sustainability

MDPI

Article

Custom Generative Artificial Intelligence Tutors in Action: An Experimental Evaluation of Prompt Strategies in STEM Education

Rok Gabrovšek and David Rihtaršič *

Faculty of Education, University of Ljubljana, Kardeljeva Ploščad 16, SI-1000 Ljubljana, Slovenia; rok.gabrovsek@pef.uni-lj.si

* Correspondence: david.rihtarsic@pef.uni-lj.si

Abstract

The integration of generative artificial intelligence, particularly large language models, into education presents opportunities for both personalised learning and pedagogical challenges. This study focuses on electrical engineering laboratory education. We developed a configurable prototype of a generative artificial intelligence powered tutoring tool, implemented it in an undergraduate electrical engineering laboratory course, and analysed 208 student–tutoring tool interactions using a mixed-methods approach that combined research team evaluation with learner feedback. The findings show that student prompts were predominantly procedural or factual, with limited conceptual or metacognitive engagement. Structured prompt styles produced clearer and more coherent responses and were rated the highest by students, while approaches aimed at fostering reasoning and reflection were valued mainly by the research team for their pedagogical depth. This contrast highlights a consistent preference–pedagogy gap, indicating the need to embed stronger instructional guidance into artificial intelligence tutoring. To bridge this gap, a promising direction is the development of pedagogically enriched AI tutors that integrate features such as adaptive prompting, hybrid strategy blending, and retrieval-augmented feedback to balance clarity, engagement, and depth. The results provide practical and conceptual value relevant to educators, developers, and researchers interested in artificial intelligence tutors that are both engaging and pedagogically sound. For educators, the study clarifies how students interact with tutors, helping align artificial intelligence use with instructional goals. For developers, it highlights the importance of designing systems that combine usability with educational value. For researchers, the findings identify directions for further study on how design choices in artificial intelligence tutoring affect learning processes and pedagogical alignment across STEM contexts. On a broader level, the study contributes to a more transparent, equitable, and sustainable integration of generative AI in education.

Keywords: generative artificial intelligence; generative AI tutors; prompting strategies; STEM education; pedagogical design

check for updates

Academic Editor: Jun (Justin) Li

Received: 26 September 2025
Revised: 22 October 2025
Accepted: 23 October 2025
Published: 25 October 2025

Citation: Gabrovšek, R.; Rihtaršič, D. Custom Generative Artificial Intelligence Tutors in Action: An Experimental Evaluation of Prompt Strategies in STEM Education. *Sustainability* **2025**, *17*, 9508. https://doi.org/10.3390/su17219508

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The integration of artificial intelligence in education has raised fundamental questions about its ability to replicate or complement human teaching. One area that has explored this question for many years is the field of intelligent tutoring systems (ITSs). Over the past few decades, ITSs have advanced personalised learning by offering structured, individual

guidance through expert-crafted content and rule-based logic, often achieving learning gains comparable to one-on-one human tutoring [1]. However, developing traditional ITSs remains highly resource intensive, typically requiring extensive collaboration and iterative testing [2,3].

Recent advances in generative artificial intelligence (GEN-AI) and large language models (LLMs) have opened new possibilities for tutoring tools that enable natural language interaction, rapid prototyping, and customisation based on prompts [4,5]. Two main approaches dominate current practice. The first includes general-purpose tools (e.g., ChatGPT), which offer accessible interactions but lack explicit pedagogical structuring unless guided thoughtfully [6–9]. The second approach involves custom-built systems based on the same technologies, but designed for specific contexts. These systems embed pedagogy more directly through instructional design and domain-specific guidance [4,5,10–16].

Despite rapid advances, evidence suggests that technical sophistication alone does not determine pedagogical effectiveness. Studies provide mixed findings regarding both learning outcomes and broader educational effects. Some report positive effects, such as improvements in language proficiency [5], programming skills, and self-regulation [9], or higher test performance with AI assistance [12,17]. However, others caution that unstructured or over-reliant use may hinder learning: for example, high school students who used GPT-4 extensively performed worse once access was removed, suggesting an over-reliance effect [18], and GEN-AI use has been correlated with lower exam scores—particularly among high-performing students [19,20]. These outcomes highlight that effective tutoring requires not only technical innovation, but also careful attention to pedagogy, human oversight, and user competence [20,21].

Recent research further supports this view, showing that learners' ability to benefit from AI tools depends strongly on their sense of agency, self-efficacy, and reflective regulation, with AI literacy development being closely linked to transformative agency rather than mere tool proficiency [22]. Current discussions increasingly highlight the importance of prompt literacy, understood as a component of digital literacy that helps learners formulate questions beyond superficial help, fostering deeper learning [23]. Prompt literacy is not merely a technical skill but is closely tied to broader literacies such as critical thinking, systems thinking, and ICT self-concept [24,25]. By encouraging experimentation, reflection, and iterative refinement, it fosters learner agency, helping students build competences such as self-efficacy, persistence, and strategic thinking, which are essential for sustainable, human-centred professions [26]. In this sense, developing prompt literacy aligns with broader educational goals of preparing learners for adaptive, future-oriented problem solving and contributes to sustainable education and professional practice [21,24–26].

In both previously mentioned approaches, prompt strategies play a central role. Prompt strategies refer to systematic ways of formulating prompts to guide AI outputs toward pedagogically meaningful responses [27]. They can be used directly by learners or embedded by educators and developers at the system level within custom-built tools. Several prompt strategies have shown high pedagogical potential. These include persona prompts, chain-of-thought reasoning, flipped questioning, template-based scaffolds, few-shot examples, retrieval-augmented generation (RAG), and game-like formats [27–37].

This dual perspective illustrates that GEN-AI is not only a matter of short-term learning gains, but also of cultivating long-term competencies and equitable access. In this sense, GEN-AI tutors can be aligned with the broader agenda of sustainable education: when thoughtfully designed, they enhance accessibility, support disadvantaged learners, and enable teachers to orchestrate more adaptive and transparent learning processes, which contribute to SDG 4 (Quality Education) and long-term societal goals [38–41].

However, realising these benefits requires a clearer understanding of how GEN-AI tutors actually function in authentic learning contexts. Sustainable use of such systems depends not only on technological accessibility, but also on pedagogically informed design choices that promote meaningful learning rather than superficial assistance. One key factor shaping this balance is the prompt itself and the way instructions, roles, and examples are formulated to guide AI behaviour. Understanding how different prompting strategies influence both the quality of AI responses and students' learning experience is therefore essential for developing tutoring systems that are pedagogically effective, equitable, and sustainable.

Existing research has rarely compared such strategies in a systematic way, especially in the context of science, technology, engineering and mathematics (STEM) education, where learners are often engaged in laboratory-based tasks that require them to combine different kinds of knowledge and skills. A few exceptions exist, including comparative work in mathematics education [42] and a recent systematic review [43,44], but overall empirical evaluations across diverse instructional settings remain scarce. If we return to the opening question of whether technology can ever provide the same level of support as a human tutor, it becomes evident that traditional ITSs have repeatedly demonstrated their pedagogical effectiveness. However, GEN-AI tutors still need to prove their capacity to achieve comparable and sustainable learning gains, which highlights a gap in the literature.

Therefore to address this gap, this study introduces configurable GEN-AI tutoring prototype that prioritises flexibility and teacher control, ensuring alignment with instructional objectives. Using this prototype, the study systematically evaluates how different prompt strategies affect the quality and reception of AI-generated responses in laboratory-based Electrical Engineering activities, serving as a representative example within the broader STEM learning environment. The research is guided by three questions:

- RQ1: What types of questions do students formulate when interacting with a GEN-AI tutor in STEM laboratory activities?
- RQ2: How do responses generated under different prompting strategies differ in their instructional characteristics, based on authors' analysis?
- RQ3: How do students perceive and rank responses produced by different prompting strategies, and to what extent are these preferences consistent across different task contexts?

This research is part of broader national and institutional efforts to explore the educational potential of generative artificial intelligence and its implications for sustainable, learner-centred innovation. It directly contributes to several ongoing projects. Within the GEN-UI initiative [45], the study informs the development of practical guidelines and sample pedagogical scenarios for the meaningful and responsible use of GEN-AI in Slovenian schools. In parallel, it supports the goals of the 21st Century Skills project [46], particularly in designing and evaluating innovative teaching models that combine sustainability, digital literacy, and pedagogical effectiveness. Finally, the research aligns with the long-term objectives of the University of Ljubljana, Faculty of Education's Education for Sustainable Development programme, which emphasises transformative learning, open innovation, and the strategic integration of AI-driven educational tools within STEM education. Through this alignment, the study aims not only to examine student–AI interaction, but also to provide insights for teacher education in the context of sustainable, digitally enriched pedagogy, building on recent frameworks that integrate digital and sustainability competencies in education [47].

2. Materials and Methods

This study employed a mixed-methods sequential design in the form of a small-scale pilot study, centred on a configurable prototype of a GEN-AI tutoring tool developed

for this purpose. The study aimed to explore three research questions: the nature of student questions (RQ1), the instructional quality of AI-generated responses under different prompt strategies (RQ2), and students' perceptions of tutor responses (RQ3). The approach combined exploratory analysis of student–AI interactions with a design-oriented evaluation of pedagogical prompt strategies, aiming to support the integration of GEN-AI in STEM education, in line with SDG 4: Quality Education. As a pilot study, the primary goal was not to generalise results to wider populations, but rather to test the prototype in a controlled laboratory setting and to generate insights that inform further tool development and provide a foundation for future, larger-scale studies.

2.1. Research Design

This study employed a mixed-methods sequential design in the form of a small-scale pilot study centred on a configurable prototype of a GEN-AI tutoring tool developed for this purpose. The design progressed through three phases: an initial exploratory phase that informed the development of a benchmark question set, followed by two complementary evaluation phases addressing different perspectives on the same dataset.

In Phase 1 (RQ1), authentic student–AI interactions were qualitatively analysed to identify the types and functions of student questions emerging during laboratory work. This exploratory phase produced a benchmark set of representative questions that served as input for the following analyses. In Phase 2 (RQ2), the research team conducted a qualitative expert evaluation of AI tutor outputs generated under seven prompting strategies, using the benchmark questions derived from Phase 1. Experts assessed which configurations produced pedagogically valuable and accurate responses, representing an informed professional perspective on the instructional usefulness of AI tutors. In Phase 3 (RQ3), the same set of responses was evaluated by students through quantitative ranking and Likert-scale rating, reflecting learners' preferences and perceived helpfulness. This phase provided insight into how students' views of usefulness align or differ from expert evaluations.

The study thus combined qualitative and quantitative methods in order to capture complementary perspectives on the same phenomenon: expert judgments representing pedagogical quality and student perceptions representing user experience. The sequential structure ensured coherence between data collection, analysis, and interpretation, supporting triangulation between professional and learner perspectives.

As a pilot study, the purpose was not to generalise the results but to test the prototype and methodological procedures in a controlled laboratory setting. The findings aim to guide future development of AI tutoring systems and to inform larger-scale research on sustainable, digitally enriched pedagogy in STEM education.

2.2. Participants and Setting

The study was conducted within an undergraduate Electrical Engineering course at the Faculty of Education, University of Ljubljana. Participants were eight first-year students enrolled in the course. All students who were present during the laboratory sessions of the course participated in the study. They were fully informed about the purpose and procedures of the research and agreed to take part voluntarily.

The empirical part of the study took place during two 90 min laboratory sessions, each mostly focused on practical problem-solving tasks related to circuit design, microcontroller programming, sensor analysis, and electrical measurements. Students worked individually on desktop PCs and interacted with the GEN-AI tutor through a chat interface accessible via browser.

Following the sessions, seven out of the eight participants engaged in a follow-up activity where they reviewed and ranked written responses generated by different tutor

configurations. One participant was absent on the day of the activity, resulting in a final sample of seven. The responses were based on a selected set of benchmark questions, drawn from the authentic queries students had originally asked during the laboratory sessions. Each benchmark question was processed by multiple tutor versions, each applying a distinct prompting strategy. Students were asked to rank the responses according to which one they would most prefer to receive in a tutoring interaction. They first rated each response on a five-point Likert scale (1 = least helpful, 5 = most helpful) and then produced a forced ranking, ordering the responses from most to least preferred without allowing ties.

In addition, the research team reflected on the expanded set of responses drawing on their teaching experience and involvement in the study. These reflections supported the interpretation of student feedback and helped to contextualise the findings.

2.3. Tutoring Tool Architecture

To support pedagogical flexibility and enable systematic experimentation, a configurable AI tutor was developed using the open-source no-code platform n8n [48]. This platform was chosen because it enables rapid prototyping, allowing quick adjustments, modifications, and use of different LLMs without extensive programming. This architecture connects a custom frontend for teachers and students to backend workflows that processes inputs, routes prompts, manages tutor behaviour, and retrieves instructional content. As shown in Figure 1, teachers access a dedicated management interface designed for pedagogical control and workflow customization. Through this interface, they can create virtual classrooms, assign or adjust tutoring workflows, upload instructional materials, define system messages, and monitor student activity in real-time.

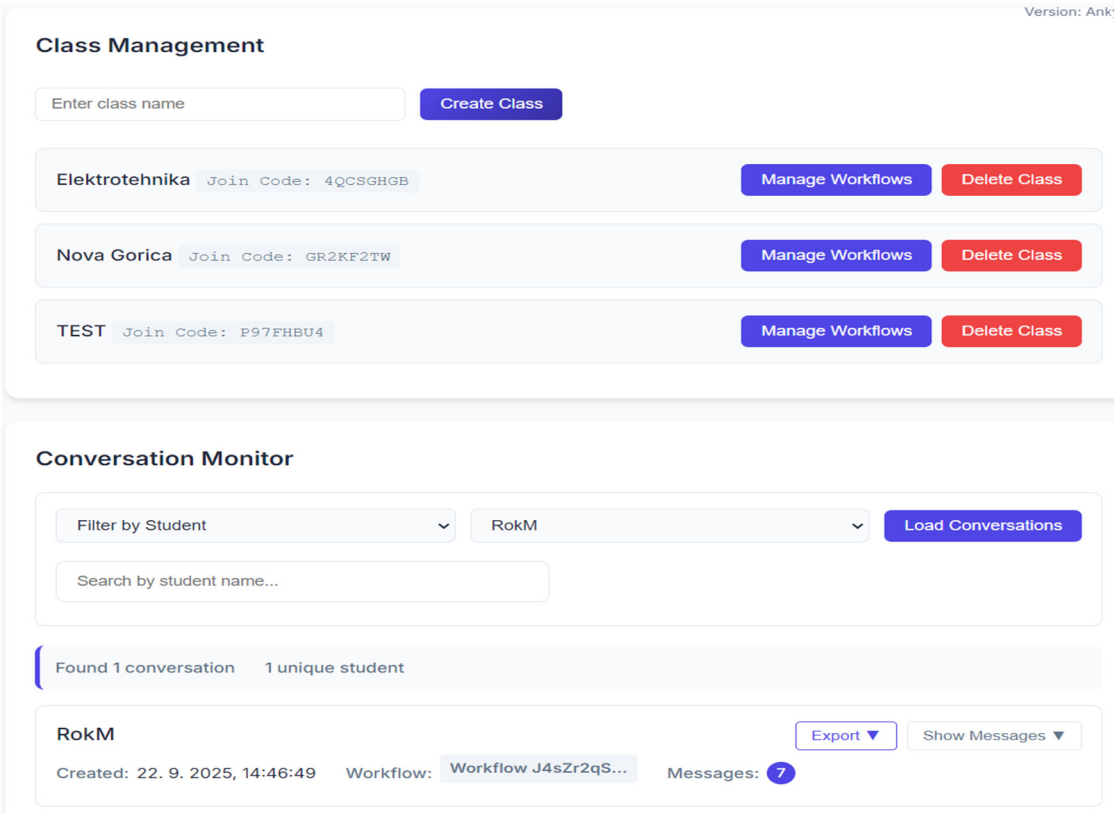


Figure 1. Part of teacher management dashboard of the GEN-AI tutoring platform, showing class creation, workflow assignment, and conversation monitoring tools.

Students, conversely, interact with the tutor through a chat interface that supports natural language input (Figure 2). This intuitive frontend lets them ask questions, receive personalised assistance, and engage with the instructional logic configured by their teacher.

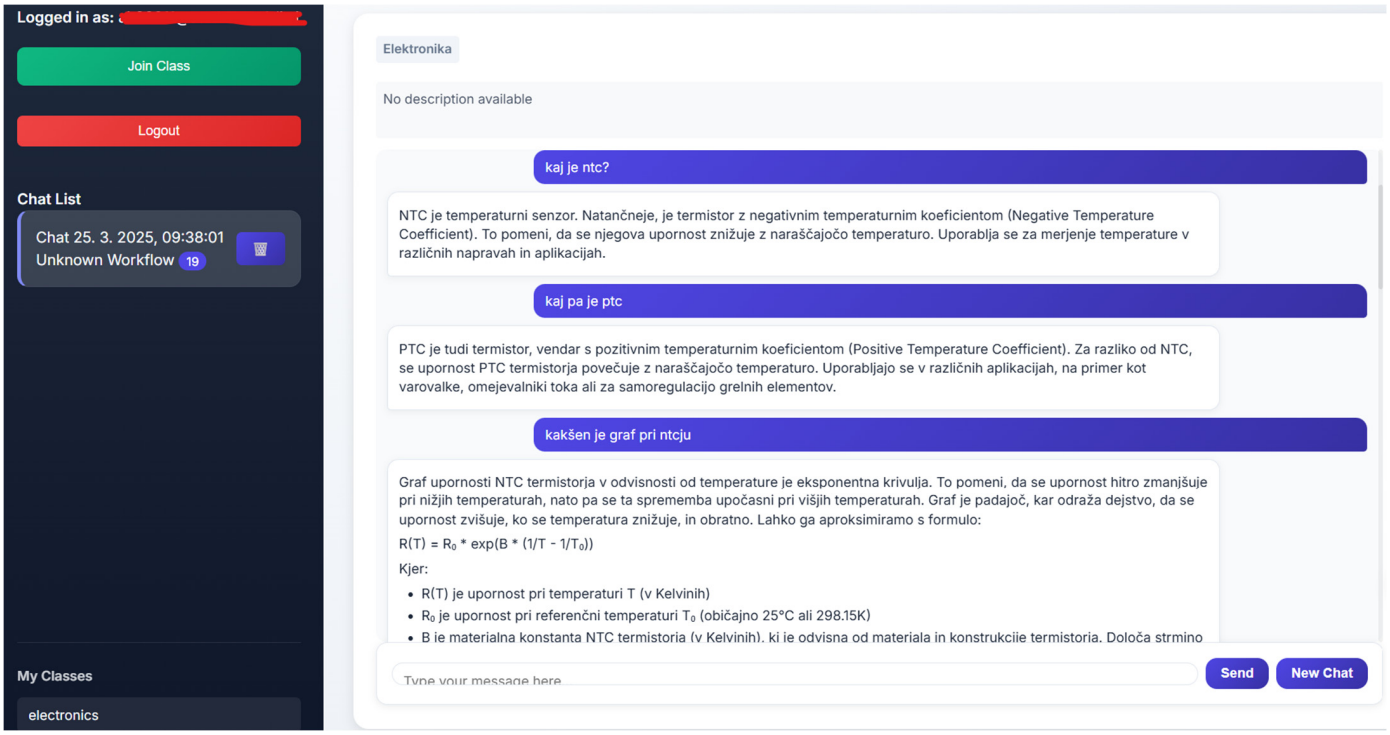


Figure 2. Student chat interface of the GEN-AI tutor, showing how learners log into their assigned class and interact with the system through natural language queries.

The core language model used in this study was Claude 3.5 Sonnet (accessed via API), although the architecture can integrate other widely used LLMs. All tutoring logic was executed through modular workflows in a no-code environment (Figure 3), ensuring rapid deployment and transparency. This design also facilitates ease of use in educational settings. The workflow applied in our study represented the minimal configuration required for experimentation, while the system remains fully extensible: advanced users may incorporate conditional logic, retrieval-augmented generation, or other automation components depending on their pedagogical goals and technical expertise. Detailed examples of prompts used for each strategy category and the student questionnaire on perceptions of and experiences with the AI tutor are provided in the Supplementary Materials [Online].

In this basic configuration, the tutor is structured around a modular input architecture (Figure 4), composed of four key elements:

1. Educator prompt (system message): A prompt defined by the teacher that sets the tutor's instructional role, tone, and strategy;
2. Student input: The learner's current query or message submitted via the chat interface;
3. Dialogue context (history): Previous messages in the conversation that help the tutor understand the current input;
4. Retrieval layer: Optional access to materials provided by teacher and stored in a vector database, invoked by the tutor when additional context is needed.

This layered design not only supports dynamic adaptation of the tutor's behaviour to across instructional contexts, but also provides a robust testbed for controlled experimentation with prompt strategies, information access, and conversational memory.

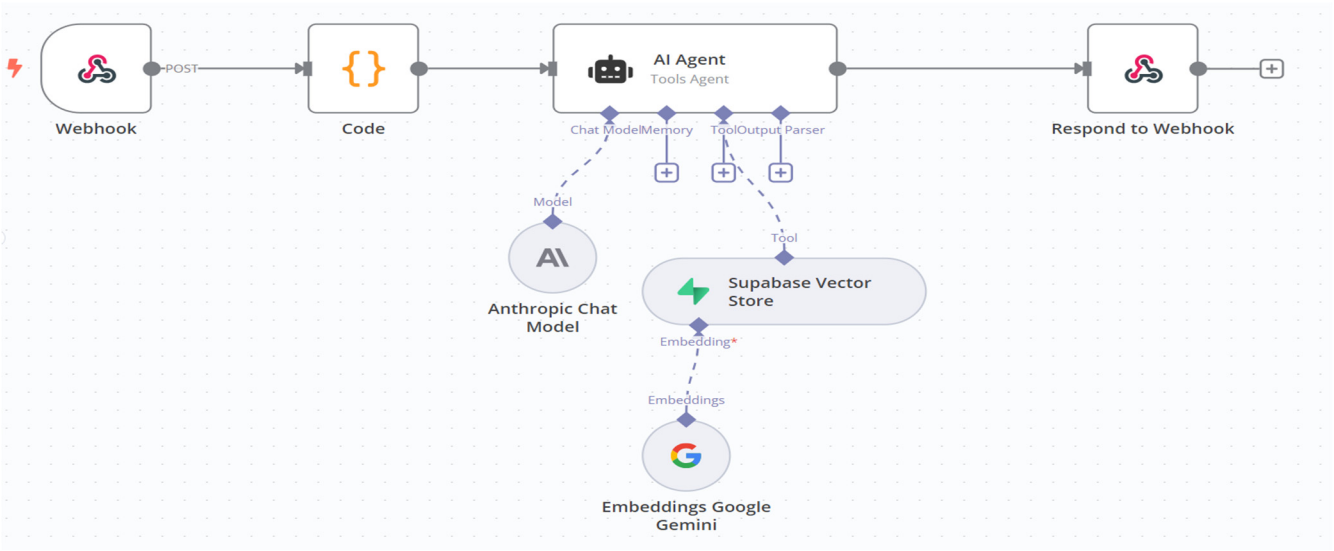


Figure 3. Example of a basic workflow in the GEN-AI tutor, showing input processing via Webhook and Code nodes, central AI Agent logic, optional vector store retrieval, and response delivery to the frontend. An asterisk (*) marks nodes (e.g., Supase Vector Store, Chat Model) that require completion or extension with an additional tool to finalize the workflow.

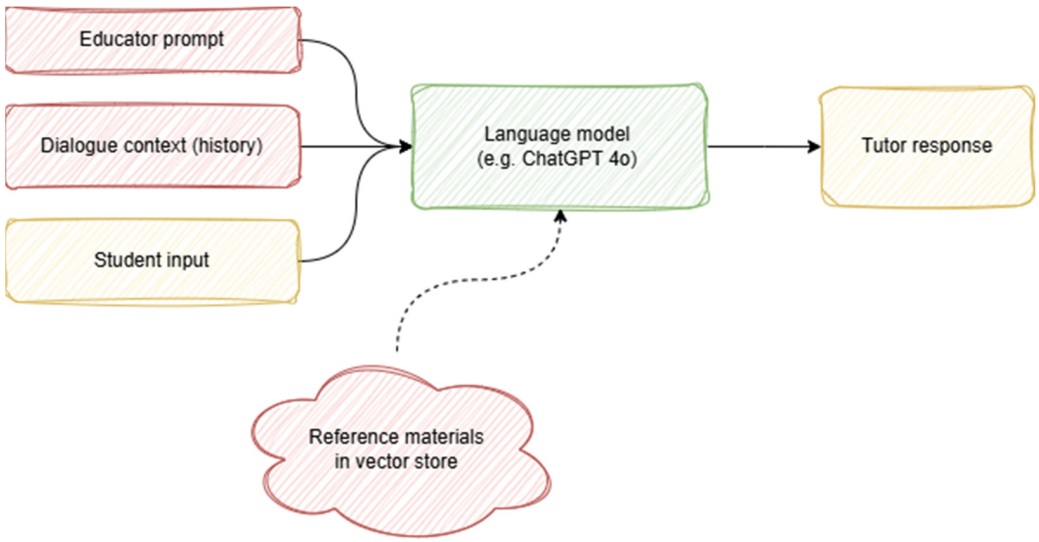


Figure 4. Inputs and information flow in the modular GEN-AI tutor architecture.

2.4. Prompt Configurations

For the experimental phase, seven distinct tutor versions were created by modifying only the system message (Table 1). These configurations addressed RQ2 and RQ3 by enabling a controlled comparison of how prompt strategies affect the clarity, usefulness, and reception of AI-generated responses. To ensure a valid comparison, all system messages were kept minimal and strategy-specific to isolate the effect of a single strategy. Prompt formulations were informed by recent literature on educational prompt engineering [27–37,43,44]. Where required, we introduced only the minimal decisions needed to activate a strategy (e.g., assigning a teacher role for persona, or providing a small number of examples for few-shot strategy). All other LLM settings (e.g., temperature, top-p, maximum tokens) were held constant across configurations. To further isolate prompt effects, retrieval-augmented generation was disabled and the evaluation was conducted in a single-turn mode without dialogue history. The seven strategies evaluated were: neutral (baseline), persona, template,

chain-of-thought, few-shot, game-based, and flipped. The core phrasing of each system message is listed in Table 1.

Table 1. Overview of LLM prompt strategies and their educational goals.

Prompting Strategy	System Message	Function/Intended Effect
Neutral (Baseline)	You are a helpful tutor.	Produces standard answers without any specific strategy or structure.
Persona	You are an experienced teacher. Provide answers as a teacher would, focusing on clarity and correctness.	Shapes the output by assigning the model a specific role or identity.
Template	If the question asks for a definition or fact, respond in the format: Term–Definition–Example–Common mistake (if any). If the question asks for a procedure or method, list the steps one by one in order.	Structures the output in a fixed format to make explanations more consistent.
Chain-of-thought	Think step-by-step. Explain your reasoning before giving the final answer.	Makes the model explain its thinking step-by-step before giving the answer.
Few-Shot	Use the logic demonstrated in the following examples to answer the next question in the same style. [example 1], [example 2], ...	Makes the model follow the examples given in the prompt and produce similar output
Game-based	Turn each question into a quiz. Ask the user first, then reveal the correct answer with a short explanation. [following rules]	Turns the output into a game-like format with questions, feedback, and interaction.
Flipped	Ask one question at a time (Socratic method) to help the student reach the answer on their own. Do not give direct answers.	Makes the model ask questions instead of answering directly, guiding the user step-by-step.

2.5. Benchmark Question Set

The benchmark question set used for evaluation was drawn from a pool of authentic questions generated by students and collected during the exploratory phase. Selection followed a dual classification scheme, designed to capture both the pedagogical function and content focus of student queries.

First, questions were selected to represent the four main functional categories commonly observed in learning interactions (Table 2).

Table 2. Functional categories of benchmark questions with representative examples.

Category	Description	Example
Factual	Definitions or straightforward knowledge recall.	"What is the definition of electrical resistance?"
Procedural	Step-by-step guidance or instructions.	"How do I calculate the current through this resistor using Ohm's law?"
Conceptual	Reasoning about underlying principles or cause–effect relationships.	"Why does increasing the resistance decrease the current in the circuit?"
Metacognitive	Self-evaluation, hypothesis formulation, or confidence checks.	"I believe the circuit should behave this way, but how can I check if I'm right?"

Second, after an initial review of the full dataset, we developed a set of categories related to content to reflect the variety of inputs typically encountered in practical STEM ac-

tivities. These content-related categories were used solely to contextualise the environment and the situations in which learners posed their questions. They served a descriptive, supportive role and were not used as selection criteria for the benchmark set nor as factors in the experimental comparison. Table 3 presents the final set of content categories alongside representative examples drawn from the original dataset.

Table 3. Content-related categories of benchmark questions with representative examples.

Category	Description	Example
Troubleshooting	Addressing unexpected outcomes, system faults, or technical errors, typically in hardware or code.	"The output voltage of the sensor only changes by 0.1 V when I change the temperature by 20 °C. What could be wrong?"
Theoretical reasoning	Aiming to understand or predict system behaviour based on underlying principles (e.g., Ohm's law).	"Why is the RMS value of an AC voltage different from its peak value?"
Measurement-focused	Related to taking or interpreting measurements, expected vs. actual values, or observed data.	"How can we measure the effective voltage of a source?"
Task clarification	Seeking clarification about instructions, tools, or expected outcomes.	"How should I approach this task? Based on the measured data and the U(t) graph, draw a phasor diagram showing the conditions at t = 0.15 s."
Code-related	Referencing programming behaviour, modifications, or including code snippets.	"Which part of the Arduino code should I change to make the robot move more slowly?"
Hardware/Equipment-related	Referring to components, wiring, or physical setup.	"Can I power the display with 9 V instead of 5 V?"
Social	Informal, conversational, or socially motivated messages.	"Thanks! What else can you do?"
Underspecified/Incomplete	Short, ambiguous, or context-dependent inputs.	"And now?"
Incorrect assumptions	Based on flawed conceptual understanding or technical misconceptions.	"If I use two thermistors in a voltage divider, will it make the sensor more sensitive?"

This two-dimensional classification enabled a deeper understanding of student engagement with the AI tutor and informed the selection of representative benchmark questions for the second phase of the study.

Selection of Benchmark Questions for Evaluation

We selected six benchmark questions from the phase 1 set of authentic student queries. To reflect the observed distribution, we included two items from the more frequent functional categories, factual, and procedural, and one item each from conceptual and metacognitive. The selection prioritised clarity, representativeness of typical laboratory queries, and unambiguous phrasing. To illustrate, the following two examples represent typical benchmark items used in the study:

- Factual: "What does the term 'effective voltage' mean in the context of an AC power source?"
- Conceptual: "The NTC sensor output changes by only 0.1 V for a 20° C temperature change. What could be wrong?"

2.6. Evaluation Procedure

To address RQ2 and RQ3, we used a dual evaluation approach combining learner rankings with research team reflections. For each of the six benchmark questions, we generated one response per strategy (seven total). By default, outputs were produced in single-turn mode without dialogue history and with retrieval disabled. For strategies that do not provide a direct answer by design (e.g., flipped, game-based), we allowed a short, scripted multi-turn exchange (up to three assistant turns) using predefined neutral student replies to progress the interaction.

Participants reviewed the seven responses side-by-side, anonymised and labelled with randomised letter codes (A–G). The mapping between letters and strategies was re-randomised for each question to prevent identification. They assigned a unique rank from 1 (most preferred) to 7 (least preferred), with no ties, based on overall preference. In addition to ranking, students provided brief written comments on each response, noting perceived strengths and weaknesses. These qualitative insights supported the interpretation of quantitative results by clarifying the reasons behind preferences or disagreements.

In parallel, the research team conducted a qualitative review of both the benchmark responses and a broader sample of student–AI interactions from the laboratory sessions. This allowed the analysis to capture not only direct comparisons across strategies, but also contextual patterns observed in authentic conversations. Although no formal scoring rubric was applied, the review was carried out by two subject-matter experts who drew on their teaching experience and involvement in the study. They discussed their evaluations collaboratively, focusing on instructional criteria such as factual accuracy, clarity of explanation, and relevance to the learner's needs. Special attention was given to how each prompt strategy supported student understanding, encouraged reasoning, and maintained engagement.

Student rankings were analysed using non-parametric tests appropriate for ordinal, related samples. A Friedman test assessed differences in mean ranks across the seven strategies, with blocks defined as one student's ranking (1–7) of all seven strategies for one question, only complete blocks were included, yielding N = 42. A significant Friedman result ($p < 0.05$) was interpreted as evidence that not all strategies were equally preferred. When this was the case, we inspected mean ranks (lower = better) and conducted pairwise comparisons using two-tailed Wilcoxon signed rank tests ($\alpha = 0.05$) to identify which specific strategies differed significantly from the top-ranked alternative. Pairwise p-values were reported without adjustment, as no correction was applied given the pilot nature of the study, which was primarily intended to explore directions for further didactic development. The degree of agreement among students for each benchmark question was assessed using Kendall's coefficient of concordance (W), which ranges from 0 (no agreement) to 1 (perfect agreement). Interpretation followed established thresholds:

- $0.00 < W < 0.20$—slight agreement
- $0.20 \leq W < 0.40$—fair agreement
- $0.40 \leq W < 0.60$—moderate agreement
- $0.60 \leq W < 0.80$—substantial agreement
- $0.80 \leq W < 1$– almost perfect agreement

All statistical analyses were conducted in IBM SPSS Statistics (Version 28). Given the small number of participants, the analyses were exploratory in nature.

2.7. Validity and Reliability

To ensure the validity and reliability of the study, both content-related and methodological considerations were addressed.

The overall study design and AI tutor setup were reviewed by two colleagues from the Department of Computer Science at the Faculty of Education, University of Ljubljana. Although not subject-matter experts in electrical engineering, they provided critical feedback on the coherence of the tutor's logic, the structure of prompts, and the methodological alignment between research phases. Their review helped confirm that the tutor's configuration, data collection procedures, and evaluation framework were internally consistent and appropriately designed for the study's aims.

Methodological reliability was further reinforced by the research team's prior experience [49] using the same GEN-AI tutor in related educational contexts, particularly in the Robotics course. In that earlier implementation the tutor was adapted for project-based learning and integrated with didactic principles such as active learning and Kolb's experiential cycle. Insights from that deployment helped ensure stable system performance and informed the present study's design.

Collectively, these measures strengthened the credibility of the research process and provided a robust foundation for interpreting the study's findings.

3. Results

3.1. Classification of Student Questions (RQ1)

During the first phase, a total of 274 student messages were collected during laboratory sessions. Of these, 66 entries were excluded from the first part of the analysis because they were either off-topic, incomplete beyond interpretation, or unrelated to the instructional objectives. The final dataset for functional classification consisted of 208 valid messages.

Each message was classified into one of four functional categories based on its role in the learning process:

- Procedural messages (n = 123; 59.1%) represented the majority of entries, focusing on instructions, task steps, or operational procedures.
- Factual messages (n = 55; 26.4%) involved requests for definitions, factual confirmation, or parameter values.
- Conceptual messages (n = 21; 10.1%) addressed underlying principles and causal mechanisms.
- Metacognitive messages (n = 9; 4.3%) reflected self-evaluation, uncertainty, or planning strategies.

In parallel, all 274 messages (including those excluded from functional coding) were analysed for their content-related characteristics to capture the situational variety of interactions in a STEM laboratory setting. Each message was manually coded and assigned to a single dominant category based on its primary theme. The distribution is summarised in Table 4.

Table 4. Distribution of student messages by content category.

Content Category	Count (n)	Percentage (%)
Code-related	66	24.1
Underspecified/Incomplete	59	21.5
Measurement-focused	35	12.8
Social comments	30	10.9
Hardware/Equipment-related	26	9.5
Troubleshooting	24	8.8
Theoretical reasoning	14	5.1
Task clarification	13	4.7
Incorrect assumptions	7	2.6

The two classification schemes provide complementary perspectives on the dataset and served as the basis for constructing the benchmark question set in Phase 2.

3.2. Research Team Evaluation of Tutor Responses Based on Prompt Strategy (RQ2)

To address RQ2, qualitative evaluation was conducted on responses produced by seven distinct tutor configurations. Clear differences emerged in how each strategy structured its responses and addressed the student queries. To illustrate these differences, representative excerpts from the tutor's output are provided below.

- Neutral responses were brief and to the point, typically providing a direct answer without much additional context or pedagogical framing.
- Template outputs followed a consistent, easily interpretable format. For factual and procedural questions, these responses were typically well organised and included common elements such as definitions, examples, or step-by-step instructions. Example: "Term—Effective voltage . . . Definition—The RMS value of an AC source ... Example—Household mains voltage is 230 V RMS ... Common mistake—Confusing RMS value with peak voltage. . ."
- Chain-of-thought responses emphasised reasoning and step-by-step explanations, guiding the student through a structured logical progression. This structure was particularly evident in conceptual and troubleshooting contexts.
- Example: "Let's begin by understanding what alternating current (AC) means. Then we can define the effective voltage, which is calculated based on the power it delivers. Finally, we'll see how it differs from peak voltage."
- Persona responses adopted a conversational and supportive tone, often framed as if written by a human teacher. This made the responses more approachable, though sometimes less concise. Example: "Great question! In the lab, we often work with AC sources, so it's important to understand the concept of effective voltage. Think of it as the value that tells you how much work the voltage can really do."
- Few-shot responses followed the style of the examples included in the prompt. When questions were similar in form, the answers mirrored the demonstrated structure, but with different question types, outputs sometimes deviated from the expected structure.
- Game responses introduced playful elements (e.g., roleplay, quizzes, exclamations). The format was versatile, with outputs ranging from short quiz interactions to roleplay scenarios and other gamified styles. Example: "Tesla smiles and asks: 'Can you guess why RMS voltage matters when plugging in devices at home?'"
- Flipped responses replaced direct answers with follow-up questions. Instead of providing explanations, the tutor prompted the student to reflect or propose an answer before continuing. Example: "What do you think effective voltage means? How might it differ from the highest voltage reached in an AC waveform?"

3.3. Student Preferences Regarding Tutor Responses (RQ3)

Student rankings were analysed using the Friedman test for related samples (blocks = student–question pairs with complete ranks, N = 42). The test revealed statistically significant differences among the seven strategies, $\chi^2(6, N = 42) = 92.14$, $p < 0.001$. Kendall's coefficient of concordance indicated a fair level of agreement among rankings ($W = 0.366$). Table 5 presents the overall average ranks. The template strategy was most preferred (average rank = 2.76).

To further examine whether preferences differed across question types, the benchmark items were grouped into their primary functional categories: factual, procedural, conceptual, and metacognitive. As shown in Table 6, the Friedman test indicated statistically significant differences in mean ranks among strategies for each category (all $p < 0.05$).

Table 5. Average Student Rankings of Tutor Configurations (lower values indicate stronger preference).

Prompt Strategy	Average Rank
Template	2.76
Persona	3.22
Neutral	3.29
Chain-of-thought	3.50
Few-Shot	3.57
Flipped	5.69
Game	5.93

Table 6. Best-ranked strategy and agreement per primary question category.

Primary Category	Best Strategy	Mean Rank	W	*p*
Factual	Template/Persona	2	0.656	<0.001
Procedural	Template	1.71	0.456	0.004
Conceptual	Chain-of-Thought	2.71	0.600	<0.001
Metacognitive	Neutral	2.57	0.397	0.011

Note: Lower mean ranks indicate higher preference; W = Kendall's coefficient of concordance; *p*-values from Friedman test.

Based on the interpretation thresholds outlined in the Section 2, student agreement was substantial for the factual category (W = 0.656), moderate for the conceptual (W = 0.600) and procedural categories (W = 0.456), and fair for the metacognitive category (W = 0.397). In addition, Wilcoxon signed rank tests ($\alpha = 0.05$) were conducted for each benchmark question to determine whether the best ranked strategy was significantly preferred over the alternatives. The results for questions Q1–Q6 are summarised in Table 7.

Table 7. Per-category winners and significant pairwise differences (Wilcoxon test).

Question	Winner	Statistically Better than ($p < 0.05$)
Factual	Template	Game, Flipped, Few-shot
Procedural	Template	Game, Chain-of-thought, Few-shot
Conceptual	Chain-of-thought	Flipped, Game
Metacognitive	Neutral	Game

These results indicate that the overall preference pattern aligned with the aggregated mean ranks. However, the per-question analysis showed that the top ranked strategy for a given question was not always significantly better than all alternatives. The weakest performing strategies, game-based and flipped, were most frequently outperformed in pairwise comparisons.

4. Discussion

4.1. RQ1—What Students Ask a GEN-AI Tutor to Do

In our STEM, laboratory-based electrical engineering setting, student–AI interactions were predominantly task-driven. As shown in Section 3.1 and Table 4, most prompts were procedural (59.1%) or factual (26.4%), focusing on carrying out steps, verifying parameters, or requesting definitions. Only a small share were conceptual (10.1%) or metacognitive (4.3%), indicating limited engagement with underlying principles or reflective strategies. The content-based classification supports this finding: most interactions related to immediate laboratory demands such as writing or adjusting code and checking measurements, while very few engaged in broader reasoning or reflection. This suggests that the tutor was

primarily used as a practical support tool to maintain task flow, with limited contribution to deeper learning processes.

This distribution reveals a clear dominance of task-focused interactions, suggesting that students relied on the tutor primarily for immediate, instrumental assistance rather than conceptual exploration. Such a pattern is consistent with prior research on help-seeking in intelligent tutoring systems, where learners tend to request direct procedural guidance when operating in time-constrained or performance-driven contexts [3].

From the perspective of digitally enriched pedagogy, this finding highlights the need for scaffolded integration of AI tutors that guide students from procedural assistance toward reflective reasoning. The near absence of metacognitive questioning was unexpected and suggests that students may perceive the AI tutor more as a technical assistant than as a learning partner.

4.2. RQ2—How Prompting Strategies Shape Instructional Value (Authors' View)

Although all prompting strategies produced factually correct answers, their instructional value varied considerably. This confirms prior work showing that prompt design strongly affects the pedagogical usefulness of LLM outputs [27,34–37] and highlights the importance of teacher-controlled configuration for aligning AI responses with learning goals [12–15].

Strategies such as chain-of-thought and flipped were judged most valuable for fostering reasoning, conceptual understanding, and metacognitive engagement. These approaches are recommended when deeper learning is the goal, but they require more time and cognitive effort, making them less efficient for routine queries. Template and persona strategies provided a more balanced trade-off: template ensured structure and predictability, while persona created a supportive and human-like tone that maintained learner confidence. By contrast, game-based, neutral, and few-shot strategies were less robust pedagogically, but can be useful for engagement, brevity, or practice in specific cases, but limited as stand-alone approaches.

These reflections indicate that no single strategy is universally optimal. Instead, the findings reveal two broad design patterns: strategies that prioritise efficiency and clarity versus those that promote reflection and reasoning. The instructional value of each configuration thus depends on how well it aligns with the intended learning process. The unexpected observation was that even minimal prompts—such as persona or template—could produce responses perceived as pedagogically coherent, underscoring the potential of simple prompt structures when coupled with thoughtful instructional framing.

4.3. RQ3—How Students Ranked Responses

Student rankings revealed modest differences overall (Table 5), yet clear tendencies by question type (Table 6). Template was most often preferred for factual and procedural questions, chain-of-thought for conceptual ones, and neutral for metacognitive tasks. Pairwise comparisons (Table 7) confirmed these patterns only partially: winning strategies outperformed some weaker alternatives, but were not consistently superior to all others, underscoring that differences remained modest.

Overall, students gravitated toward direct and well-structured responses (template, persona, neutral), while dialogue-heavy strategies such as flipped and game-based were least appreciated. In the lab context, this suggests that learners valued quick, definitive answers that supported efficient task completion more than formats designed to encourage extended reflection. Similar dynamics have been observed in prior studies [28,29], where step-by-step or Socratic interactions enhanced engagement but were not consistently perceived as more helpful, and did not always improve long-term outcomes. Especially, Blasco

and Charisi [29] found that while Socratic tutoring and step-by-step reasoning enhanced student interaction, they did not significantly improve learning gains or retention. Learners often perceived direct, outcome-oriented responses as more useful, even when dialogic approaches held greater pedagogical potential. Our findings align with this pattern: in time-constrained, performance-driven environments such as laboratory work, students appear to prioritise efficiency and clarity over exploratory dialogue. This underscores the importance of context when evaluating AI tutor designs, what is pedagogically ideal may not always align with learners' immediate needs or preferences.

4.4. The Preference–Pedagogy Gap

The comparison of author evaluations (RQ2) and student rankings (RQ3) highlights a consistent gap between pedagogical strength and learner preference. Strategies judged most valuable by authors, such as chain-of-thought and flipped, did not align with students' preference for concise, task-oriented responses. This reflects a broader pattern reported in previous research: fluency and responsiveness often mask the absence of scaffolding needed for deeper learning [10,19], and learners may reject demanding strategies even when they support understanding [23,28,29]. This gap suggests that teacher guidance and careful instructional design remain essential when integrating AI tutors into STEM education.

In the metacognitive category, no strategy outperformed the neutral configuration. Its concise style was judged as most effective, further illustrating that efficiency is prioritised over pedagogical depth. For AI tutor design, these findings underscore the need to bridge the preference–pedagogy gap by combining immediacy with elements that promote reflection and conceptual growth.

4.5. Design Implications: Richer Prompts and Multi-Step Orchestration

Future development should aim to design AI tutors that are not only theoretically well-grounded but also practically applicable in diverse educational contexts. Among the promising directions are hybrid prompting, retrieval-augmented generation (RAG), and adaptive scaffolding.

To advance these directions, research should further investigate enriched prompt designs that blend strategies to balance clarity, engagement, and depth. In this study, prompts were intentionally kept simple. In practice, however, providing richer contextual information, such as learner background, prior responses, or task-specific cues, could enable more precise and pedagogically rich interactions [12,13].

A promising approach is hybrid prompting, where multiple strategies are integrated within a single configuration to leverage their complementary strengths [28,33]. For example, a persona strategy can enhance accessibility and engagement, a template can provide clear structure, and short follow-up questions can stimulate reflection. Selective few-shot examples can further model the desired response style. Another approach that combines different strategies involves multi-agent workflows, in which specialised tutors contribute complementary strengths such as factual accuracy, conceptual reasoning, or motivational support [15].

Additional improvement could come from retrieval-augmented generation (RAG), enabling tutors to draw on teacher-provided materials and tailor responses to the learner's prior knowledge and current context [37]. In practice, this allows precise, context-aware support, such as answering a question tied to a specific assignment or lab task.

The tutoring tool used in this study already supports key mechanisms for such enhancements, including teacher oversight, real-time switching between strategies, and integration of RAG or hybrid prompts. Future iterations could further incorporate adaptive prompting mechanisms that scaffold learners from preferred, easy-to-follow strategies

toward more conceptually demanding forms of reasoning. By monitoring learner responses and progress, the tutor could gradually shift from procedural guidance to reflective questioning, encouraging deeper cognitive engagement while maintaining accessibility and motivation. Such progression-oriented adaptation could bridge the gap between learner preferences and pedagogical value.

4.6. Limitations and Future Work

The findings of this pilot study should be interpreted in light of its limitations.

First, the study was conducted with small number of participants within a single STEM domain (electrical engineering), which constrains the generalisability of the results. However, the laboratory-based nature of the tasks shares similarities with other STEM contexts, where students engage in experimental and measurement-based reasoning. Moreover, electrical engineering as a discipline typically integrates multiple STEM dimensions, physics, mathematics, programming, and electronics, making it a representative domain for studying how learners interact with AI tutors in experimental problem-solving settings. Still, replication across additional disciplines such as physics or chemistry would provide stronger empirical support for the preference–pedagogy patterns identified here.

Beyond methodological aspects, several technical and ethical considerations also influence the applicability of the proposed tutoring approach. From a technical perspective, the operational cost of using large language models depends on the number of processed tokens, which include both input and output text. In this configuration, pricing was set at USD 3 per one million input tokens and USD 15 per one million output tokens. To illustrate, one million tokens roughly corresponds to about 750,000 English words or approximately 600–700 pages of text. While this cost structure remains manageable for small-scale or pilot classroom implementations, larger deployments would require careful monitoring of usage or the adoption of more cost-efficient models. Costs can increase when the full conversational context is preserved across multiple exchanges, which should be considered when designing longer or multi-turn interactions.

Regarding scalability, the tutoring tool was designed as a locally hosted application, which stores all interactions and configurations on the teacher's computer and connects to a selected LLM provider only through an API. This decentralised architecture ensures strong privacy control and allows teachers to adapt the tutor to their instructional context without external dependencies. For large-scale or institutional implementations, other platforms offering centralised management and shared infrastructure would be more appropriate, whereas the present tool prioritises teacher autonomy, transparency, and local data control. In the long term, developing a centralised yet teacher-centric platform, combining ease of use with the same level of transparency and configurability, could represent a valuable direction for future work.

On the ethical side, issues of bias, data privacy, and transparency remain central. Bias is less prominent in factual, experimentally based STEM domains such as electrical engineering, yet subtle linguistic or representational biases can still appear in examples or explanations. Data privacy risks are reduced by the local storage architecture, as no student data are permanently transmitted to external servers. Nevertheless, responsible data handling and ongoing transparency about how prompts and responses are processed remain essential to ensure trustworthy educational use of AI tutors.

Finally, the results concerning AI response quality (RQ2) reflect the performance of a specific large language model (Claude 3.5 Sonnet) and are therefore subject to temporal variation and potential model drift. Achieving high-quality responses typically requires iterative prompt adjustment, as individual models vary in their sensitivity to specific phrasing and structure [27]. In this study, prompts were iteratively refined for the selected model

while kept deliberately simple to ensure that the intended pedagogical strategy determined the observed effects. Although future updates may affect the overall quality of model responses, the relative differences between prompting strategies are likely to remain broadly consistent, as they reflect pedagogical design rather than model-dependent behaviour.

5. Conclusions

This study examined how students interact with a generative AI tutor in a STEM laboratory environment and how different prompting strategies shape the instructional value of tutor responses.

Regarding RQ1, student–AI interactions were predominantly task-driven, with most prompts being procedural or factual. Learners used the tutor mainly as a tool for immediate assistance rather than as a partner in conceptual exploration.

For RQ2, the analysis of prompting strategies showed that while all configurations produced factually correct responses, their pedagogical depth varied. Chain-of-thought and flipped prompts encouraged reasoning and reflection but required more effort and time, whereas template and persona provided clarity and structure with less cognitive demand.

In RQ3, student rankings confirmed this trade-off: learners favoured concise, outcome-oriented responses over exploratory dialogue, revealing a clear preference–pedagogy gap between what students value and what supports deeper learning.

These findings have practical implications for the design of AI tutors in education. To develop effective educational AI systems, it is essential that researchers and developers understand the strengths and limitations of different prompting strategies. By recognising how various configurations influence reasoning, engagement, and clarity, they can make informed decisions when adapting AI tutors to specific learning contexts. Building on this understanding, hybrid prompting, adaptive scaffolding, and retrieval-augmented generation (RAG) can be combined to design tutors that are both efficient and pedagogically meaningful. Such deliberate orchestration enables AI systems to balance immediacy with depth, serving not only as task assistants but as facilitators of reflective and sustainable learning.

While this was a small-scale pilot, its results provide valuable groundwork for future studies exploring how generative AI can be integrated into authentic classroom practice. Further research should replicate these findings across disciplines and larger samples, examine long-term learning effects, and address ethical and technical considerations such as data privacy and model drift.

Ultimately, as AI tutoring tools evolve and AI literacy expands, we can expect the preference–pedagogy gap to narrow, paving the way toward realising the vision of inclusive and sustainable education (SDG 4).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su17219508/s1, including examples of prompts used for each strategy category; Student questionnaire on perceptions of and experiences with the AI tutor.

Author Contributions: Conceptualisation, R.G. and D.R.; methodology, R.G. and D.R.; validation, R.G. and D.R.; formal analysis, R.G. and D.R.; investigation, R.G. and D.R.; resources, R.G. and D.R.; data curation, R.G. and D.R.; writing—original draft preparation, R.G. and D.R.; writing—review and editing, R.G. and D.R.; visualisation, R.G. and D.R.; supervision, R.G. and D.R.; project administration, R.G. and D.R.; funding acquisition, R.G. and D.R. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge the financial support of The Slovenian Research and Innovation Agency (ARIS, Founder ID: 501100004329) under the research core funding Strategies for Education for Sustainable Development Applying Innovative Student-Centred Educational Approaches (No.

P5-0451) and the project Developing the Twenty-First-Century Skills Needed for Sustainable Development and Quality Education in the Era of Rapid Technology-Enhanced Changes in the Economic, Social, and Natural Environment (No. J5-4573). The study was also funded by the JR MLADI 2025–2026—Public Call for Funding of Educational Programs for Children and Youth to Strengthen Digital Competences and Promote Science and Technology Careers (Ref. No. 430-20/2024-3150), financed by the Ministry of Digital Transformation of the Republic of Slovenia, and the GEN-UI Project—Generative Artificial Intelligence in Education (NRP 3350-24-3502), funded by the Republic of Slovenia, the Ministry of Education, and the European Union—NextGenerationEU (Founder ID: 100031478) by lead organisation: Anton Martin Slomšek Institute.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and was reviewed and approved by the Ethics Commission of the Faculty of Education of the University of Ljubljana (Approval code: 38/2024; approval date: 10 December 2024).

Data Availability Statement: The data presented in this study are available upon request from the author. The data are not publicly available due to privacy issues.

Acknowledgments: During the preparation of this manuscript, the authors used the ChatGPT software tool (models GPT-4 and GPT-5) for translation support and for improving style and readability. All AI-generated text was reviewed and edited by the authors, who take full responsibility for the final content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. VanLehn, K. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educ. Psychol.* **2011**, *46*, 197–221. [CrossRef]
2. Weitekamp, D.; Harpstead, E.; Koedinger, K.R. An Interaction Design for Machine Teaching to Develop AI Tutors. In Proceedings of the CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020.
3. Aleven, V.; McLaren, B.M.; Roll, I.; Koedinger, K.R. Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems. *Int. J. Artif. Intell. Educ.* **2016**, *26*, 205–223. [CrossRef]
4. Calo, T.; MacLellan, C. Towards Educator-Driven Tutor Authoring: Generative AI Approaches for Creating Intelligent Tutor Interfaces. In Proceedings of the Eleventh ACM Conference on Learning @ Scale, Atlanta, GA, USA, 18–20 July 2024.
5. Liu, S.; Guo, X.; Hu, X.; Zhao, X. Advancing Generative Intelligent Tutoring Systems with GPT-4: Design, Evaluation, and a Modular Framework for Future Learning Platforms. *Electronics* **2024**, *13*, 4876. [CrossRef]
6. Cain, J.; Rajan, A.S. Proof of Concept of ChatGPT as a Virtual Tutor. *Am. J. Pharm. Educ.* **2024**, *88*, 101333. [CrossRef] [PubMed]
7. Limo, F.A.F.; Tiza, D.R.H.; Roque, M.M.; Herrera, E.E.; Murillo, J.P.M.; Huallpa, J.J.; Flores, V.A.A.; Castillo, A.G.R.; Peña, P.F.P.; Carranza, C.P.M.; et al. Personalised tutoring: ChatGPT as a virtual tutor for personalised learning experiences. *Przestrz. Społeczna Soc. Space* **2023**, *23*, 293–312.
8. Pardos, Z.A.; Bhandari, S. ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS ONE* **2024**, *19*, e0304013. [CrossRef]
9. Chen, A.; Wei, Y.; Le, H.; Zhang, Y. Learning-by-Teaching with ChatGPT: The Effect of a Teachable ChatGPT Agent on Programming Education. *Br. J. Educ. Technol. arXiv* **2024**, arXiv:2412.15226. [CrossRef]
10. Nye, B.D.; Mee, D.; Core, M.G. Generative Large Language Models for Dialog-Based Tutoring: An Early Consideration of Opportunities and Concerns. In Proceedings of the LLM@AIED Workshop, Tokyo, Japan, 3–7 July 2023.
11. Shetye, S. An Evaluation of Khanmigo, a Generative AI Tool, as a Computer-Assisted Language Learning App. *Stud. Appl. Linguist. TESOL* **2024**, *24*, 38–53. [CrossRef]
12. Jiang, Z.; Jiang, M. Beyond Answers: Large Language Model-Powered Tutoring System in Physics Education for Deep Learning and Precise Understanding. *arXiv* **2024**, arXiv:2406.10934. [CrossRef]
13. Aperstein, Y.; Cohen, Y.; Apartsin, A. Generative AI-Based Platform for Deliberate Teaching Practice: A Review and a Suggested Framework. *Educ. Sci.* **2025**, *15*, 405. [CrossRef]
14. Reicher, H.; Frenkel, Y.; Lavi, M.J.; Nasser, R.; Ran-milo, Y.; Sheinin, R.; Shtaif, M.; Milo, T. A Generative AI-Empowered Digital Tutor for Higher Education Courses. *Information* **2025**, *16*, 264. [CrossRef]

15. Schmucker, R.; Xia, M.; Azaria, A.; Mitchell, T. Ruffle&Riley: Insights from Designing and Evaluating a Large Language Model-Based Conversational Tutoring System. In Proceedings of the International Conference on Artificial Intelligence in Education, Recife, Brazil, 8–12 July 2024.
16. Mzwri, K.; Turcsányi-Szabo, M. The Impact of Prompt Engineering and a Generative AI-Driven Tool on Autonomous Learning: A Case Study. *Educ. Sci.* **2025**, *15*, 199. [CrossRef]
17. Yang, X.; Liu, X.; Gao, Y. The Impact of Generative AI on Students' Learning: A Study of Learning Satisfaction, Self-efficacy and Learning Outcomes. *Educ. Tech. Res. Dev.* **2025**, 1–14. [CrossRef]
18. Bastani, H.; Bastani, O.; Sungu, A.; Ge, H.; Kabakcı, O.; Mariman, R. Generative AI Can Harm Learning. 2024. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4895486 (accessed on 28 June 2025).
19. Wecks, J.O.; Voshaar, J.; Plate, B.J.; Zimmermann, J. Generative AI Usage and Exam Performance. *arXiv* **2024**, arXiv:2404.19699.
20. Lehmann, M.; Cornelius, P.B.; Sting, F.J. AI Meets the Classroom: When Do Large Language Models Harm Learning? *arXiv* **2024**, arXiv:2409.09047.
21. Kotsiovos, J.; Chicone, R.; Doyle, S. Optimizing Student Success: The Impact of Generative AI in Teaching and Learning. In Proceedings of the 22nd Theory of Cryptography Conference (TCC 2024), Milan, Italy, 2–6 December 2024.
22. Avsec, S.; Rupnik, K. From Transformative Agency to AI Literacy: Profiling Slovenian Technical High School Students through the Five Big Ideas Lens. *Systems* **2025**, *13*, 562. [CrossRef]
23. Puech, R.; Macina, J.; Chatain, J.; Sachan, M.; Kapur, M. Towards the Pedagogical Steering of Large Language Models for Tutoring: A Case Study with Modeling Productive Failure. *arXiv* **2024**, arXiv:2410.03781. [CrossRef]
24. Maloy, R.W.; Gattupalli, S. Prompt Literacy. In *EdTechnica: The Open Encyclopedia of Educational Technology*; EdTech Books: Provo, UT, USA, 2024. [CrossRef]
25. Kurent, B.; Avsec, S. Synergizing Systems Thinking and Technology-Enhanced Learning for Sustainable Education Using the Flow Theory Framework. *Sustainability* **2024**, *16*, 9319. [CrossRef]
26. Rupnik, D.; Avsec, S. Student Agency as an Enabler in Cultivating Sustainable Competencies for People-Oriented Technical Professions. *Educ. Sci.* **2025**, *15*, 469. [CrossRef]
27. Cain, W. Prompting Change: Exploring Prompt Engineering in Large Language Model AI and Its Potential to Transform Education. *TechTrends* **2024**, *68*, 47–57. [CrossRef]
28. White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D.C. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv* **2023**, arXiv:2302.11382. [CrossRef]
29. Blasco, A.; Charisi, V. AI Chatbots in K-12 Education: An Experimental Study of Socratic vs. Non-Socratic Approaches and the Role of Step-by-Step Reasoning. 2024. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5040921 (accessed on 28 June 2025).
30. Li, Y.; Feng, S.; Wang, D.; Zhang, Y.; Yang, X. A Persona-Aware Chain-of-Thought Learning Framework for Personalized Dialogue Response Generation. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC 2024), Hangzhou, China, 1–3 November 2024.
31. Zhao, Y.; Cao, H.; Zhao, X.; Ou, Z. An Empirical Study of Retrieval Augmented Generation with Chain-of-Thought. *arXiv* **2024**, arXiv:2407.15569. [CrossRef]
32. Tan, C.W. Large Language Model-Driven Classroom Flipping: Empowering Student-Centric Peer Questioning with Flipped Interaction. *arXiv* **2023**, arXiv:2311.14708.
33. Mao, Y.; He, J.; Chen, C. From Prompts to Templates: A Systematic Prompt Template Analysis for Real-World LLMapps. *arXiv* **2025**, arXiv:2504.02052. [CrossRef]
34. Kakarla, S.; Borchers, C.; Thomas, D.; Bhushan, S.; Koedinger, K.R. Comparing Few-Shot Prompting of GPT-4 LLMs with BERT Classifiers for Open-Response Assessment in Tutor Equity Training. *arXiv* **2025**, arXiv:2501.06658.
35. Xu, D.; Xie, T.; Xia, B.; Li, H.; Bai, Y.; Sun, Y.; Wang, W. Does Few-Shot Learning Help LLM Performance in Code Synthesis? *arXiv* **2024**, arXiv:2412.02906. [CrossRef]
36. Kong, A.; Zhao, S.; Chen, H.; Li, Q.; Qin, Y.; Sun, R.; Zhou, X.; Wang, E.; Dong, X. Better Zero-Shot Reasoning with Role-Play Prompting. *arXiv* **2023**, arXiv:2308.07702.
37. Li, Z.; Wang, Z.; Wang, W.; Hung, K.; Xie, H.; Wang, F.L. Retrieval-augmented generation for educational application: A systematic survey. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100417. [CrossRef]
38. Deng, X.; Yu, Z. A Meta-Analysis and Systematic Review of the Effect of Chatbot Technology Use in Sustainable Education. *Sustainability* **2024**, *15*, 2940. [CrossRef]
39. United Nations. The Sustainable Development Goals Report 2024. Available online: https://unstats.un.org/sdgs/report/2024 (accessed on 28 June 2024).
40. Nedungadi, P.; Tang, K.Y.; Raman, R. The Transformative Power of Generative Artificial Intelligence for Achieving the Sustainable Development Goal of Quality Education. *Sustainability* **2024**, *16*, 9779. [CrossRef]

41. UNESCO. Guidance on Generative AI in Education and Research. UNESCO Report, 2023. Available online: https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research (accessed on 28 June 2024).
42. Schorcht, S.; Buchholtz, N.; Baumanns, L. Prompt the Problem–Investigating the Mathematics Educational Quality of AI-Supported Problem Solving by Comparing Prompt Techniques. *Front. Educ.* **2024**, *9*, 1386075. [CrossRef]
43. Olea, C.; Tucker, H.; Phelan, J.; Pattison, C.; Zhang, S.; Lieb, M.; Schmidt, D.; White, J. Evaluating Persona Prompting for Question Answering Tasks. In Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing (AIS 2024), Sydney, Australia, 22–23 June 2024.
44. Chen, E.; Wang, D.; Xu, L.; Cao, C.; Fang, X.; Lin, J. A Systematic Review on Prompt Engineering in Large Language Models for K-12 STEM Education. *arXiv* **2024**, arXiv:2410.11123. [CrossRef]
45. GEN-UI Initiative. Available online: https://gen-ui.si/ (accessed on 10 September 2025).
46. University of Ljubljana, Faculty of Education. 21st Century Skills Project. Available online: https://www.pef.uni-lj.si/razvijanje_vescin_21_stoletja/ (accessed on 10 September 2025).
47. Falegnami, A.; Romano, E.; Tomassi, A. The Emergence of the GreenSCENT Competence Framework. In *The European Green Deal in Education*; Routledge: London, UK, 2024; pp. 204–216.
48. n8n.io. Available online: https://n8n.io/ (accessed on 10 September 2025).
49. Rihtaršič, D.; Gabrovšek, R. Innovative Use of a GUI Tutor in Robotics and Electronics Projects within a Pedagogical Context. In *Education in the Age of Generative Artificial Intelligence: International Guidelines and Research*; Čotar Konrad, S., Flogie, A., Eds.; Založba Univerze na Primorskem: Koper, Slovenia, 2025; ISBN 978-961-293-431-6. [CrossRef]

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