education sciences

MDPI

Article

A Lecture-Specific AI-Based Tutor for Higher Education: Pedagogical Design and Empirical Evaluation

Samuel Tobler and Katja Köhler *

Center for Active Learning, ETH Zurich, 8092 Zurich, Switzerland; samuel.tobler@biol.ethz.ch
* Correspondence: katja.koehler@biol.ethz.ch

Abstract

Generative AI tools are increasingly used in higher education, yet most available systems lack pedagogical grounding, course alignment, and insight into student learning. This paper presents the development, implementation, and evaluation of the bioTutor, an open-source, course-specific AI chatbot designed to support constructivist learning in large university classrooms. The system integrates a curated knowledge base, a didactically structured interaction design, and a learning analytics dashboard for instructors that summarizes anonymized student-chatbot conversations. To assess students' perceptions of usefulness, ease of use, learning relevance, output quality, and result demonstrability, we developed an education-adapted extension of the Technology Acceptance Model (edTAM) and applied it in an introductory biology course with 407 enrolled students. During a 23-week deployment, students generated more than 10,000 interactions across over 1000 conversations. Questionnaire data indicate high usability, strong perceived usefulness, and broad interest in adopting similar tools. Usage patterns show that the bioTutor was employed both for learning and exam preparation. These findings suggest that students perceived pedagogically structured, course-grounded AI chatbots as useful for learning and exam preparation, while the lecturer dashboard provided aggregated insights into students' questions and recurring difficulties. The open-source framework enables adaptation to other disciplines and provides a scalable foundation for further research on didactically informed AI systems in higher education.

Keywords: generative AI; chatbot; higher education; intelligent tutoring system; learning analytics; constructivist pedagogy; educational technology

1. Introduction

Applications based on generative Artificial Intelligence (AI) have become an integral part of our daily lives and, consequently, are increasingly being integrated into educational settings (Sharma, 2024; Heeg & Avraamidou, 2023), and students rely on such applications in their academic routines (Abdaljaleel et al., 2024; Valeri et al., 2025; Zhang et al., 2024). From reviewing lecture content, solving problem sets, generating or revising code, or preparing for exams, to conducting literature research, tools like ChatGPT (OpenAI, 2026), GitHub Copilot (GitHub, n.d.), or Gemini (Google, 2026) are widely used and applied across various disciplines (Baig & Yadegaridehkordi, 2024). In line with this, a recent study conducted at a European technical university showed that a majority of students use ChatGPT or similar applications on a regular basis for study-related tasks (Balabdaoui et al., 2024). While these developments pose certain challenges, such as assessing the academic merit of student work, ensuring independent task completion, and upholding academic

Academic Editor: Mike Joy

Received: 1 April 2026
Revised: 13 May 2026
Accepted: 18 May 2026
Published: 21 May 2026

integrity (Kiryakova & Angelova, 2023), generative AI-based tools also offer an opportunity to address longstanding issues including personalizing learning experiences or enhancing student engagement in higher education (Vieriu & Petrea, 2025). Moreover, this technology presents a unique chance not only to support students in their learning processes but also to collect insights about these processes, thereby enabling educators to tailor teaching to students' knowledge levels to support learning (Chaiklin, 2003).

1.1. Challenges Associated with Available AI-Based Tools

The number of students relying on tools like ChatGPT is steadily increasing (Baig & Yadegaridehkordi, 2024), even though the underlying Large Language Models (LLMs) have well-known and extensively documented limitations (cf. Alkaissi & McFarlane, 2023; Claman & Sezgin, 2024). While available LLMs are becoming increasingly accurate and make fewer errors (Achiam et al., 2023), the information they provide is not always factually correct (e.g., Fukuda et al., 2025; Chow et al., 2024; Theophilou et al., 2023). Moreover, the additional information often introduced by the chatbot during conversations can not only cognitively overwhelm learners but may also deviate from the direction of learning intended by the instructors (García-López et al., 2025). In biology for instance, if a lecture covers the topic of mitosis with a focus on chromosome segregation and the molecular mechanisms involved, it might be didactically more valuable for subsequent conversations to focus on these aspects rather than shifting to meiosis as a comparison if not yet discussed in class. Therefore, the provided answers can diverge from the intended focus of the instruction (Achiam et al., 2023; García-López et al., 2025). Concomitantly, studies show that students generally wish to use generative AI tools based on course materials, as these may offer higher-quality feedback (Balabdaoui et al., 2024).

Another challenge of chatbot-based interactions for knowledge acquisition is the underlying didactic design. LLM-based chatbots generally follow a similar pattern: the user enters a question or task and receives a corresponding answer within seconds. As a result, the associated thought process is skipped (Bauer et al., 2025). Moreover, the abundance of information and the speed at which it can be accessed may lead to reduced cognitive efforts (Stadler et al., 2024) and therefore eventually in reduced learning gains (Azeem & Abbas, 2025; Kosmyna et al., 2025; Li et al., 2025; Mueller & Oppenheimer, 2014). In line with these findings, offering appropriate learning scaffolds when using tools like ChatGPT in education and actively integrating different learning modes, may be inevitable for providing an effective learning environment.

This emphasis on scaffolded chatbot use reflects principles from learning theory, particularly research on «productive failure» or «problem-solving before instruction» (Kapur, 2014; Loibl et al., 2017; Loibl & Rummel, 2014). This research suggests that the highest learning gains are achieved when students first actively engage in a problem-solving phase before the correct solution is discussed (Sinha & Kapur, 2021). If instruction is provided immediately instead, students are not expected to identify own knowledge gaps (Loibl & Rummel, 2014), to explore different potential solutions (Kapur & Bielaczyc, 2012) and, ultimately, to construct their knowledge (Bada & Olusegun, 2015), which may lead to lower learning gains. Applied to the challenge of chatbot use, this suggests that students may be deprived of the opportunity to actively engage with their own questions before a solution is presented (Bai et al., 2023; Fan et al., 2025), although being aware of such questions and having the opportunity to reflect on them were shown to enhance performance (Spitzer et al., 2025). If knowledge no longer needs to be constructed, students' conceptual understanding might diminish on a long-term when using chatbots without appropriate pedagogical design.

https://doi.org/10.3390/educsci16050812

In addition to the challenge of leveraging chatbots effectively for student learning, a further limitation arises from the fact that conventional chatbots do not grant instructors access to student interaction data, thereby restricting insights into students' learning processes. Understanding which problems students struggle with and which concepts are still unclear could help instructors not only reflect on their own teaching practices but also reveal direct ways to improve their instruction (Buck & Trauth-Nare, 2009; Leenknecht et al., 2021). Thus, there is emergent need for the development of didactically motivated open source chatbots that allow access to information on students' learning status.

1.2. Didactically Motivated Chatbots for University Lectures

In large university courses, it can be challenging to address all students' specific needs, consider their prior knowledge, educational and cultural background, and provide personalized instruction (Northedge, 2003). Intelligent tutoring systems based on machine learning, more specifically artificial intelligence, however, hold the potential to tackle precisely this challenge (Tobler & Köhler, 2025; Steenbergen-Hu & Cooper, 2014). However, setting up a system that combines a knowledge-based and controlled environment for student interactions, a didactically structured design to support learning, and lecturer-facing insights into student questions remains challenging.

When Schwarz and Bransford proposed in 1998 that learning is more effective when a meaningful activity precedes instruction (i.e., one that engages with the subject) they laid the groundwork for the current trends in learning theory: Empirical research has repeatedly shown that working on (own) problems instead of first receiving instruction significantly enhances learning, even if the correct solution is not found (Sinha & Kapur, 2021). Thereby, relevant prior knowledge is activated to help anchor new concepts more broadly and learners are supported in identifying their own knowledge gaps, thereby potentially boosting engagement and motivation. These insights are further supported by the conceptual change theory (Vosniadou, 2014), which posits that knowledge develops through the integration of new information with existing conceptual frameworks.

In contrast to the design of specific instructions or problem sets, the starting points for new learning scenarios are less clearly defined when working with chatbots as the students initiate the first step. To counter this, we propose to start the conversations with students introducing their own problems for clarification which then build the groundwork for follow-up conversations.

Yet, this challenge also entails considerable potential, as it enables the individualization of the problem-solving approach. By designing the chatbot so that students' own open questions form the starting point of the interaction, their actual problems become the central focus of learning. This is in line with research on instructional design, showing that formulating one's own questions and working through them can be an effective way to learn (e.g., Chi et al., 1994; Chin & Osborne, 2008; Dori & Herscovitz, 1999).

Moreover, questions and answers generated by students may not only help to activate prior knowledge but also to promote knowledge construction (Chi et al., 1994; Chin & Brown, 2000) and increase motivation and interest (Chin & Kayalvizhi, 2005). Thus, combining question formulation with generating own answers and generative AI-based evaluation of those responses allows personalized feedback without depriving students of the opportunity to construct knowledge independently. By continuing the conversation in the style of a Socratic dialogue, (i.e., a form of interaction that promotes critical thinking through guided questions; Wang et al., 2008; Katsara & De Witte, 2019), in which the chatbot follows up with further questions based on the students' responses, a continuous process of problem solving could be achieved that builds on the learner's prior knowledge.

Particularly in the educational context, however, it is crucial that the context and content of such an LLM-based chatbot's answer can be guided to a certain extent to counter potential hallucination or output unpredictability (Kasneci et al., 2023). One way to achieve this is a combination of human quality control with transparent explanations of LLM associated limitations and the integration of a curated knowledge base, thereby controlling that the chatbot's interactions remain reliable and aligned with the didactic goals. Since OpenAI's Assistants API (OpenAI, 2023) enables users to provide a custom knowledge base (e.g., lecture notes, scripts, or other written materials for answering student questions), integrating this API as the underlying chatbot model presents a valuable solution to tackle the problems described above.

Finally, implementing such a chatbot for a specific course should aim to remove potential barriers and ensure accessibility for all students. It is important that students can use the chatbot without performance pressure or fear of negative consequences, even if lecturers gain insights into their learning behavior (Zobel et al., 2023). By ensuring that no personal data is recorded (i.e., anonymous access) or by restricting instructors' access to data, these considerations might be mitigated. In other words, the minimal data set would include only students' inputs and the chatbot's outputs to make conversations traceable, supplemented by randomly generated identity markers and timestamps to group related interactions meaningfully.

The described chatbot operationalizes this didactic approach through a predefined instruction prompt and a structured user-interface sequence. The conversations with the AI-based tutor begin with a student-generated question and the student's own answer, before being continued through follow-up questions in a Socratic style. More concretely, students were first asked to formulate a question. If the question was clear and related to the lecture context, they were asked to provide their own initial answer. The chatbot was then instructed to provide concise feedback, avoid directly revealing answers, rely on the provided lecture materials, and refine students' reasoning through follow-up questions before providing explanations. An exemplary conversation illustrating this interaction structure is shown in Figure 1. The full instruction prompt is provided in Appendix A. The chatbot was tested before deployment to examine whether it showed the expected behavior. Yet, the consistency of this interaction pattern was not systematically quantified across all recorded conversations.

Conversation No. 13 (translated from German)	**Category**
BIOTUTOR: Hi! Which question would you like to solve today?	Conversation starter
STUDENT: What is glycosylation?	Question
BIOTUTOR: That's a good question. Now try to find an answer yourself.	Question evaluation; General feedback
STUDENT: The attachment of sugar to proteins or lipids.	Content-based answer
BIOTUTOR: Your answer is correct! Glycosylation indeed refers to the attachment of sugar to proteins or lipids, which can affect their stability, function, and interactions. It's great that you summarised that so well. What do you think might be the importance of this sugar attachment for cell communication or intracellular transport?	Question evaluation; Content feedback: rephrasing; Content feedback: addition; General feedback; Follow-up question
STUDENT: The sugars serve to identify the specific molecule?	Content-based answer
BIOTUTOR: Your answer is correct — sugar residues can act as "identity markers" that help recognise specific molecules. They also play a role in cell contact and signalling pathways. What do you think — what kinds of changes in glycosylation could affect cell function?	Question evaluation; Content feedback: rephrasing; Content feedback: addition; Follow-up question
STUDENT: I don't know	Off-content answer
BIOTUTOR: That's completely fine. Learning often starts right where we still have questions. Think about this: if a molecule's "identifier" no longer works properly because its sugar structures have changed, what might that mean for cell functions? What effects could such a change have on how cells interact?	Question evaluation; General feedback; Question rephrasing
STUDENT: That molecules end up in the wrong place or are processed incorrectly?	Content-based answer
BIOTUTOR: That's correct — altered glycosylation can indeed cause molecules to reach the wrong location or be processed incorrectly, which can impair cell function. Well observed! How could such a malfunction disturb communication between cells or metabolism in the organism? Think about what happens when signals aren't properly transmitted.	Question evaluation; Content feedback: rephrasing; Content feedback: addition; General feedback; Follow-up question; Content feedback: assistance
STUDENT: No idea.	Off-content answer
BIOTUTOR: No problem; it's completely normal that not everything is clear yet. Try to imagine what happens when important molecules are mislocalised. How could changes in cell communication and signal transmission have an effect? For example, what happens to the system when the "messages" don't get through? What consequences could that have for the organism?	Question evaluation; General feedback; Question rephrasing; Content feedback: assistance
STUDENT: Depending on the amount and type of cell — maybe the cell dies?	Content-based answer
BIOTUTOR: That's an interesting idea. Indeed, changes in glycosylation can have negative effects that may even trigger cell death. This disrupts normal cell communication and function, which in some cases can promote disease. In your opinion, how could such a process lead to functional disorders in a tissue?	Question evaluation; Content feedback: rephrasing; Content feedback: addition; Follow-up question
STUDENT *ends the conversation*	

Figure 1. Exemplary conversation with the bioTutor. Student inputs and bioTutor outputs are grouped in different categories to provide an overview of the course of the conversation.

2. Technical Architecture of the Chatbot and Lecturer Dashboard

2.1. Developing a Chatbot for University Classrooms

Based on these considerations, we developed an AI-based chatbot (bioTutor) that follows the described didactic design and provides students with an opportunity to individually test and deepen their understanding of lecture content. In addition, we aimed to create a technically simple solution to ensure transferability to other courses without significant adaptation efforts. Therefore, the entire software is published as open source code and can be freely accessed and adjusted to enable researchers or educational developers to advance or implement such a chatbot. Programmed as an RShiny application (Chang et al., 2024), the chatbot can be hosted either on local servers or on commercially available platforms (such as Posit's Shinyapps server; https://shinyapps.io, accessed on 20 May 2025), making it possible to provide a customizable standalone solution tailored to individual use cases.

The chatbot conversations are powered by OpenAI's Assistant API (OpenAI, 2023), which provides a specifically instructed assistant that functions as a conversational agent. The Assistant API further allows for direct integration of custom documents as a knowledge base, specific instruction on chatbot behavior, and selection of generative pre-trained transformer (GPT) models with different pricing ranges as the underlying LLM.

A further challenge was to make student questions and the corresponding LLM responses available for later analysis by instructors. Thus, this data needed to be stored in an easily accessible and usable format. This was achieved by saving both the input and output of the LLM during processing to an external data storage. In the present case, we opted for Google Sheets (Google, 2025), as these databases can be easily created, edited, and accessed via an API. However, the free version of Google Sheets has some limitations, such as a slower API and rate limits, which may take effects when working with large student cohorts simultaneously. Yet, if working with asynchronous conversations, Google Sheets proved to be a valuable solution.

To allow reconstruction of individual conversations from the saved questions and answers, each chatbot session is assigned a randomized unique ID (UID) and for each interaction the timestamp is recorded. No additional data is collected, in order to maintain student anonymity. Consequently, there is no login required and the chatbot is accessible as a web application via a shared link. A schematic overview of the chatbot's software architecture is shown in Figure 2, its implementation as a web application is shown in Figure 3 (left).

2.2. A Dashboard Informing Lecturers on Student Learning

Being able to record students' conversations, which is typically not possible with commercially available chatbots, enables further data analysis to gain insights into learning and teaching, e.g., identifying unclear points and addressing these by adjusting instruction based on students' questions. To support this, a lecturer dashboard was developed that is directly linked to the chatbot's conversations.

It provides AI-generated weekly summaries of challenging topics and analyses of conversations, offering instructors an initial overview of potentially recurring difficulties. In addition, usage statistics and access to anonymized raw data are provided, allowing more in-depth analysis of the conversations when needed.

The dashboard's technical implementation allows setting up a comparable application without much technical knowledge. We chose the RShiny programming framework for the dashboard (Chang et al., 2024) to develop a data-driven and responsive application that can be hosted on a server with minimal effort. The dashboard is linked to the same Google Sheet as the chatbot and connected to OpenAI's API for using GPT-based LLMs.

Unlike the chatbot, however, the dashboard is password-protected and not tied to a specific course. Thus, different course-specific dashboards can be loaded depending on the user login, simplifying the administrative provision of the dashboard for multiple courses and lecturers.

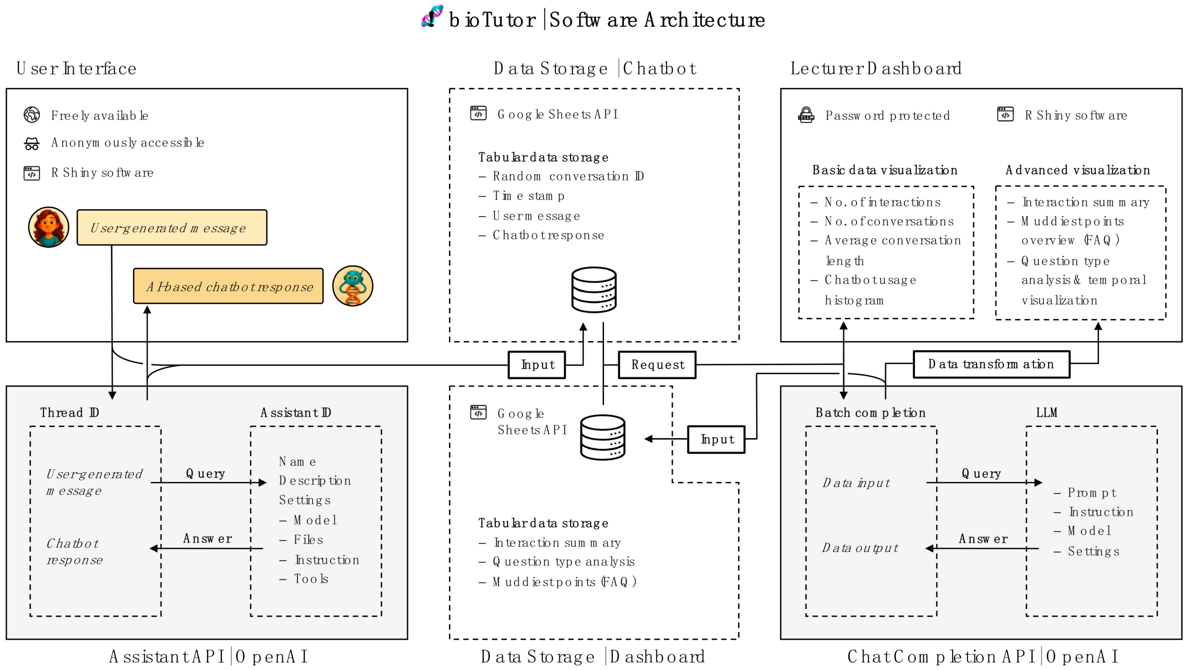

Figure 2. Software architecture and dependencies of the bioTutor.

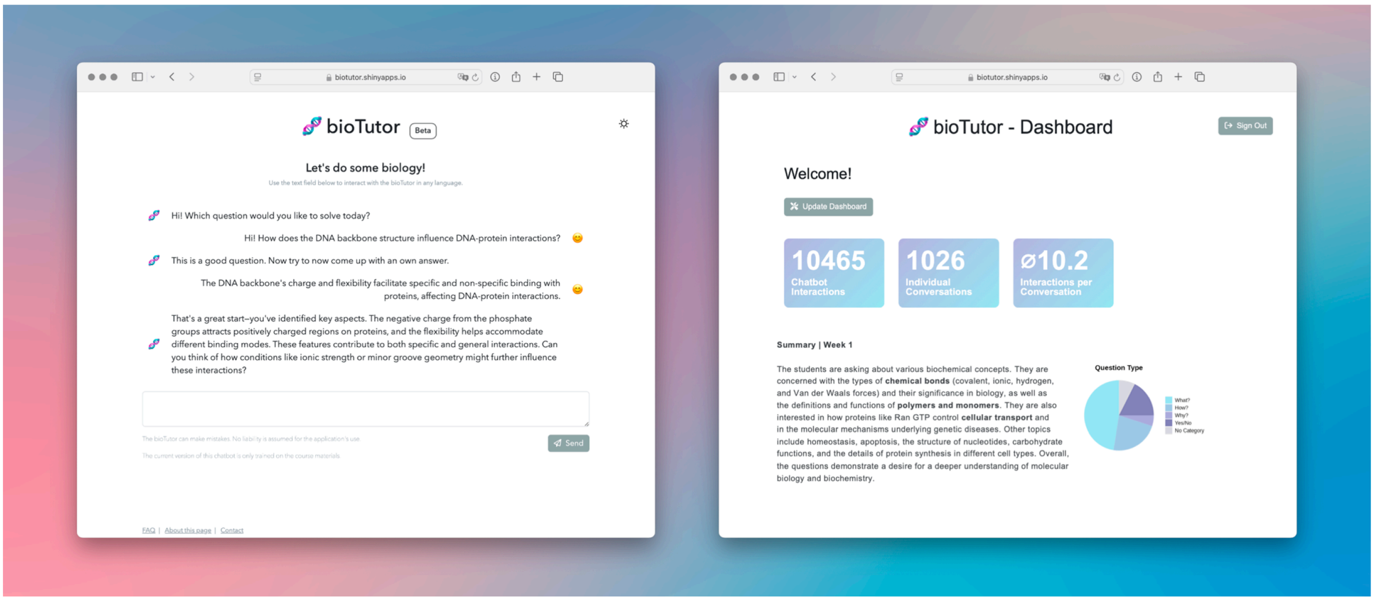

Figure 3. User interface of the bioTutor chatbot (**left**) and lecturer dashboard (**right**).

Upon logging in, the conversation data is directly retrieved from the chatbot-linked Google Sheet. Statistical data analyses are calculated automatically, including the number of questions, number of conversations, or the average conversation length. In-depth qualitative analyses must be manually triggered with a button, prompting a batch of diverse analyses of questions and answers via OpenAI's Chat Completion (OpenAI, 2023). By doing so, the dashboard generates weekly summaries of all interactions, a list of the most frequently asked questions, and an analysis of question types. The question type analysis enables the identification of the nature of questions, for example, whether students asked primarily about definitions or about biological mechanisms (Tawfik et al., 2020).

An overview of the dashboard's software architecture is shown in Figure 2. Figure 3 (right) shows a screenshot of the integration of the dashboard as a web application. The code for the dashboard and chatbot is available on GitHub.

3. Using the bioTutor to Support Life Science Education

The developed applications described in Section 2 were introduced as part of a large lecture with 407 students at a highly ranked technical university in Switzerland. At the beginning of the semester, the chatbot «bioTutor» and its functionalities were presented to the students by the lecturer, and the link to access it was provided digitally in their Learning Management System (LMS). The students were free to use the bioTutor and its use was anonymous, voluntary, and not tied to passing the course. Students were informed in advance and reminded when using the chatbot that the conversations in the bioTutor would be recorded anonymously to gain insights into their learning process and used for research purposes. Thereby, only student inputs, chatbot outputs, timestamps, and randomly generated session IDs were stored. Access to these data was restricted to the research group and educators of the course. Since each interaction with the bioTutor automatically generated an anonymous entry in the conversation dataset, students could not opt out of recording while using the tool. In addition, students were anonymously surveyed about their use of the chatbot as part of the end-of-semester feedback of the lecture. The university's ethics committee approved all data collections and surveys prior to their implementation (23 ETHICS-277).

3.1. Methods

3.1.1. Participants

The students who had access to the chatbot were enrolled in either the Health Sciences and Technology or Human Medicine study program and inscribed to an introductory biology lecture. Of the 316 students who finished the end-of-semester feedback, 94 students (29.7%) stated that they had used the chatbot. Of these 94 students, 73 completed the questionnaire and were therefore included in the data analysis. Data collection was completely anonymous and no demographic data about the participant population was collected[1].

3.1.2. Procedure

At the end of the semester, students had the opportunity to complete an anonymous questionnaire about the course. Part of this questionnaire dealt with the use and implementation of the bioTutor. Answering the questions was voluntary and there was no compensation. Students could complete the questionnaire independent of location or time. The submission deadline for the questionnaire was the end of the semester; however, the bioTutor could still be used until the end of the exam phase (i.e., two months later). Thus, it cannot be ruled out that not all students who used the chatbot also gave feedback.

3.1.3. Education-Adapted Technology Acceptance Model (edTAM)

To analyze how students accepted and used the bioTutor, the Technology Acceptance Model 2 (TAM2; Venkatesh & Davis, 2000) was theoretically extended and adapted for educational contexts, resulting in the education-adapted Technology Acceptance Model (edTAM; Figure 4). The structure of the edTAM follows TAM2 and adjusts it for technologies in the educational context to reflect the usefulness of a technology for learning processes. As such, the edTAM items are concerned with factors ultimately influencing students' chatbot usage behavior. Perceived usefulness and perceived ease of use were measured using the UMUX-Lite questionnaire (Lewis et al., 2013). Learning relevance encompasses the subjective assessment of using the technology for one's own learning success. Output quality describes the subjective perception of the usefulness of the technology's output, while result demonstrability deals with the extent to which the results of using the technology become tangible and communicable.

Table 1. Overview of the questionnaire items.

	Item	Explanation
1.1	The bioTutor chatbot was straightforward to use.	Adapted from the UMUX-Lite questionnaire1. Item concerning perceived ease of use.
1.2	The capabilities of this chatbot met my requirements.	Adapted from the UMUX-Lite questionnaire1. Item concerning perceived usefulness.
2.1	I would have appreciated having a similar chatbot for exercises in other lectures.	Item concerning behavioral intention.
2.2	The follow-up questions posed by the chatbot were helpful in deepening my understanding of the topic.	Item concerning learning relevance.
2.3	The chatbot's answers were useful in clarifying my outstanding questions.	Item concerning output quality.
2.4	The time invested in learning with the chatbot did not pay off.	Item concerning result demonstrability (negatively formulated).

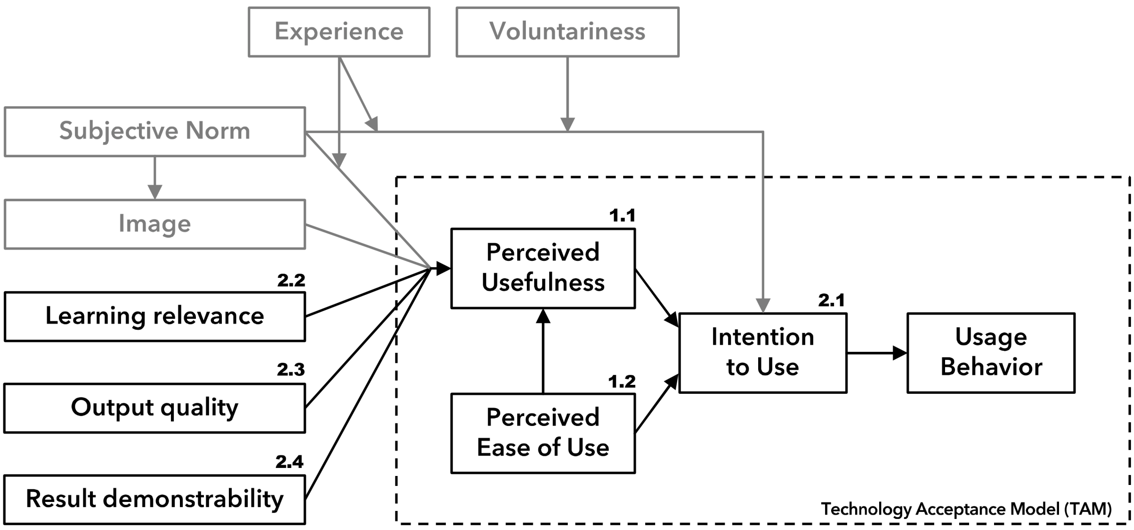

Figure 4. Education-adapted Technology Acceptance Model. Small numbers indicate with which item the individual aspects of the edTAM were examined (cf. Table 1).

For assessing learning relevance (the personal estimation of relevance of the technology for own learning gains), output quality (the subjective assessment of the chatbot's analysis quality and follow-up conversation), and result demonstrability (whether the benefits of the technology are clearly visible or tangible), one item was used each (Table 1). Although individual items may not cover the entire dimension of a theoretical construct (Diamantopoulos et al., 2012), they still provide initial insights and allow for efficient surveys (Bergkvist & Rossiter, 2007). However, as these constructs were each assessed with a single item, they should be interpreted as exploratory indicators rather than robust construct measurements. Further validation with multi-item scales is therefore needed to confirm construct validity and reliability.

The TAM2's subjective norm, which includes the experience and voluntariness of technology use, as well as the user's image, were not analyzed in this context as using the tool was non-mandatory, anonymous, and not tied to grades or public visibility among peers. Therefore, the edTAM framework used in this study should be understood as a context-specific adaptation rather than a complete validation of all TAM2 components.

3.1.4. Data Analysis

All items of the edTAM questionnaire were answered on a 5-point Likert scale. An overview of the items used is shown in Table 1. The Likert-scale responses were transformed into numeric values and subsequently analyzed descriptively. The negatively formulated item 2.4 was reverse-coded before analysis. No missing data were present in the analyzed questionnaire dataset.

For the exploratory correlation analysis between questionnaire items, Likert responses were treated as approximately interval-scaled and Pearson correlations were calculated using the *rcorr* function from the *Hmisc* (Harrell, 2025) package. Statistical significance was evaluated using conventional thresholds (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$). No correction for multiple testing was applied, as the analysis was exploratory and descriptive. These data and the chatbot usage data were analyzed and visualized using the RStudio software environment (R Core Team, 2016, Version 2025-09-02+418).

3.2. Empirical Investigation and Analysis of Student Evaluations of the bioTutor

The analysis of the questionnaire results showed that students were generally satisfied with the bioTutor, both in terms of usability and perceived usefulness (combined Median = 4.5, Figure 5). More than 85% of students indicated that using the bioTutor was straightforward, and more than 75% stated that the bioTutor adequately met their requirements. As such, one student stated, "I find it very useful that it asks follow-up questions, as this allows one to test one's knowledge." These findings are also reflected in the data on learning relevance, output quality, and result demonstrability: 77% of students reported that the bioTutor's follow-up questions were helpful in deepening the learned material, and nearly 70% agreed that the responses were helpful in clarifying open questions, even though the chatbot itself was designed to mainly support students in finding answers but not directly providing them. Only 23% of students indicated that learning with the chatbot was not worth the invested time. Possible reasons for this might include the didactic design that enforces active examination with the topic rather than providing fast solutions or technical limitations including server outages or longer waiting times for chatbot responses during high demand. Qualitative analysis of student feedback on open questions regarding the bioTutor partially supports this hypothesis. For example, one student stated that "[. . .] for some questions I find this [asking of explanations] somewhat unnecessary and would prefer to simply receive an answer," and another student stated that "the follow-up questions are not useful when one cannot answer one's own question and the chatbot asks for the solution."

The finding that, despite the widespread availability of other free-to-use chatbots like ChatGPT or Gemini, a large majority of over 80% of students disclosed their preference to have similar tools in other lectures further supports the conclusion that the implementation of a lecture-based and didactically motivated chatbot addresses a significant need in current university education.

https://doi.org/10.3390/educsci16050812

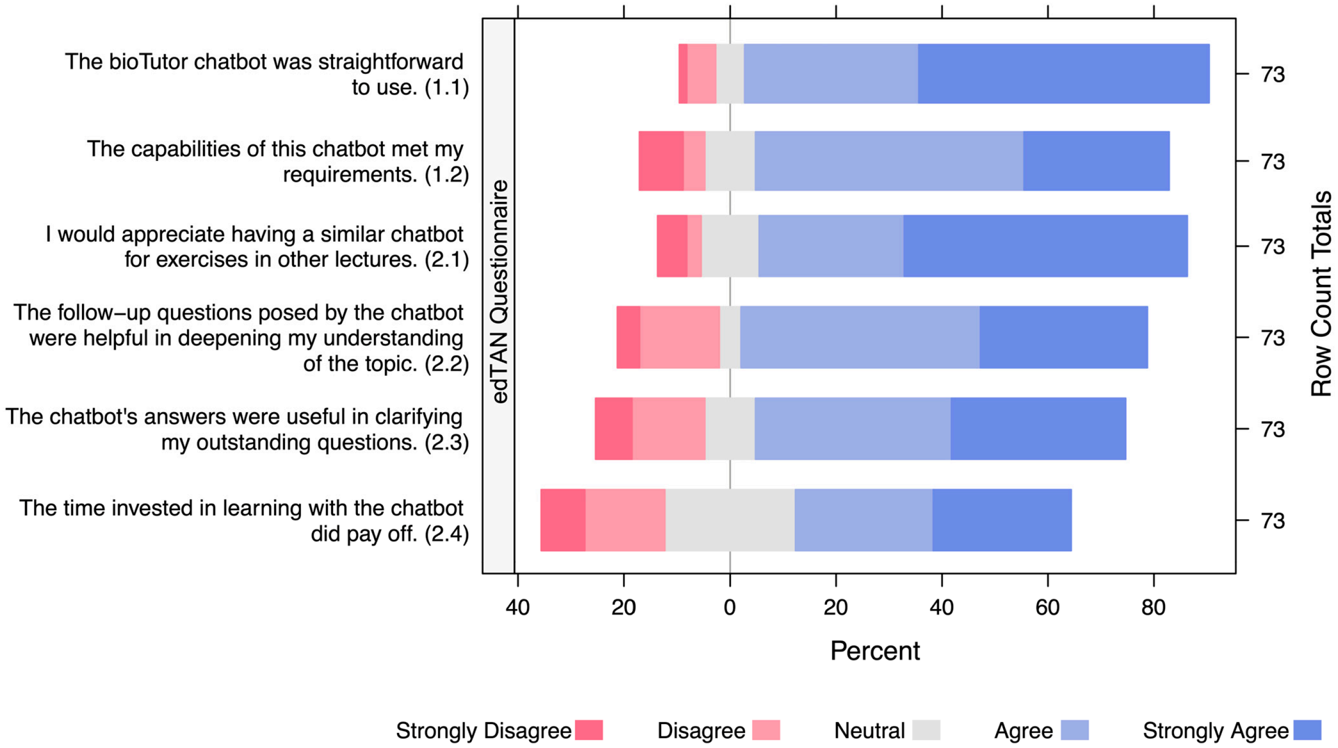

Figure 5. Descriptive questionnaire item analysis. Numbers in brackets after the items display the item number (cf. Table 1). Neutral values are centered around the null value (cf. Heiberger & Robbins, 2014). Note that item 2.4 has been inverted for better readability of the graph.

An additional exploratory correlation analysis of the individual item values according to the edTAM framework showed that the majority of the selected questionnaire items were significantly correlated with each other, indicating preliminary coherence between the selected items (Table 2). However, these correlations do not constitute validation of the edTAM framework, and statistical evaluations of the edTAM questionnaire regarding construct validity and reliability need to be confirmed in future studies and larger cohorts. In summary, the developed chatbot bioTutor demonstrates high acceptance among students, which can be attributed to its usability and implementation functionality.

Table 2. Correlation analysis of questionnaire items.

Item	1.1	1.2	2.1	2.2	2.3
1.2	0.62 ***	–	–	–	–
2.1	0.70 ***	0.58 ***	–	–	–
2.2	0.66 ***	0.70 ***	0.66 ***	–	–
2.3	0.55 ***	0.81 ***	0.59 ***	0.69 ***	–
2.4	0.21 (ns)	0.23 (ns)	0.08 (ns)	0.27 *	0.36 **

Note. Values indicate correlation coefficient r. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; ns = not significant.

3.3. Evaluation and Discussion of Students' Chatbot Usage

To analyze the chatbot usage, the time period during which the tool was available was divided into three phases: the semester with in-person lectures and exercises (13 weeks), the learning and exam preparation phase after the end of lectures (7 weeks), and the period after the exam until the tool was deactivated (3 weeks). The chatbot bioTutor was introduced to the students in the second week of the semester; no lectures were held in the first week. The tool was presented for five minutes during an official lecture hour without compulsory attendance and linked afterwards on the course's LMS, along with a guide explaining the structure of the tool and its intended use. Students were not further reminded of the existence of the tool, and no extrinsic motivation for its use was provided (e.g., no integration into lectures, grade bonuses, or specific exercises).

During the entire 23-week period in which the bioTutor was available, 10,465 student-formulated interactions with the chatbot were recorded. These interactions included questions asked by students, student answers to their own questions, as well as responses to and interactions with the bioTutor's questions and statements. These interactions occurred in 1026 individual conversations, which resulted in a median of 6 student interactions per conversation (Mean ± SD = 10.2 ± 11.6).

Since data were recorded anonymously, it was not possible to determine how many individual users actually used the chatbot. However, the end-of-semester questionnaire (Section 3.1.1) provides a complementary self-report estimate: approximately 30% of the students who completed the course feedback reported having used the bioTutor. This value should not be interpreted as a verified usage rate for the full cohort, but only as the proportion of feedback respondents who self-reported chatbot use. Combining this self-reported number with the log data suggests an average of approximately 11 conversations per self-reported user. However, this estimate is descriptive only, as the anonymous usage data cannot be linked to individual questionnaire responses.

The distribution of usage data displaying when students used the bioTutor most intensively (Figure 6), showed that during the semester, the usage remained relatively constant at a Median of 4 conversations per day (Mean ± SD = 5.4 ± 3.8; without outliers outside of the 1.5 IQR). At two timepoints, the usage spiked, coinciding with the mid- and end-of-semester formative assessments (learning objective checks). After the winter break and especially shortly before the summative exam in mid-February, the usage increased significantly and remained high for a longer period. Interestingly, the chatbot was still used in March, even though the associated lecture had already ended and students were aware that the materials underlying the chatbot's knowledge foundation were not specific to any other course.

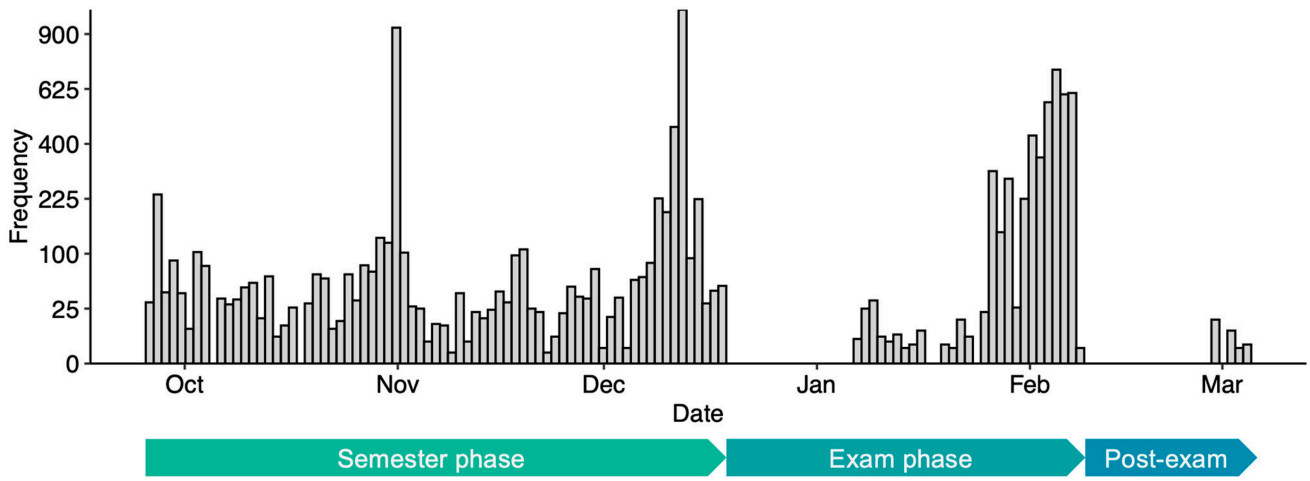

Figure 6. bioTutor Chatbot Usage Analysis. The usage data of the chatbot is displayed during the semester, the exam phase, and the time after the exam. The chatbot was not used during the Christmas holiday break and in the period after the students took the exam and before the next semester started (mid-February to end of February). The y-axis is plotted on a square-root scale to improve visibility; the x-axis shows the number of individual interactions with the chatbot.

The results, thus, indicate that the bioTutor was already used during the semester for preparing or reviewing lecture content but more frequently applied in the preparation for mid-terms or exams. Nevertheless, students appeared to perceive benefits of the chatbot for their own learning and may have used it to review or engage with contents from other lectures, supporting the claim of its perceived usefulness.

4. General Discussion

Recent research on the use of generative AI in education has highlighted both its potential to enhance student learning (Heeg & Avraamidou, 2023; Vieriu & Petrea, 2025) as well as its associated challenges, such as limited pedagogical grounding and contextualization (Alkaissi & McFarlane, 2023; Claman & Sezgin, 2024), or reduced cognitive processing depth or limited transfer knowledge gain (Bauer et al., 2025). The present study extends this line of work by introducing and evaluating an open-source framework for a generative AI chatbot aiming to address these challenges through didactically informed conversations. Thus, the system combines a chatbot based on a lecture-specific knowledge base with principles of instructional design to provide a potential learning support. Moreover, it includes a learning analytics dashboard that may provide educators with aggregated and potentially actionable insights into student questions and recurring difficulties. The results of its implementation in a large undergraduate classroom indicate high student acceptance, perceived usefulness, and usage during learning and exam preparation phases, suggesting that pedagogically grounded digital support may be perceived as useful for learning and exam preparation.

4.1. Student Acceptance, Usage Patterns, and Insights for Lecturers

Despite being entirely voluntary and not incentivized, 94 of the 316 students who completed the end-of-semester feedback reported having used the chatbot. This indicates that a substantial subgroup of the feedback respondents engaged with the tool, although the anonymous design prevents verification of the exact number of individual users in the full cohort. Their usage thereby peaked during key assessment phases, such as the midterm or the final exam, suggesting that students perceived the tool as a meaningful measure to support their individual learning process and exam preparation.

The chatbot's high usability and usefulness ratings, as measured by the herein proposed education-adapted technology acceptance model (edTAM), underline its perceived relevance. This is further supported by the finding that more than 75% of the students who used the chatbot rated its answers as helpful and that its follow-up questions, which are tailored to students' prior knowledge, were perceived as supporting their knowledge construction. These results are also in line with prior research on ITS in higher education highlighting the importance of adaptive and dialogue-based learning support (Mirata et al., 2020). Furthermore, considering that over 80% of the chatbot users stated their interest in having similar tools in other courses indicates that the students valued the generative AI-based chatbot potentially due to its didactic design and tailoring to specific course content. Moreover, the didactic design of the tool was intended to support processes related to transferable competencies that are essential in contemporary higher education, such as critical thinking and analytical skills (Köhler & Tobler, 2025), by motivating students to identify their own knowledge gaps, formulate their own questions, and critically reflect on their answers. These competencies were not directly measured in the present study. Thus, the enforced requirement that students formulate their own questions, propose answers, and then engage in a follow-up Socratic dialogue distinguishes the bioTutor from chatbots like ChatGPT which typically provide direct answers without motivating students to first reflect on their understanding (cf. Fan et al., 2025).

Importantly, the bioTutor was not intended to replace student–lecturer or student–teaching assistant interactions but rather to provide an additional layer of support that is always available and is associated with a lower inhibition threshold to ask questions one might not otherwise dare to ask.

The lecturer dashboard provided additional insights into the nature of student questions and their development over time. These aggregated and automatically analyzed chatbot interaction data may inform teaching practice by highlighting recurring questions, challenging topics, and potential knowledge gaps. Whether and how lecturers used these insights to adapt their teaching remains to be investigated.

4.2. Learnings and Practical Implications

To ensure successful use of the chatbot by students, it is worthwhile to introduce the chatbot to the students early and in person. Thereby, time should be dedicated to clearly explaining the purpose of the bioTutor, especially focusing on its functionalities and its didactic design. For instance, someone looking for a tool providing quick answers to an open question or trying to use materials to generate multiple-choice questions will not achieve their goal with the bioTutor and thus, might be disappointed by its use. In addition to the in-person explanation, a digital guide could be provided so that students can inform themselves about the functionalities at any time. Furthermore, a list of potential conversation starter questions could be prepared to ease the first interactions with the new tool.

Another aspect that is crucial for the development of a lecture-based chatbot is the requirement for diligent material preparation and pre-processing. For instance, complicated figures in slide sets require descriptions and lecture transcripts might require human revision prior to be used as content-grounding material (cf. Santos et al., 2022). In fact, one of the main challenges in designing the content-specific bioTutor was the preparation of input material that comprehensively matched the lecture content.

4.3. Limitations and Future Work

Whereas the present study provides initial insights into students' acceptance and usage behavior when interacting with the bioTutor, several limitations must be considered. As such, the study does not provide evidence on whether the bioTutor influenced learning outcomes, since a quasi-experimental setting that would have allowed comparison between different student groups was not feasible in the present course context. Accordingly, the present study cannot isolate whether these evaluations were specifically attributable to its pedagogical design, as no control or comparison group was included. Excluding some students from a potential learning opportunity for exam preparation might have created inequality, which we deliberately aimed to avoid. Further, no demographic data were collected from the questionnaire respondents. Thus, it remains unclear whether the bioTutor was perceived differently by students with different backgrounds, prior knowledge, or language proficiency. Future studies may therefore integrate the system into appropriate experimental settings to investigate potential effects on learning, also in relation to different demographic variables and prior knowledge.

In addition, the present study did not investigate whether or how lecturers used the information provided by the dashboard to adapt their teaching. Thus, the lecturer dashboard should be understood as a prototype for aggregated instructional insight rather than as a fully validated learning analytics tool. Although the dashboard provides AI-generated summaries of student questions, recurring topics, question types, and usage patterns, the accuracy, usefulness, and pedagogical reliability of these summaries were not systematically evaluated in the present study. Future studies should therefore examine how instructors interpret and use such dashboard-based information, whether it supports instructional decision-making, and how the quality of AI-generated summaries can be evaluated, for example through expert ratings or comparisons with human-coded conversation analyses.

Moreover, adapting the bioTutor to other courses, disciplines, and curricula could provide additional insights into how to adjust the chatbot and lecturer dashboard to better support learning and teaching. Further, whereas the herein presented edTAM framework represents an exploratory application, it should be further validated, as several constructs were assessed with single items only.

Finally, the field of machine learning, particularly in the context of generative AI, is evolving rapidly (Song & Wang, 2020). Hence, it remains a challenge to continuously analyze and examine which tools or options are best suited to achieve one's own didactic and pedagogical goals. Continuous monitoring and software adjustments may therefore be required to respond to technological developments, such as switching from the Assistant API to the Response API (OpenAI, 2026). Future work should therefore also address the long-term maintainability and adaptability of the system.

4.4. Conclusions

The bioTutor represents one possible low-threshold approach to implementing a didactically structured chatbot (constructivist dialogue with follow-up questions and prior knowledge adaptation) that also collects teaching-relevant insights from anonymized student interactions. As an open-source and scalable system, the bioTutor and lecturer dashboard can be adapted to different educational contexts with comparatively simple technical means. Finally, the developed edTAM framework for assessing the acceptance of technological innovations in education showed promising initial results, although further validation is required in order to confirm its reliability and validity.

Author Contributions: Conceptualization: S.T. and K.K.; Data curation: S.T.; Formal Analysis: S.T.; Funding acquisition: S.T. and K.K.; Investigation: S.T. and K.K.; Methodology: S.T.; Project administration: S.T.; Resources: S.T.; Software: S.T.; Supervision: S.T., K.K.; Validation: S.T. and K.K.; Visualization: S.T.; Writing—original draft: S.T. and K.K.; Writing—review & editing: S.T. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the ETH Innovedum Fund (approval number: 4453).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of ETH Zurich, Switzerland (23 ETHICS-277, approved on 10 April 2024).

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

Data Availability Statement: Data Availability Statement: The data supporting the findings of this study consist of anonymized student-chatbot interaction data and survey responses. These data are not publicly available but can be made available upon reasonable request. The source code for the bioTutor and lecturer dashboard is openly available at: https://github.com/samueltobler/bioTutor/ (accessed on 18 February 2026).

Acknowledgments: We thank Olivia Bossert for her support with the preparation of material for the chatbot's knowledge base; and Robin Bauknecht and Jasmin Frei for their feedback during the development of bioTutor. We also thank all the student teaching assistants supervising in the related course for putting the tool through its paces and providing valuable feedback, as well as all undergraduate students for their participation. We used OpenAI's ChatGPT-5.2 to improve the manuscript's grammar and clarity.

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

Appendix A

The following instruction prompt was used to guide the behavior of the bioTutor:

As a friendly Socratic biology teacher, evaluate the student's input based on factual correctness, providing concise and relevant feedback. Encourage students expressing uncertainty while refraining from revealing the answer. Do not directly tell the answer, even if students ask you to provide it. Rather motivate them to think again critically. Craft follow-up questions to deepen understanding, using information from provided files without introducing terms not present in them. Use natural conversational language and do not add references and do not cite files.

Steps:

1. Evaluate the student's input for factual accuracy.
2. Provide concise, encouraging feedback and shortly elaborate on the concept.
3. Create a follow-up question based on information in the provided files.
4. Avoid using terms not found in the files or referencing file names in your response.

Output Format:

- Limit responses to 100 words.
- Use natural conversational language and no emojis.
- Use the same language as the user.

Notes:

- Encourage students expressing uncertainty without providing direct answers.
- Remove any file name or document reference numbers from responses.

Note

[1] In a prior research project, a similar student cohort from the same lecture was recruited (Tobler et al., 2022). In the latter, the demographic data indicated that the students were 19.7 $\pm$ 1.6 years old (mean $\pm$ SD). 71.6% of the participants were female, 62.2% of the students had a STEM-related background from prior education.

References

Abdaljaleel, M., Barakat, M., Alsanafi, M., Salim, N. A., Abazid, H., Malaeb, D., Mohammed, A. H., Hassan, B. A. R., Wayyes, A. M., Farhan, S. S., El Khatib, S., Rahal, M., Sahban, A., Abdelaziz, D. H., Mansour, N. O., AlZayer, R., Khalil, R., Fekih-Romdhane, F., Hallit, R., . . . Sallam, M. (2024). A multinational study on the factors influencing university students' attitudes and usage of ChatGPT. *Scientific Reports*, *14*(1), 1983. [CrossRef]

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., . . . McGrew, B. (2023). Gpt-4 technical report. *arXiv*, arXiv:2303.08774.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), e35179. [CrossRef] [PubMed]

Azeem, S., & Abbas, M. (2025). Personality correlates of academic use of generative artificial intelligence and its outcomes: Does fairness matter? *Education and Information Technologies*, *30*(13), 18131–18155. [CrossRef]

Bada, S. O., & Olusegun, S. (2015). Constructivism learning theory: A paradigm for teaching and learning. *Journal of Research & Method in Education*, *5*(6), 66–70.

Bai, L., Liu, X., & Su, J. (2023). ChatGPT: The cognitive effects on learning and memory. *Brain-X*, *1*(3), e30. [CrossRef]

Baig, M. I., & Yadegaridehkordi, E. (2024). ChatGPT in the higher education: A systematic literature review and research challenges. *International Journal of Educational Research*, *127*, 102411. [CrossRef]

Balabdaoui, F., Dittmann-Domenichini, N., Grosse, H., Schlienger, C., & Kortemeyer, G. (2024). A survey on students' use of AI at a technical university. *Discover Education*, *3*(1), 51. [CrossRef]

Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review*, *37*(2), 45. [CrossRef]

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*(2), 175–184. [CrossRef]

Buck, G. A., & Trauth-Nare, A. E. (2009). Preparing teachers to make the formative assessment process integral to science teaching and learning. *Journal of Science Teacher Education*, *20*(5), 475–494. [CrossRef]

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's Educational Theory in Cultural Context*, *1*(2), 39–64.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web Application Framework for R* (R package version 1.8. 1.9001) [Computer software]. Available online: https://cran.r-project.org/web/packages/shiny/index.html (accessed on 20 March 2026).

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. [CrossRef]

Chin, C., & Brown, D. E. (2000). Learning deeply in science: An analysis and reintegration of deep approaches in two case studies of grade 8 students. *Research in Science Education*, *30*(2), 173–197. [CrossRef]

Chin, C., & Kayalvizhi, G. (2005). What do pupils think of open science investigations? A study of Singaporean primary 6 pupils. *Educational Research*, *47*(1), 107–126. [CrossRef]

Chin, C., & Osborne, J. (2008). Students' questions: A potential resource for teaching and learning science. *Studies in Science Education*, *44*(1), 1–39. [CrossRef]

Chow, R., Hasan, S., Zheng, A., Gao, C., Valdes, G., Yu, F., Chhabra, A., Raman, S., Choi, J. I., Lin, H., & Simone, C. B., II. (2024). The accuracy of artificial intelligence ChatGPT in oncology examination questions. *Journal of the American College of Radiology*, *21*(11), 1800–1804. [CrossRef]

Claman, D., & Sezgin, E. (2024). Artificial intelligence in dental education: Opportunities and challenges of large language models and multimodal foundation models. *JMIR Medical Education*, *10*(1), e52346. [CrossRef] [PubMed]

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, *40*(3), 434–449. [CrossRef]

Dori, Y. J., & Herscovitz, O. (1999). Question-posing capability as an alternative evaluation method: Analysis of an environmental case study. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *36*(4), 411–430. [CrossRef]

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, *56*(2), 489–530. [CrossRef]

Fukuda, H., Morishita, M., Muraoka, K., Yamaguchi, S., Nakamura, T., Habu, M., Yoshioka, I., Awano, S., & Ono, K. (2025). Evaluating the accuracy and performance of ChatGPT-4o in solving Japanese national dental technician examination. *International Dental Journal*, *75*(4), 100847. [CrossRef] [PubMed]

García-López, I. M., González, C. S. G., Ramírez-Montoya, M. S., & Molina-Espinosa, J. M. (2025). Challenges of implementing ChatGPT on education: Systematic literature review. *International Journal of Educational Research Open*, *8*, 100401. [CrossRef]

GitHub. (n.d.). *GitHub copilot* [Software]. Available online: https://github.com/features/copilot (accessed on 18 February 2026).

Google. (2025). *Google sheets API overview*. Google Developers. Available online: https://developers.google.com/sheets/api (accessed on 18 February 2026).

Google. (2026). *Gemini API overview*. Google AI for Developers. Available online: https://ai.google.dev/gemini-api (accessed on 18 February 2026).

Harrell, F. (2025). *Hmisc: Harrell miscellaneous* (R package version 5.2-4) [Computer software]. CRAN. Available online: https://cran.r-project.org/web/packages/Hmisc/index.html (accessed on 20 March 2026).

Heeg, D. M., & Avraamidou, L. (2023). The use of Artificial intelligence in school science: A systematic literature review. *Educational Media International*, *60*(2), 125–150. [CrossRef]

Heiberger, R. M., & Robbins, N. B. (2014). Design of diverging stacked bar charts for Likert scales and other applications. *Journal of Statistical Software*, *57*(5), 1–32. [CrossRef]

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, *38*(5), 1008–1022. [CrossRef]

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, *21*(1), 45–83. [CrossRef]

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. [CrossRef]

Katsara, O., & De Witte, K. (2019). How to use socratic questioning in order to promote adults' self-directed learning. *Studies in the Education of Adults*, *51*(1), 109–129. [CrossRef]

Kiryakova, G., & Angelova, N. (2023). ChatGPT—A challenging tool for the university professors in their teaching practice. *Education Sciences*, *13*(10), 1056. [CrossRef]

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv*, arXiv:2506.08872.

Köhler, K., & Tobler, S. (2025). Yes, we can: Empowering 21st century skills in a large introductory classroom through project-based learning. *ETH Learning and Teaching Journal*, *5*(1).

Leenknecht, M., Wijnia, L., Köhlen, M., Fryer, L., Rikers, R., & Loyens, S. (2021). Formative assessment as practice: The role of students' motivation. *Assessment & Evaluation in Higher Education*, *46*(2), 236–255.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When there's no time for the SUS. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2099–2102). ACM.

Li, S., Liu, J., & Dong, Q. (2025). Generative artificial intelligence-supported programming education: Effects on learning performance, self-efficacy and processes. *Australasian Journal of Educational Technology*, *41*(3), 1–25. [CrossRef]

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, *29*(4), 693–715. [CrossRef]

Loibl, K., & Rummel, N. (2014). Knowing what you don't know makes failure productive. *Learning and Instruction*, *34*, 74–85. [CrossRef]

Mirata, V., Hirt, F., Bergamin, P., & van der Westhuizen, C. (2020). Challenges and contexts in establishing adaptive learning in higher education: Findings from a Delphi study. *International Journal of Educational Technology in Higher Education*, *17*, 32. [CrossRef]

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, *25*(6), 1159–1168. [CrossRef]

Northedge, A. (2003). Rethinking teaching in the context of diversity. *Teaching in Higher Education*, *8*(1), 17–32. [CrossRef]

OpenAI. (2023). *New models and developer products announced at DevDay*. Available online: https://openai.com/index/new-models-and-developer-products-announced-at-devday/ (accessed on 18 February 2026).

OpenAI. (2026). *ChatGPT* (large language model) [Software]. Available online: https://chat.openai.com/ (accessed on 18 February 2026).

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available online: https://www.R-project.org (accessed on 20 March 2026).

Santos, G. A., De Andrade, G. G., Silva, G. R. S., Duarte, F. C. M., Da Costa, J. P. J., & de Sousa, R. T. (2022). A conversation-driven approach for chatbot management. *IEEE Access*, *10*, 8474–8486. [CrossRef]

Sharma, S. (2024). Benefits or concerns of AI: A multistakeholder responsibility. *Futures*, *157*, 103328. [CrossRef]

Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research*, *91*(5), 761–798. [CrossRef]

Song, P., & Wang, X. (2020). A bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years. *Asia Pacific Education Review*, *21*(3), 473–486. [CrossRef]

Spitzer, M. W. H., Moeller, K., Richter, E., Frede, L., & Bardach, L. (2025). A red button does it all: A red button nudges students to close their knowledge gaps in an intelligent tutoring system. *Learning and Instruction*, *99*, 102181. [CrossRef]

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, *160*, 108386. [CrossRef]

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*(2), 331. [CrossRef]

Tawfik, A. A., Graesser, A., Gatewood, J., & Gishbaugher, J. (2020). Role of questions in inquiry-based instruction: Towards a design taxonomy for question-asking and implications for design. *Educational Technology Research and Development*, *68*(2), 653–678. [CrossRef]

Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., Testa, A., Boiano, R., Hernandez-Leo, D., Ruskov, M., Taibi, D., Gabbiadini, A., & Ognibene, D. (2023). Learning to prompt in the classroom to understand AI limits: A pilot study. In *International conference of the Italian association for artificial intelligence* (pp. 481–496). Springer Nature.

Tobler, S., & Köhler, K. (2025). Towards personalized education in life sciences: Tailoring instruction to students' prior knowledge and interest through machine learning. *Trends in Higher Education*, *4*(4), 68. [CrossRef]

Tobler, S., Sinha, T., Köhler, K., Hafen, E., & Kapur, M. (2022). The impact of prior knowledge in narrative-based learning on understanding biological concepts in higher education. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44). University of California.

Valeri, F., Nilsson, P., & Cederqvist, A. M. (2025). Exploring students' experience of ChatGPT in STEM education. *Computers and Education: Artificial Intelligence*, *8*, 100360. [CrossRef]

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, *46*(2), 186–204. [CrossRef]

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' aca-demic development. *Education Sciences*, *15*(3), 343. [CrossRef]

Vosniadou, S. (2014). Examining cognitive development from a conceptual change point of view: The framework theory approach. *European Journal of Developmental Psychology*, *11*(6), 645–661. [CrossRef]

Wang, S. Y., Tsai, J. C., Chiang, H. C., Lai, C. S., & Lin, H. J. (2008). Socrates, problem-based learning and critical thinking—A philosophic point of view. *The Kaohsiung Journal of Medical Sciences*, *24*(3), S6–S13. [CrossRef] [PubMed]

Zhang, J. S., Yoon, C., Williams, D. K. A., & Pinkas, A. (2024). Exploring the usage of ChatGPT among medical students in the United States. *Journal of Medical Education and Curricular Development*, *11*, 23821205241264695. [CrossRef] [PubMed]

Zobel, T., Steinbeck, H., & Meinel, C. (2023, October 11–13). *Towards inclusive education: A review of and guide to accessible features in chatbots for moocs*. 2023 IEEE Learning with MOOCS (LWMOOCS), Cambridge, MA, USA.