**, *16*, 20. [CrossRef]
3. Kontche Steve, M. Ethical Considerations for Companies Implementing LLMs in Education Software. *Int. J. Innov. Sci. Res. Technol.* **2024**, *9*, 1856–1861. [CrossRef]
4. März, M.; Himmelbauer, M.; Boldt, K.; Oksche, A. Legal Aspects of Generative Artificial Intelligence and Large Language Models in Examinations and Theses. *GMS J. Med. Educ.* **2024**, *41*, Doc47. [CrossRef]
5. Perković, G.; Drobnjak, A.; Botički, I. Hallucinations in LLMs: Understanding and Addressing Challenges. In Proceedings of the 2024 47th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 20–24 May 2024; pp. 2084–2088.
6. Nayak, P.; Gogtay, N.J. Large Language Models and the Future of Academic Writing. *J. Postgrad. Med.* **2024**, *70*, 67–68. [CrossRef] [PubMed]
7. Sorin, V.; Brin, D.; Barash, Y.; Konen, E.; Charney, A.; Nadkarni, G.; Klang, E. Large Language Models and Empathy: Systematic Review. *J. Med. Internet Res.* **2024**, *26*, e52597. [CrossRef] [PubMed]
8. Marche, S. Will ChatGPT Kill the Student Essay? The Atlantic. Available online: https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/ (accessed on 6 February 2025).
9. Ravi, A.; Neinstein, A.; Murray, S.G. Large Language Models and Medical Education: Preparing for a Rapid Transformation in How Trainees Will Learn to Be Doctors. *ATS Sch.* **2023**, *4*, 282–292. [CrossRef]
10. Peláez-Sánchez, I.C.; Velarde-Camaqui, D.; Glasserman-Morales, L.D. The Impact of Large Language Models on Higher Education: Exploring the Connection between AI and Education 4.0. *Front. Educ.* **2024**, *9*, 1392091. [CrossRef]
11. Diab Idris, M.; Feng, X.; Dyo, V. Revolutionizing Higher Education: Unleashing the Potential of Large Language Models for Strategic Transformation. *IEEE Access* **2024**, *12*, 67738–67757. [CrossRef]

12. Tsai, M.L.; Ong, C.W.; Chen, C.L. Exploring the Use of Large Language Models (LLMs) in Chemical Engineering Education: Building Core Course Problem Models with Chat-GPT. *Educ. Chem. Eng.* **2023**, *44*, 71–95. [CrossRef]

13. Bonner, E.; Lege, R.; Frazier, E. Large Language Model-Based Artificial Intelligence in the Language Classroom: Practical Ideas for Teaching. *Teach. Engl. Technol.* **2023**, *2023*, 23–41. [CrossRef]

14. Vrdoljak, J.; Boban, Z.; Vilović, M.; Kumrić, M.; Božić, J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. *Healthcare* **2025**, *13*, 603. [CrossRef]

15. Siino, M.; Falco, M.; Croce, D.; Rosso, P. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access* **2025**, *13*, 18253–18276. [CrossRef]

16. Pereira, A.F.; Ferreira Mello, R. A Systematic Literature Review on Large Language Models Applications in Computer Programming Teaching Evaluation Process. *IEEE Access* **2025**, *13*, 113449–113460. [CrossRef]

17. Shahzad, T.; Mazhar, T.; Tariq, M.U.; Ahmad, W.; Ouahada, K.; Hamam, H. *A Comprehensive Review of Large Language Models: Issues and Solutions in Learning Environments*; Springer International Publishing: Berlin/Heidelberg, Germany, 2025; Volume 6, ISBN 0123456789.

18. Thurzo, A.; Strunga, M.; Urban, R.; Surovková, J.; Afrashtehfar, K.I. Impact of Artificial Intelligence on Dental Education: A Review and Guide for Curriculum Update. *Educ. Sci.* **2023**, *13*, 150. [CrossRef]

19. Rathore, B. Future of Textile: Sustainable Manufacturing & Prediction via ChatGPT. *Eduzone Int. Peer Rev. Acad. Multidiscip. J.* **2023**, *12*, 52–62.

20. Alves, B.C.; Freitas, L.A.d.; Aguiar, M.S.d. Chatbot as Support to Decision-Making in the Context of Natural Resource Management. In Proceedings of the Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais, Maceió, Brazil, 20–24 July 2021; pp. 29–38.

21. Parović, M. Could Artificial Intelligence (AI) Contribute to a Just Energy Transition? *Energ. Ekon. Ekol.* **2024**, *26*, 25–30. [CrossRef]

22. Prieto, S.A.; Mengiste, E.T.; Soto, B.G. de Investigating the Use of ChatGPT for the Scheduling of Construction Projects. *Buildings* **2023**, *13*, 857. [CrossRef]

23. Rathore, D.B. Future of AI & Generation Alpha: ChatGPT beyond Boundaries. *Eduzone Int. Peer Rev. Multidiscip. J.* **2023**, *12*, 63–68.

24. Jungwirth, D.; Haluza, D. Artificial Intelligence and Ten Societal Megatrends: An Exploratory Study Using GPT-3. *Systems* **2023**, *11*, 120. [CrossRef]

25. Jungwirth, D.; Haluza, D. Artificial Intelligence and the Sustainable Development Goals: An Exploratory Study in the Context of the Society Domain. *J. Softw. Eng. Appl.* **2023**, *16*, 91–112. [CrossRef]

26. Giudici, M.; Abbo, G.A.; Belotti, O.; Braccini, A.; Dubini, F.; Izzo, R.A.; Crovari, P.; Garzotto, F. Assessing LLMs Responses in the Field of Domestic Sustainability: An Exploratory Study. In Proceedings of the 2023 Third International Conference on Digital Data Processing (DDP), London, UK, 27–29 November 2023; pp. 42–48. [CrossRef]

27. Preuss, N.; Alshehri, A.S.; You, F. Large Language Models for Life Cycle Assessments: Opportunities, Challenges, and Risks. *J. Clean. Prod.* **2024**, *466*, 142824. [CrossRef]

28. Agostini, D.; Picasso, F. Large Language Models for Sustainable Assessment and Feedback in Higher Education. *Intell. Artif.* **2024**, *18*, 121–138. [CrossRef]

29. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Unveiling Sustainability in Ecommerce: GPT-Powered Software for Identifying Sustainable Product Features. *Sustainability* **2023**, *15*, 12015. [CrossRef]

30. Lee, A.V.Y.; Tan, S.C.; Teo, C.L. Designs and Practices Using Generative AI for Sustainable Student Discourse and Knowledge Creation. *Smart Learn. Environ.* **2023**, *10*, 59. [CrossRef]

31. Silva, C.A.G.d.; Ramos, F.N.; de Moraes, R.V.; Santos, E.L. dos ChatGPT: Challenges and Benefits in Software Programming for Higher Education. *Sustainability* **2024**, *16*, 1245. [CrossRef]

32. Abulibdeh, A.; Zaidan, E.; Abulibdeh, R. Navigating the Confluence of Artificial Intelligence and Education for Sustainable Development in the Era of Industry 4.0: Challenges, Opportunities, and Ethical Dimensions. *J. Clean. Prod.* **2024**, *437*, 140527. [CrossRef]

33. de Villiers, C.; Dimes, R.; Molinari, M. How Will AI Text Generation and Processing Impact Sustainability Reporting? Critical Analysis, a Conceptual Framework and Avenues for Future Research. *Sustain. Account. Manag. Policy J.* **2024**, *15*, 96–118. [CrossRef]

34. Kamalov, F.; Santandreu Calonge, D.; Gurrib, I. New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. *Sustainability* **2023**, *15*, 12451. [CrossRef]

35. Holmes, W.; Porayska-Pomsta, K.; Holstein, K.; Sutherland, E.; Baker, T.; Shum, S.B.; Santos, O.C.; Rodrigo, M.T.; Cukurova, M.; Bittencourt, I.I.; et al. Ethics of AI in Education: Towards a Community-Wide Framework. *Int. J. Artif. Intell. Educ.* **2022**, *32*, 504–526. [CrossRef]

36. Boelaert, J.; Coavoux, S.; Ollion, É.; Petev, I.; Präg, P. Machine Bias. How Do Generative Language Models Answer Opinion Polls? *Sociol. Methods Res.* **2025**, *54*, 1156–1196. [CrossRef]

37. Subaveerapandiyan, A.; Vinoth, A.; Neelam, T. Netizens, Academicians, and Information Professionals' Opinions About AI with Special Reference to ChatGPT. *Libr. Philos. Pract.* **2023**, 1–15. Available online: https://digitalcommons.unl.edu/libphilprac/7596 (accessed on 23 September 2025).

38. Wangsa, K.; Karim, S.; Gide, E.; Elkhodr, M. A Systematic Review and Comprehensive Analysis of Pioneering AI Chatbot Models from Education to Healthcare: ChatGPT, Bard, Llama, Ernie and Grok. *Future Internet* **2024**, *16*, 219. [CrossRef]

39. Kosaraju, D. Zero-Shot Learning: Teaching AI to Understand the Unknown. *Int. J. Res. Rev.* **2021**, *8*, 482–487. [CrossRef]

40. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November 2022; Volume 35, pp. 1–14.

41. Kanika, B.K. 9 Frameworks to Master ChatGPT Prompt Engineering. Medium. Available online: https://medium.com/@KanikaBK/9-frameworks-to-master-chatgpt-prompt-engineering-e2fac983bc61 (accessed on 23 September 2025).

42. Lange, R.T. Inter-Rater Reliability. In *Encyclopedia of Clinical Neuropsychology*; Springer: New York, NY, USA, 2011; ISBN 978-0-387-79948-3.

43. European School Education Platform 4 C's: Communication, Collaboration, Creativity and Critical Thinking | European School Education Platform. Available online: https://school-education.ec.europa.eu/en/learn/courses/4-cs-communication-collaboration-creativity-and-critical-thinking (accessed on 24 February 2025).

44. Pant, A.B. (Ed.) Odds Ratio BT. In *Dictionary of Toxicology*; Springer Nature: Singapore, 2024; pp. 737–738. ISBN 978-981-99-9283-6.

45. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Lawrence Erlbaum Associates: New York, NY, USA, 1988; ISBN 0805802835.

46. Stojiljković, Z. Blockades and Strikes in Serbia. *SEER* **2024**, *27*, 283–284. [CrossRef]