

To Chat or to Quiz?: Examining the Pedagogical Benefits and Risks of AI Tutors in Facilitating High School Science Learning from Videos

Xiaoyu Liu
University of Maryland
College Park, United States
dasqxyl@umd.edu

Yunyan Li
University of Michigan
Ann Arbor, United States
yunyanli@umich.edu

Ying Xu
University of Michigan
Ann Arbor, United States
yxying@umich.edu

Xinyi Lu
University of Michigan
Ann Arbor, United States
lwlxy@umich.edu

Joshua Littenberg-Tobias
GBH
Boston, United States
joshua_tobias@wgbh.org

Soobin Jeon
University of Michigan
Ann Arbor, United States
sbjeon@umich.edu

Shawn Y Stevens
GBH
Boston, United States
shawn_stevens@wgbh.org

Xu Wang
University of Michigan
Ann Arbor, United States
xwanghci@umich.edu

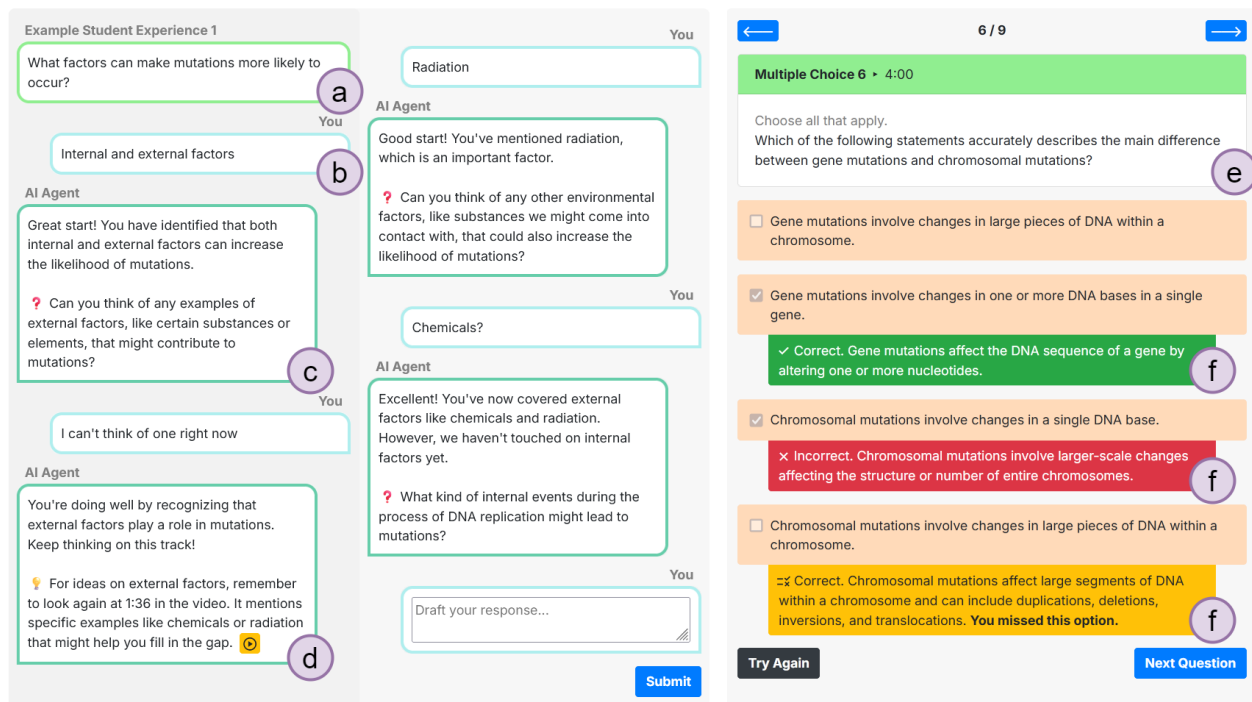


Figure 1: In this project, we investigate two methods of embedding AI-based tutoring in high school science videos. *ScienceChat* embeds questions in a science video and supports learning in two modes. In Chat Mode, learners answer open-ended questions (a) by constructing free responses (b). Based on the student answer, an AI agent may ask follow-up questions (c) or give hints (d). In Quiz Mode, learners answer multiple choice questions (e) and receive instant feedback (f).

Abstract

Chat-based tutoring is becoming increasingly powerful with the advances in generative AI. There is a prospect of using chat tutors to provide personalized and pedagogically beneficial feedback, which was only possible in traditional quiz-style tutors where students' answers were more easily assessed. This work contrasts the pedagogical benefits and risks of genAI powered chat and worksheet tutoring in high school science. We propose *ScienceChat*, a chat-based tutoring system that embeds open-ended questions throughout a science video, where students chat with an AI agent and receive feedback. We ran a within-subject experiment with high school students, who learnt with *ScienceChat* and a quiz tutor that posed multiple-choice questions. We found significant learning gains between pre- and post-tests in both conditions. Students in *ScienceChat* had higher engagement and used more science vocabularies in the post-test. We also observed frustrated moments when AI tried to elicit details that students thought were unnecessary.

CCS Concepts

• **Applied computing** → **Computer-assisted instruction; Interactive learning environments.**

Keywords

Chat-Based Tutoring, Multiple-Choice Questions, High School, Video-Based Learning

1 Introduction

Chat-based tutoring systems are becoming increasingly powerful with the advances in general-purpose large language models (LLMs) [23, 32, 39, 44, 48]. Chat-based tutoring describes scenarios where the tutor agent poses a question, the student constructs an answer, and the tutor agent provides a personalized and pedagogically beneficial feedback [30, 37]. In contrast, traditional intelligent tutoring systems (such as cognitive tutors [8, 18] and constraint-based tutors [33]) often pose more scripted questions to learners, such as multiple-choice or fill-in-the-blank questions, where existing algorithms can reliably provide automated feedback. These cognitive tutor-style questions also resemble the assignment worksheets that many students get in their science classrooms nowadays.

A key factor that differentiates between chat-based tutors and question-answering tutors (which we will refer to as "quiz tutors") is how much the tutor requires the learner to construct an answer independently from scratch. Additionally, chat-based tutors are more social and flexible (e.g., they can handle questions that students ask them) [14, 16, 38, 41], whereas quiz tutors are more targeted (e.g., they ensure that the students will get a learning experience that aligns with the lesson's objectives). Despite decades of research

demonstrating the benefits of tutoring systems of both kinds, there is a lack of empirical evidence comparing the two modes, especially when powerful generative AI tools could both enhance learners' learning experiences in chat-based tutoring systems, and facilitate the creation of quiz tutors.

In this project, we aim to investigate how students interact with and learn from generative AI-powered chat-based versus quiz-based tutoring systems. We situate our study in the context of high school science learning from videos, for the following reasons. First, students begin to learn more complex science phenomenon in high school in comparison to middle school, which may benefit more from sensemaking and explanation processes [19]. Second, students are not as independent as college students and would benefit from scaffolding and targeted practices [42]. Lastly, videos are commonly used tools in high school science classrooms where teachers may assign videos and facilitate classroom discussion based on the videos [24, 49].

To address the question of "to chat or to quiz", we developed *ScienceChat*, a video-based learning platform that supports two modes of learning. A set of questions are programmed within the platform. When learners watch videos in *ScienceChat*, the video would pause at question markers and the learner would answer a question about the content they have just watched. The two modes in *ScienceChat* supports chat tutoring and quiz tutoring respectively. In Chat Mode, learners answer open-ended questions, and there is an AI agent that gives feedback to the learner. In Quiz Mode, learners answer multiple-choice questions and receive both corrective feedback and explanations.

ScienceChat requires teachers to provide an initial list of open-ended questions that they would like students to learn based on the video, e.g., "What factors can make mutations more likely to occur?" For each question, teachers also need to provide a set of grading components that they would like to see in the student answer, e.g., "Component 1: External factors include chemicals or radiation", "Component 2: Internal factors include events that cause issues during DNA replication". With these inputs, the AI chat agent will present students with the original open-ended question and guide them to learn through a back-and-forth conversation. In the process, the agent may ask follow-up questions or provide hints. With the exact same teacher input, *ScienceChat* can also automatically generate a set of multiple-choice questions with similar question stems, up to 4 options per question, and a feedback message for each option. In Quiz Mode, students answer the multiple-choice questions at the same timestamps as in Chat Mode.

We conducted a within-subject experiment with 14 high school students within the grade range 9–12. Each participant watched two science videos (on topics of mutations and geoengineering) either in Chat Mode or Quiz Mode, counterbalanced. We administered a pre-test and a post-test quiz per video to measure learning outcome, and an experience survey per video to probe into their experience. Each study lasted for 90 minutes.

The following is a summary of our findings and implications:

- We observed significant learning gains from pre-test to post-test for both conditions. On a 8-point quiz, students' average score increased from 2.9 to 6.2, an average gain of ~41% of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '25, June 03–05, 2018, Woodstock, NY

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06

<https://doi.org/XXXXXX.XXXXXXX>

the total score. However, we did not witness a significant difference between the two conditions on learning gains. This suggests that when teachers are limited on time, a carefully designed LLM pipeline could deliver pedagogically beneficial quiz questions which are ready for immediate use, e.g., on platforms that are already popular among teachers such as EdPuzzle [3]. Given that using AI-powered chat tutors requires more computational resources, it needs further evidence to advocate for their use.

- We did observe qualitative evidence that for students who had lower pre-test scores, they seemed to benefit more from Chat Mode. Since our study sample is relatively small, future work is required to investigate the interaction between tutoring methods and students' prior knowledge levels.
- In the experience survey, students rated Chat Mode to be more thought-provoking, provided more new knowledge, and encouraged deeper reflections about the video. However, students also rated Chat Mode to incur higher cognitive loads.
- We observed that answering questions in Chat Mode was more difficult and took longer for the students. On the other hand, students reviewed previous video segments more frequently and recalled more details when answering questions in Chat Mode, whereas many students got the multiple-choice question right on first try. We observed qualitative evidence of learning in particular in Chat Mode, e.g., when students showed misconceptions in their answers, the AI agent guided them to correct the mistakes.
- We also noticed situations where the AI chat agent was annoying and distracting for the students, e.g., when it tries to elicit unnecessary details from the students, when it incorrectly assesses the student's answer, or when it reveals the correct answer in the follow-up question, making the student confused.

2 Related Works

2.1 Chat-Based Tutoring

Chat-based tutoring refers to the instructional strategy in which tutors ask students open-ended questions to provoke their thinking, with feedback and / or follow-up questions offered to them to further guide them towards constructing their own understanding of a learning topic [39, 48]. This approach is grounded in socio-cultural theories, which posit that learning is a socially mediated process where children would acquire knowledge and skill through guided interactions with more knowledgeable partners. Such external scaffolding will expand a child's development potential from the limited circle that they can achieve without the intervention of others (referred to as "the zone of proximal development" by Vygotsky [15]). In this sense, through active question answering, students can achieve a level of learning that they would not have otherwise reached.

In chat-based tutoring, the benefits of asking questions are greatly enhanced when combined with responsive dialog, which often includes a mix of hints, clarifications and explanations [13]. During a pedagogical dialog, questioning helps tutor assess the children's current level of understanding, while the responses that children

receive back prompt them to reconsider their initial responses, identify areas for revisions, and refine their answers towards the expected direction. These principles, which support overall comprehension through conversational exchange, can also be applied to the learning of specific domains. For instance, studies show that interactive dialog interventions in classroom settings not only boosts students' language comprehension, but also deepens their knowledge in disciplines such as math and science [17, 45].

Rapid advancements in machine learning technologies have enabled AI to amplify the learning benefits of chat-based tutoring. AI could provide additional dialog opportunities with children, especially when a human conversation partner is unavailable. Before the advent of LLMs, studies have already started to apply strategies grounded in human-to-human chat-based tutoring strategies to digital tutoring systems. An early example for this line of work was AutoTutor [12]. This pedagogical agent was designed to engage in natural language conversations with students. AutoTutor's dialog strategies include asking students questions and using special techniques to elicit their thinking. These techniques include prompts that encourage students to elaborate on their responses, hints that break down complex questions for easier understanding, and guidance that help students revise their responses towards the expected answer. Studies have suggested that these intelligent tutoring systems can significantly enhance students' learning outcomes [29]. In some studies, the benefits observed from intelligent tutors are even comparable to those generated by a trained human tutor [20].

LLMs offer new possibilities to make these intelligent tutors even more customized to student responses. Unlike traditional intelligent tutors (e.g., AutoTutor), LLM-enabled chat-based tutoring is not limited by preset dialog structures or pre-scripted responses. Therefore, it is possible that these LLM-powered tutoring systems can independently generate questions based on provided learning materials and provide highly relevant and coherent responses based on past interactions [46]. This adaptive feature makes LLMs well-suited for open-ended conversations with students, especially in science disciplines where one of the most important learning objectives is problem solving.

2.2 Evaluation of Tutoring System Strategies

A few large-scale meta-analyses on the effect of different tutoring systems have been conducted in recent times. [43] categorizes existing tutoring systems into five subgroups depending on the granularity of their instruction strategy:

(1) Human tutoring

A human leads the tutoring session instead of a computer system.

(2) Substep-based tutoring

The system provides live scaffolding and feedback through conversation as the learner solves the problem.

(3) Step-based tutoring

The system prompts the learner to input the final answer and pre-defined intermediate steps, then provides feedback based on the answer and steps.

(4) Answer-based tutoring

The system prompts the learner to input the final answer

only, then infers their reasoning and provides feedback based on their answer.

(5) No tutoring

These subgroups are ordered in decreasing levels of granularity, i.e. the minimum amount of reasoning that can be contained within each round of system-learner interaction. For example, human tutoring is not constrained by any computer system and can go as detailed as possible. On the other hand, answer-based systems will only perform a one-off evaluation on the final answer, with little to no insight about how the learner actually arrived at that answer. It is thus possible to map the three subgroups in the middle with the two types of tutoring considered in our work ($\supseteq$ means "is a superset of"):

- Quiz-based tutoring $\supseteq$ Answer-based tutoring / step-based tutoring
- Chat-based tutoring $\supseteq$ Substep-based tutoring

In the [43] study, it is found that (1) the learning effect plateaus at step- and substep-based tutoring, making both of them indistinguishable from human-tutoring; (2) both step- and substep-based tutoring demonstrate better learning outcomes than answer-based tutoring. These conclusions fail to provide evidence for a comparison between chat- and quiz-based tutoring, as it appears that the performance of quiz-based tutoring is dependent on how much learner response is provided through the user interface (answer- vs. step-based tutoring).

A subsequent meta-analysis [21] adopted the same system categorization as [43]. They further confirmed the effect plateau for step- and substep-based tutoring by excluding (previously included) studies conducted under non-conventional teaching settings (e.g., studies which involved watching pre-recorded tutoring lectures).

It is also worth mentioning that both meta-analyses (including all of their aggregated studies) come before the general availability of LLMs like ChatGPT, so their influence on the potential outcome of chat-based tutoring was yet to be considered.

2.3 LLM Applications in Tutoring

In recent years, LLMs have started to demonstrate human-like performance in reasoning [9] and problem-solving [34, 36]. One line of work has explored the potential of LLMs as high quality tutors [28, 40]. Given the effectiveness of LLMs in feedback generation [26, 31, 35, 40], question answering [23] and expertise in teaching methodologies [28], LLM-based tutors are empowering modern Intelligent Tutoring Systems (ITS) to produce fast and targeted feedback to natural language user inputs. However, one major concern of using LLMs for high-stake activities like school teaching and tutoring is hallucination [27]. Recent work have suggested using Retrieval-Augmented Generation (RAG) to reduce the chance of producing erroneous information in LLM-based systems [10, 11]. By restricting generation only to information retrieved from the knowledge base, RAG could enhance the reliability of the LLM's output [25]. In this work, we aim to extend these explorations in LLM-based tutor systems to facilitate students' science video learning.

3 Formative Investigation

We performed an IRB-approved formative study to understand how teachers would like AI tutors to behave. In the formative study sessions, teachers were asked to develop lesson plans and comment on AI-generated tutoring questions. We recruited a total of 7 secondary science teachers to participate in our study. They reported an average of 14.6 years of science teaching experience, ranging from 6 to 26 years. Their teaching expertise included biology, chemistry, physics, engineering, and earth and space science. Each study session lasted 60 minutes online (through a Zoom meeting). After the session, they were compensated with a \$50 gift card.

During a session, a teacher was asked to watch a science video, and create lesson plans that involve teaching the video. Particularly, teachers were asked to design some questions that could encourage their students' higher-order thinking about the video. Following the lesson plan creation, we showed the participant with a set of AI-generated (using GPT-4) open-ended and multiple choice questions based on the same video that they just watched. We asked them to review the questions, annotate on them, and comment on how these questions intersect with their instructional goals.

We analyzed the data using affinity diagrams and here is a summary of findings:

- Teachers all really liked the idea of embedding an AI agent in the video which asks open-ended questions and engages in back-and-forth conversations with the students, and preferred for the agent to ask open-ended questions instead of multiple-choice questions.
- When creating questions to help teach the video, teachers would like to embed questions within the video at regular intervals to maintain student engagement and focus.
- Teachers want the questions asked to align with NGSS standards and specific to the concepts presented in the video. Teachers would like to ask questions about important terminologies/vocabularies. Teachers want to make sure the questions are answerable based on the video content and do not want the students to get frustrated by the difficulty of the questions.
- Teachers want to provide immediate feedback to students' answers. They are currently doing this through in-class discussions.
- Teachers want to teach critical thinking, e.g., they want the students to understand the reasoning behind the explanations, apply the concepts in real-world situations, and explain the concepts and phenomenon in their own words.

These findings inspired our design of *ScienceChat* as a conversational space for students to talk about science concepts in their own words under the guidance of an AI agent. However, it is of LLM's nature to be divergent, which often leads to hallucinations. In the study, we noticed that when teachers reviewed LLM-generated questions and dialogs, they especially wanted to avoid cases where the AI agent may ask a question that is beyond the student's knowledge level, or introduce new information that is unrelated to the learning material. Therefore, one key design consideration is for the *ScienceChat*'s AI chat agent to stay grounded in the video content.

4 ScienceChat

We introduce *ScienceChat*, an AI-based tutoring platform for science videos that supports both chat-based and quiz-based learning. Teachers can import course material (e.g., videos), create questions, specify expected components in student answers, and test out potential student experience with the chat agent in the Editor Mode. Students learn in the dedicated Student Mode. *ScienceChat* leverages the RAG capability of GPT-4o [4] to generate follow-up feedback for each student response in the Chat Mode.

4.1 Example Teacher’s User Journey

Mary is a high school biology teacher. In the next class, she decides to play a video on mutations [7]. To facilitate their learning, she has designed some open-ended questions for students to discuss based on the video content. To make sure that each student can receive instant, customized feedback based on their answers, Mary uses the *ScienceChat* system to set up and present her questions. An overview of her interaction with the system is shown in Figure 2.

Mary can upload new videos to the system or provide links to existing online videos. Mary can also provide learning standards, such as the Next Generation Science Standards (NGSS) [5] or state-wide disciplinary standards, which will be incorporated into the AI chat agent’s prompts.

In Chat Editor Mode, Mary can import her pre-designed open-ended questions, and specify the expected components in students’ answers. The chat agent will use such information to guide the students. On the same interface, Mary can initiate an example conversations with the tutoring agent as if she were a student. In case she discovers anything that was not covered, or notices that the tutoring agent might need some additional knowledge in order to provide effective feedback, she can go back to edit the questions and / or the expected components, and the changes will be automatically applied to the tutoring agent.

With Mary’s input, which includes the open-ended questions and the expected components, the system can automatically generate multiple-choice questions and feedback for her. Mary can also enter Quiz Editor Mode to create new questions from scratch. Once Mary is ready with the system setup, she can send the link to the student mode to her students.

Students will enter the Student Mode interface of *ScienceChat* (Figure 3), where all interface widgets intended for Editor Mode are hidden.

4.2 Chat Agent Implementation

While appearing as one monolithic component, the AI agent in Chat Mode is actually composed of multiple specialized agents communicating with each other, each configured with a set of different chain-of-thought prompts [47] (see Appendix A). This multi-agent approach allows different agents to query different contexts (i.e., different subsets of the video transcript) in order to deliver more accurate evaluation of student answers.

In each turn of a dialog, the student response and intermediate evaluation results are embedded into these prompts as parameters in order to generate the next follow-up message. The logic flow diagram for the multi-agent approach is shown in Figure 4.

4.2.1 Agent Initialization. Internally, each specialized agent calls one or more *assistants* provided by the OpenAI Assistants API [1]. The assistants are capable of performing RAG on a text document ("context") and provide responses that align with the information contained in it. When the user first adds a video or a grading component, one or more assistants are initialized accordingly (Appendix A.1):

- The **video assistant** uses the full video transcript as the context, with temperature=1 (default value).
- The **context assistant** also uses the full video transcript as the context, but with temperature=0.
- The **grading component assistant** uses a subset of the video transcript that is most relevant to the grading component as the context. This is extracted through through a call to the *context assistant* (Appendix A.9).

Once the assistants are created, they are stored to the database and retrieved whenever called by a specialized agent. If the video transcript is modified by the user, all assistants will need to be re-initialized as well, but we do not expect this to be an action that needs to be performed frequently after the initial video import.

4.2.2 Main Student Evaluation Logic. The evaluation of a student’s progress is based on how many grading components they have fulfilled so far. If there exists some grading component(s) that the student has not yet satisfied, then they will be prompted by the agents to amend or improve their answer, one component at a time. Hence, the student will always have a component that they are currently working towards ("the current component").

4.2.3 Preliminary Classification. When a student response is first received by *ScienceChat*, it will check (1) whether the response is asking a question (Appendix A.3) and (2) whether the response contains any substance (Appendix A.2). The result of these classifications will determine which agent(s) to call in subsequent steps.

4.2.4 Q&A Agent. If the question check returns true, i.e. the student is asking a question instead of giving an answer, the Q&A Agent will call the *video assistant* to answer their question (Appendix A.6). This typically happens when the student is experiencing confusion or is looking for clarification about certain term or concept. Q&A interactions do not count towards the maximum number of attempts for a grading component (see Section 4.2.6).

4.2.5 Hint Agent. If the question check returns false (i.e. the student is not asking a question) and the substance check returns false (i.e. the student response does not contain any substantial effort, like simply saying "I don’t know"), then the Hint Agent will call the *grading component assistant* (Appendix A.5) to (1) give them a hint based on the current component and (2) remind them to rewatch the video at the component’s timestamp.

The other condition in which the Hint Agent is utilized is when the student has already made n unsatisfactory answers towards the current component (see Section 4.2.6). This usually means that they are having difficulty grasping the core concept entailed by this component. The student can be hinted towards a grading component at most once.

4.2.6 Rubric Agent. If the student is not asking a question and their response is substantial, then the Rubric Agent will check whether

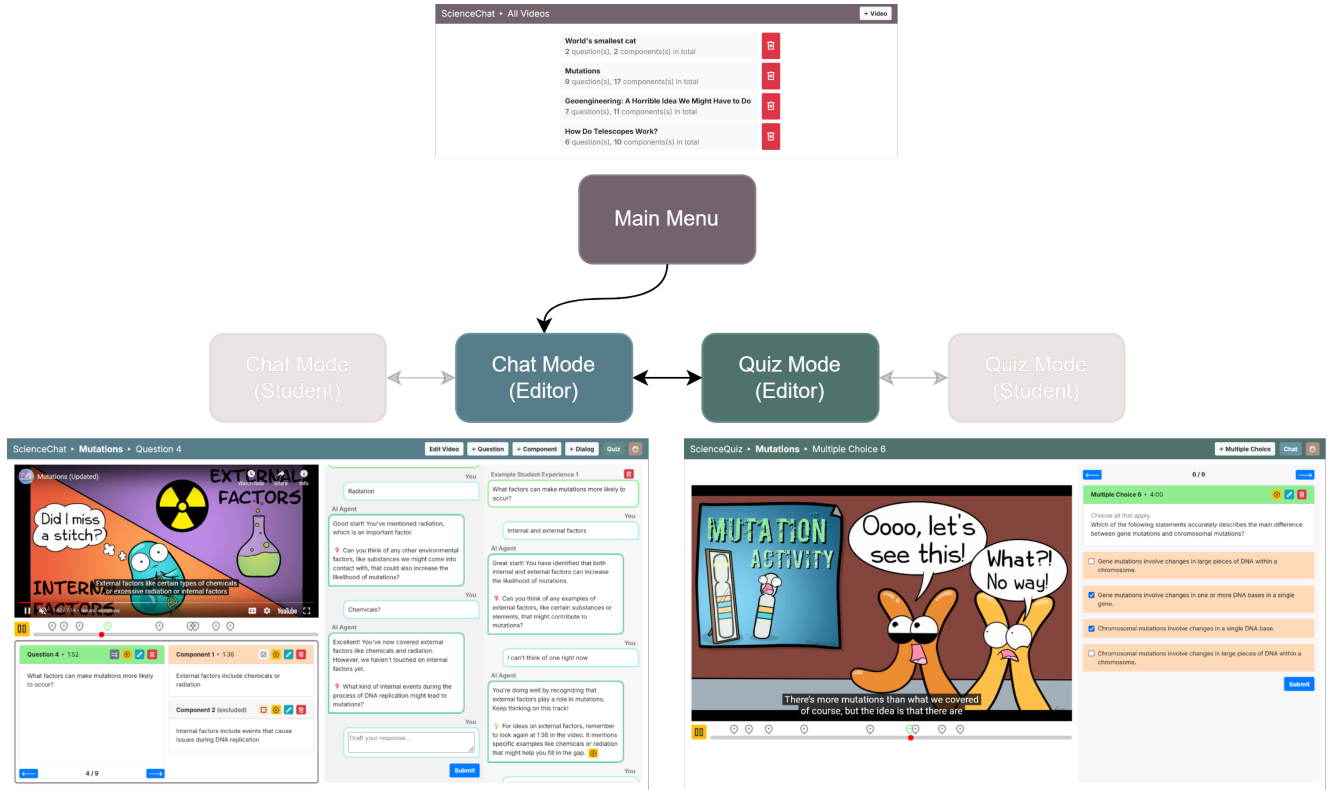


Figure 2: Teachers can import videos, create questions, specify the expected components in student answers in the Editor Mode. With these basic inputs, the system automatically initializes the chat agent in Chat Mode, and automatically generates the multiple-choice questions and feedback messages in Quiz Mode.

this response fulfills the current component. This is done through a call to GPT-4o chat completion API with the current component’s context contained in the prompt (Appendix A.4).

If the current component is fulfilled, the agent will mark it as “done” and move on to asking about the next one. Otherwise, as long as the total number of attempts on the current component does not exceed a predefined value (the variable n in Figure 4), the agent will keep probing the student. In both cases, the Rubric Agent will call the *grading component assistant* of the component to be asked about (Appendix A.7).

4.2.7 End of Dialog. If there are no more grading components left unfulfilled, the agent will call the *video assistant* to end the dialog (Appendix A.8). The student will receive a concluding statement which tells them that they have completed the question.

4.2.8 Revealing Unfulfilled Grading Component. If the student remains struggling on the current component even after a hint from the Hint Agent, the next agent will reveal this component to the student before proceeding to the next component. This behavior applies to Hint, Rubric and End of Dialog agents.

4.2.9 System Performance. The multi-agent approached makes use of request parallelization, faster LLMs (GPT-4o mini vs. GPT-4o) and faster API endpoints (normal chat completion vs. Assistants API) to optimize the system turnaround time. With the current

implementation, *ScienceChat* in Chat Mode can generate a follow-up response in under 5 seconds on average (see Table 2). This is on par with general chatbot services like ChatGPT.

4.3 System Iteration

Based on findings from three pilot study sessions we conducted, we iterated on *ScienceChat* in the following ways.

4.3.1 Reducing Maximum Student Attempts before Giving a Hint. Originally, we set that the maximum number of student attempts per grading component to be $n = 2$ (see Section 4.2.6), after which a hint will be given. We found that given the amount and complexity of grading components available, students frequently become frustrated by the extended length of dialogs and loses focus in their answers. So for subsequent study sessions, we lowered n to 1 to reduce the upper bound of each dialog’s depth.

4.3.2 Reducing Length of AI Agent’s Feedback Message. ?

4.3.3 Semantic Tags in AI Agent’s Feedback Message. We found that the lack of visual cues in the structure of AI agent’s response was another factor that made students’ understanding difficult, even if the feedback message was relatively short.

To optimize the appearance of feedback messages, we prompted the AI agent to annotate different sections of its response with

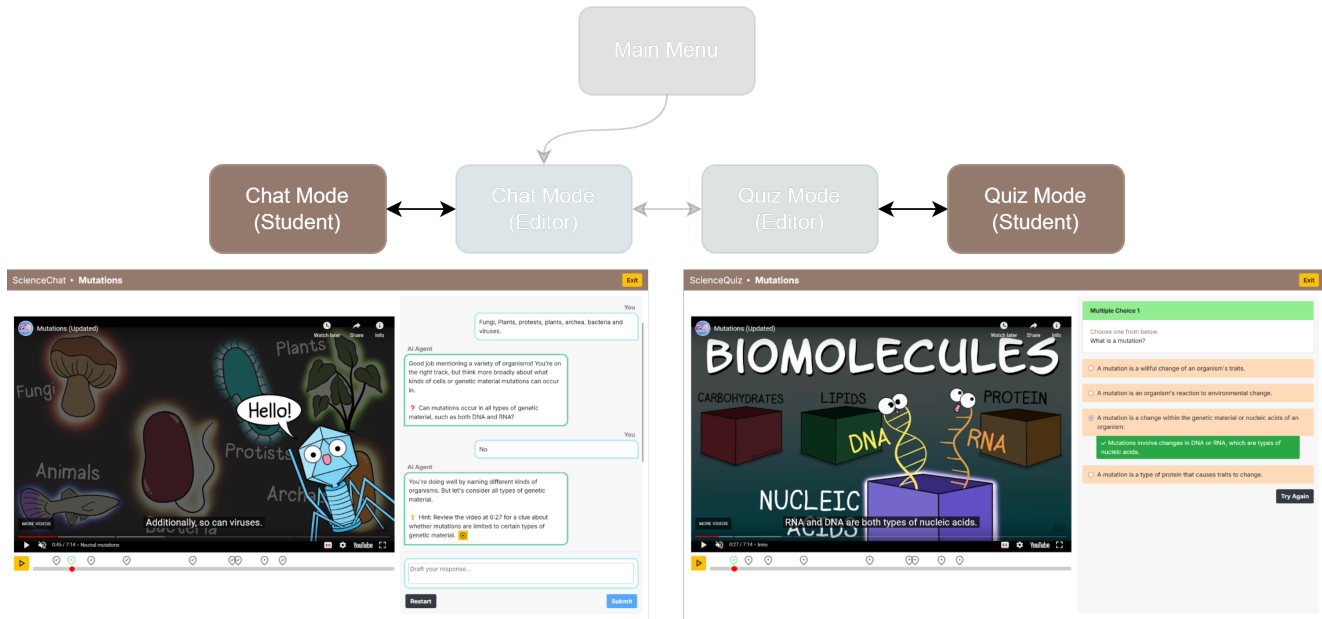


Figure 3: Student Mode of the *ScienceChat* system. Students can learn in both the Chat Mode and the Quiz Mode. The platform display a video, and there are questions markers on the video player. The video would pause at each question marker. The student will answer open-ended questions guided by an AI agent in the Chat Mode and multiple-choice questions in the Quiz Mode. In the Chat Mode (left), the agent will ask follow-up questions if their answers are unsatisfactory. The agent may also give the student a hint and prompt them to revisit a specific point in the video. In the Quiz Mode (right), the student receives immediate feedback after making a selection.

a number of pre-defined semantic tags. During front-end rendering, these tags are (1) converted into emojis for accentuation and (2) used to divide the agent response into paragraphs. A list of semantic tags is given in Table 1 and an example is given in Figure 5.

4.3.4 More Accessible Video Replay Button. The video replay button that comes with each hint message was originally positioned at the upper-right corner of AI agent’s dialog bubble, with the intention of increasing its visibility and utilization. However, post-session interview with the pilot participants revealed that they either did not notice there was a replay button at all, or they found it easier to interact with the video player itself than to mouse over to the button’s position.

Based on their feedback, we moved the video replay button to the end of the hint paragraph, so it coincides with their eyes’ focus once they finish reading the entire message (Figure 6).

4.4 System Implementation and Deployment

All front-end interfaces are designed with React . js [6] framework. The back-end server is implemented with the Django [2] framework. The student evaluation and response generation logic for Chat Mode AI agent is implemented in Python. *ScienceChat* is deployed to a VM instance on Google Cloud Platform.

5 Study Design

We conducted an IRB-approved study with high school students. Using *ScienceChat* as an instrument, we explored the following research questions:

- RQ1 Which of the two modes of science video learning would result in higher learning gains?
- RQ2 Which of the two modes of science video learning do high school students prefer?
- RQ3 What are the advantages and pitfalls of implementing an AI chat-based tutoring system for learning high school science videos?

5.1 Participant Recruitment

We dispersed participant recruitment form and study description document in multiple high schools. After we recruited an initial group of participants, we encouraged them to further spread our recruitment message and/or refer their schoolmates to sign up for the study. To be eligible for participation, each student must be between Grades 9–12 (inclusive) and have their parent or legal guardian sign a consent form prior to the study session. In total, 14 high school students participated in the study: 3 participants were from Grade 10, 4 from Grade 11, and 7 from Grade 12. Each study session lasted approximately 90–100 minutes and participants may choose to attend in person (in a classroom; 3 participants) or online (through a Zoom meeting; 11 participants). After the session, they were compensated with a \$50 gift card.

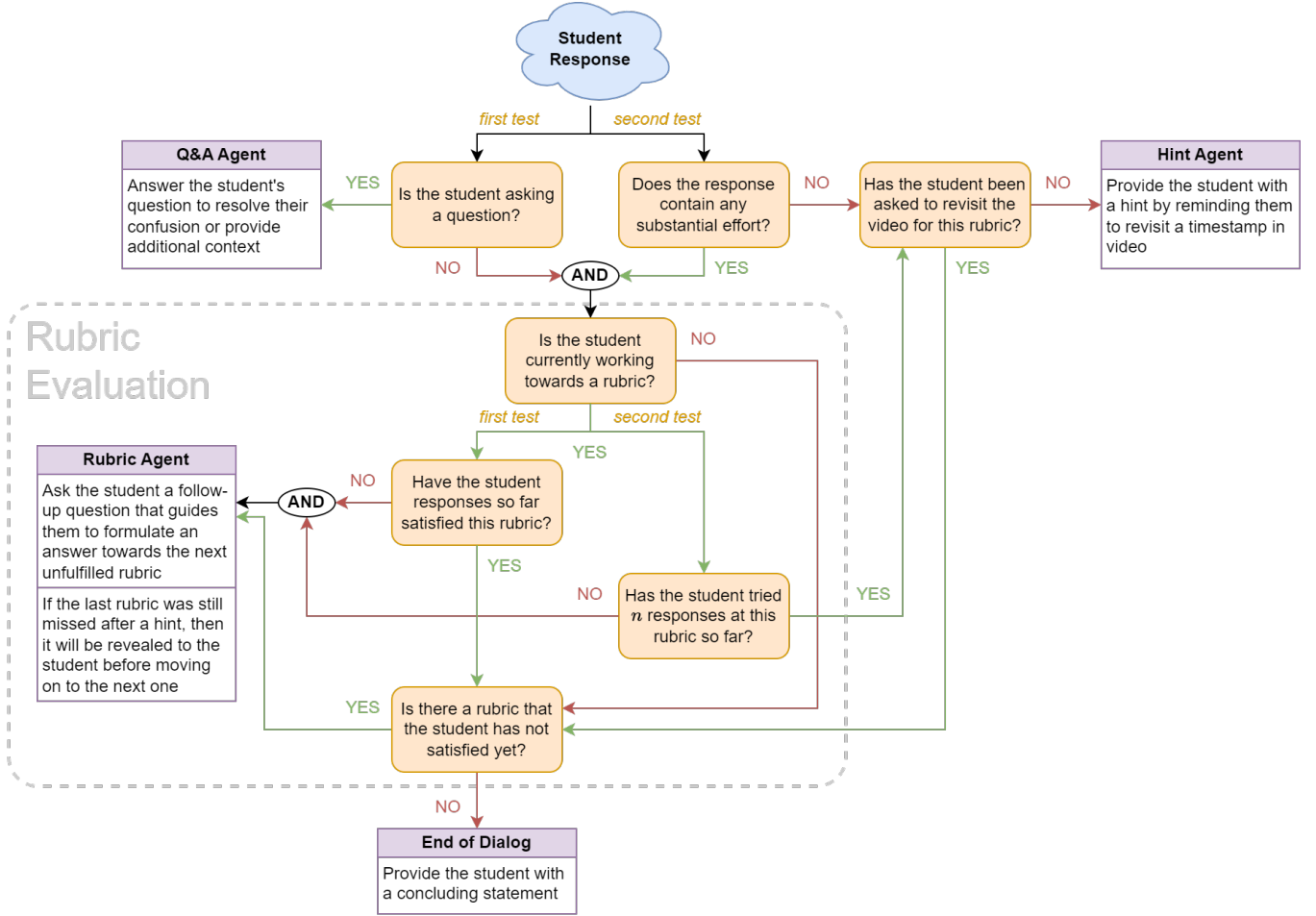


Figure 4: Logic flow diagram for the multi-agent follow-up generation process in Chat Mode.

Table 1: List of semantic tags used by *ScienceChat* AI agent, their corresponding emoji replacement and context.

Semantic Tag	Emoji	Context
[ANSWER]	Exclamation mark + "What you missed:"	Hint / Rubric / End of Dialog Agent reveals a grading component
[END]	Confetti	End of Dialog Agent gives a concluding statement
[HINT]	Light bulb	Hint Agent gives a hint
[EXPLAIN]	Check mark	Q&A Agent answers the student's question
[QUESTION]	Question mark	Rubric Agent asks a follow-up question
[COMMENT]	(None)	All other agent outputs

5.2 Study Material

5.2.1 Science Video Selection. We consulted high school teachers and current high school students to decide on the subject, topic and difficulty of science videos to be used for this study, and settled on two popular science videos of comparable depth: one on mutations [7] ("video M") and the other on geoengineering [22] ("video G").

5.2.2 Question Generation. Our collaborating high school teacher created the initial set of open-ended questions to be embedded in the videos. The Mutations video has 9 questions and the Geoengineering video has 7, which is based on the number of important concepts that can be learned from each video. Each question comes with 1–3 grading components. Both videos are about 8 minutes long, it took the teacher about 15 minutes to gather the inputs for each video. We manually mapped the questions to a timestamp in

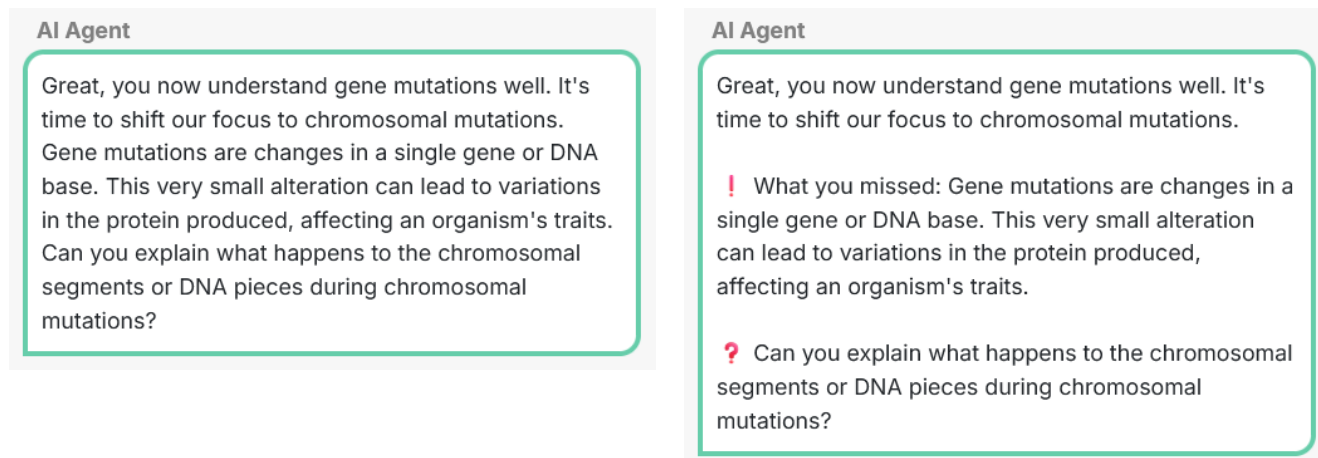


Figure 5: An example AI agent feedback message before (left) and after (right) the introduction of semantic tags.

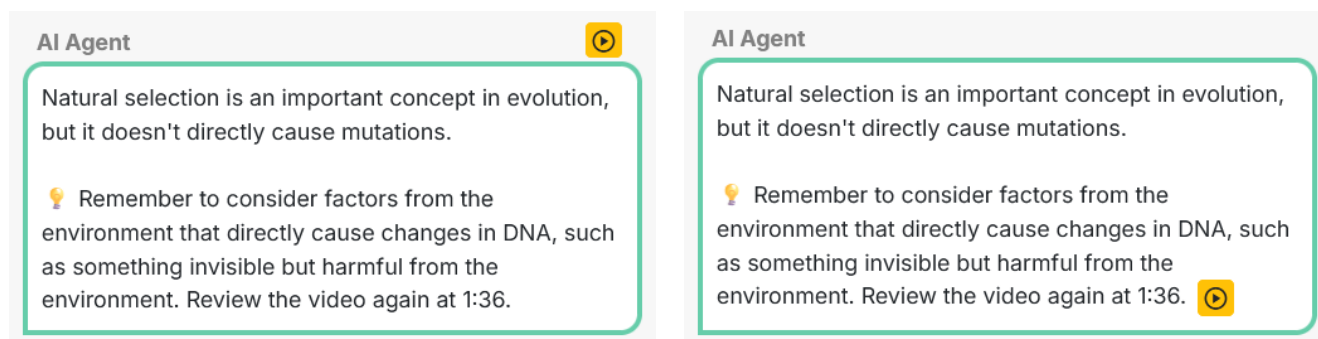


Figure 6: The placement of video replay button before (left) and after (right) the iteration.

the video. Special care was taken to place the questions between sentence breaks in the video's narration; this ensures that in Student Mode, the video player will not cut off a sentence halfway when pausing for a question. Finally, the data are imported into *ScienceChat*'s Chat Editor Mode interface directly.

The multiple choice questions are all AI-generated from the corresponding opened-ended question using the system's conversation prompt (Appendix A.10). The opened-ended questions in Chat Mode and the multiple-choice questions in Quiz Mode are mapped one-to-one and positioned at the same timestamps.

5.2.3 Pre-/Post-Test Quizzes. To investigate RQ1, we invited the high school teachers to design two sets of quizzes ("question set A / B"), along with answer keys, based on each video's content. Both versions have 5 questions covering the same essential concepts, but the question stems are slightly varied to function as pre- and post-test questions. Each test is worth 8 points. During the study sessions, the order in which each participant receives the two quiz sets is shuffled. One may receive question set A before the video (pre-test) and B after the video (post-test), or vice versa. After all sessions are complete, all quizzes are graded anonymously according to the answer key. We consider the difference in pre- and post-test scores as an indicator of students' knowledge gain.

5.2.4 Experience Surveys. To investigate RQ2 and RQ3, we designed a Likert-scale survey for Chat Mode and Quiz Mode respectively. The survey asks participants to rate to what extent they agree or disagree to a series of statements about their learning experience in *ScienceChat*. The survey is given after they complete the corresponding learning mode.

5.3 Study Procedure

Before each study session, the following items are determined at random:

- The order in which the participant is given video M and video G;
- The order in which the participant interacts with Chat Mode and Quiz Mode;
- The order in which the participant is given question set A and B as pre- / post-test.

We first gave the participant an overview of the study setup and procedure, and obtained their consent to record the session. In learning one video, participants first answered the pre-test quiz. They were given 8 minutes to complete it to the best of their ability. They will then login to the *ScienceChat* system. If they were assigned to learn this video in the Chat Mode, they will access

the Chat Student Mode interface. We give them a brief tutorial on how to navigate the system interface. They have 25 minutes to watch through the video and complete all open-ended questions. We observed the participant’s interaction with *ScienceChat* with minimal interruption, except only when they needed help with troubleshooting. After the learning, the participant was asked to fill out the Chat Mode experience survey in 3 minutes. Finally, the participant answered the post-test quiz. Again, they were given 8 minutes to finish and submit the quiz. They will repeat a similar process for the second video, except that they would learn the video in the other mode, in this case, it’ll be the Quiz Mode.

After learning of both videos, we performed a brief interview with the participant to learn more about their experience. When discussing their positive or negative interactions with the system, we encouraged them to point at specific dialogues or feedback messages.

5.4 Data Analysis

We extracted the following sources of quantitative and qualitative data from the study sessions.

- (1) We reviewed the performance and utilization of *ScienceChat*’s chat agent based on raw data extracted from the system logs.
- (2) We evaluated participants’ preference between Chat Mode and Quiz Mode based on their responses to the experience survey.
- (3) We compared participants’ quiz performance before and after watching the science video to quantify their knowledge gains.
- (4) We transcribed and analyzed the study session transcripts and created affinity diagrams to identify important themes about students’ experiences and preferences of the Chat Mode and Quiz Mode in the system.

6 Findings

6.1 System Logs

In total, 14 participants submitted **284 responses** to the AI agent in **109 conversations**. The descriptive data is displayed in Table 2. The average AI response time is less than 5 seconds. Since we used GPT-4o’s streaming API, on average participants would begin to see generated content appearing on their screen within 3 seconds. Participants’ average response length is 12.9 words, and AI’s response is much longer, with an average of 51.7 words. The number of messages per conversation is 2.6.

Table 3 provides a summary for the frequency that each specialized agent generated a response. Rubric Agent was called most frequently.

6.2 Pre- / Post-Test Performance

We observed significant learning gains from pre-test to post-test for both conditions. On a 8-point quiz, students’ average score increased from 2.9/8 to 6.2/8, an average gain of 41% of the total score. However, we did not observe a significant difference between the two conditions on learning gains.

Table 2: The average AI response time is less than 5 seconds. Since we used the streaming API, on average, learners would begin to see updated content on their screen within 3 seconds. Participants’ average response length is 12.9 words, and AI’s response is much longer, with an average of 51.7 words. The average number of learner messages per conversation is 2.6.

Statistic	Min	Avg.	Max
AI agent turnaround time (second)	3	4.86	17
Learner response length (word)	1	12.9	68
AI agent response length (word)	27	51.7	95
Number of learner responses per dialog	1	2.6	7

Table 3: Utilization frequency of each specialized agent type in the study.

Agent Type	Frequency	
Q&A Agent	4	1.4%
Hint Agent	65	22.9%
Rubric Agent	110	38.7%
End of Dialog	105	37.0%
Total	284	100.0%

6.3 Experience Survey

To address RQ2 and RQ3, we compared the students’ Likert-scale survey scores on their experiences with the Chat Mode and the Quiz Mode. Students generally responded more favorably to questions about the advantages of both modes, with average scores exceeding 3.5 on a 7-point scale. When comparing the average scores between the Chat and Quiz mode, we found students thought the Chat Mode more thought-provoking, provided new knowledge, and encouraged deeper thought about the video. In particular, by running a *t*-test, we found statistical significance in how students viewed the open-ended questions as more thought provoking than multiple choice questions (p -value = 0.004) and how the AI agent’s follow-up questions made them think harder about the video compared to feedback from multiple choice questions (p -value = 0.00).

Although not statistically significant, the Quiz Mode had higher average scores in certain areas compared to the Chat Mode. Students found reviewing multiple-choice responses easier and more helpful for their learning than reviewing the AI agent’s dialog history. They also preferred learning science videos in the Quiz Mode and were more likely to recommend the Quiz Mode experience to friends compared to the Chat Mode.

Regarding the disadvantages of both modes, there were statistically significant differences, with students responding less negatively to the Quiz Mode than the Chat Mode. When asked about their preferences, students wanted fewer open-ended questions compared to multiple-choice questions during the video experience (p -value = 0.009). They felt the AI agent in the Chat Mode elicited more unnecessary details compared to Quiz Mode’s multiple-choice

feedback (p -value = 0.03). Additionally, students reported that open-ended questions required more mental effort than multiple-choice ones (p -value = 0.049).

6.4 Post-Session Interview

Below, we present the most prominent themes that we identified from the transcripts of participant interview sessions.

6.4.1 (+) Chat agent's follow-up questions are enhancing participants' understanding/pointing them in the right direction. Many participants (P2, P8, P9, P11, P13, P14) mentioned that the chat agent in Chat Mode helped them formulate a deeper understanding of the subject matter and gain new knowledge that they were not familiar with before. P11 shared, "[Before] the open ended questions, I didn't know anything about that [video] topic. But now I feel a lot more well versed on it. [...] I already learned a lot of stuff about that topic." An example dialogue is P1's discussion about the impact of mutations, in which the AI chat agent made an effort to ensure that that participant knew both whether and why some mutations can be neutral.

Question	Are the results of mutation good or bad?
Participant	They are random, they can be good or bad.
AI Agent	[COMMENT] Great start! You've nicely mentioned that mutations can be both harmful and beneficial. However, not all mutations are either. [QUESTION] Can you think of how some mutations might not impact the organism at all, particularly with regard to changes in amino acids?
Participant	Yes, mutation can be good or bad and sometimes have no impact.
AI Agent	[COMMENT] Good, you're on the right track! You've acknowledged that some mutations don't have any effect. [QUESTION] Can you explain how certain mutations might not cause any changes to the amino acids in an organism?
Participant	Certain mutations might not cause any change in amino acids as they are not significant enough to impact an organism.
AI Agent	[COMMENT] Excellent work! You have clearly articulated that mutations can be helpful, harmful, or neutral. You also explained that some mutations are insignificant in effect. [END] (concluding statements omitted)

Participants also liked the AI agent for pinpointing which parts of their answer is right (P4, P6, P8) or wrong (P2, P5, P8, P11, P13), which served as useful indicators for their next response. P8 said, "It told me what [parts] I wrote [were] right, which made me feel that I was going on the right track." P5 added, "Especially when I'm very close to the answer, I like how it tells me where to go and other points that I'm getting wrong in." In particular, P9 experienced the following dialog, in which the AI agent was actively working to correct the participant's misconception about the universality of mutations.

Question	What organisms are affected by genetic mutations?
Participant	Fungi, Plants, protests, plants, archea, bacteria and viruses.
AI Agent	[COMMENT] Good job mentioning a variety of organisms! You're on the right track, but think more broadly about what kinds of cells or genetic material mutations can occur in. [QUESTION] Can mutations occur in all types of genetic material, such as both DNA and RNA?
Participant	No
AI Agent	You're doing well by naming different kinds of organisms. But let's consider all types of genetic material. [HINT] Hint: Review the video at 0:27 for a clue about whether mutations are limited to certain types of genetic material.
Participant	yes, they can occur in all organisms with DNA and RNA
AI Agent	[COMMENT] Excellent! You've correctly identified that any organism with RNA or DNA can experience mutations. Great job expanding your answer and understanding the broader concept of genetic material mutations. [END] (concluding statements omitted)

6.4.2 (+) Chat agent's feedback leads participants to notice video contents that they have previously missed. Many participants (P2, P3, P6, P7, P9, P12) reported that AI agent's response helped them notice important video details that they would have otherwise neglected. P3 elaborated their experience: "[The AI agent] makes me want to look back into the video and see the little things that is said and allows me to pick them up. These are all the important things that can help me learn more about what is being discussed." P9 shared a more specific example:

Question	What are some potential side effects of aerosol injections?
Participant	They could disrupt rain patterns and/or tear holes in the ozone-layer.
AI Agent	[COMMENT] Great start! You've correctly noted that aerosol injections could disrupt rain patterns and harm the ozone layer. To fully answer the question, think about the broader impacts these disruptions might have. [QUESTION] How might changes in rainfall patterns specifically affect agriculture and food production?
Participant	(Participant rewinds the video manually to look for answers) The changes in annual rainfall could lead to drought and famine.
AI Agent	[COMMENT] Excellent work! You've done a great job identifying key side effects of aerosol injections and their potential to harm agriculture through changed rainfall patterns. [END] (concluding statements omitted)

P9 speculated in hindsight, "I assume this is a 3-part answer, so [the AI agent] is looking for effects on the agricultural system as well as the ozone layer and temperature. But I probably wouldn't have gone back to look at how it affected the food industry if it didn't suggest it."

6.4.3 (+) *AI agent's hint messages with video replay button are helpful and specific.* Several participants who used the video replay button commented on its convenience and accuracy (P2, P4, P5, P10, P11, P12). P4 compared *ScienceChat*'s design with an online quizzing platform that they had used in school before: "I remember for (platform name), when you say 'rewatch', [the video] would automatically go back by 5 seconds each time. That's not really a precise timing, right? But for this one, there's actually a timing like 'You have to go here'; 'That's where it starts'. I think that's pretty good." P11 further pointed us to a dialog in their session:

Question	What are aerosol injections?
Participant	They are when acids are injected into the atmosphere to control global warming effects. They are an example of global engineering
AI Agent	[COMMENT] Great start! You've correctly mentioned injecting substances into the atmosphere as part of geoengineering. However, you need to be more specific about what exactly is injected. [QUESTION] Can you think of a specific type of particle that is released into the atmosphere during aerosol injections, and why these particles stay suspended in the air?
Participant	Sulphate Particles are released into the air
AI Agent	[COMMENT] Nicely done! You've identified sulfur particles accurately. There's a tiny detail you need to include about why these sulfur particles stay in the air for a longer period. [HINT] Think about the size of these particles and how that might affect their behavior in the atmosphere. Review the video at 3:26.
Participant	(Participant clicks on the replay button and re-watches the video from the given timestamp) The particles are small and so they can stay in the air for extended periods of time
AI Agent	[COMMENT] Good job on recognizing that sulfur particles are released and that they stay in the air due to their size. [END] (concluding statements omitted)

P11 recalled, "There was this question where I really wasn't getting it, and then [the AI agent] gave me a timestamp which I could like reflect on, which covered what I was confused about. [...] The summary that I was given in the video was pretty accurate to what they were expecting."

6.4.4 (−) *AI agent's feedback is vague / confusing.* Many participants (P1, P2, P5, P8, P9, P13) experienced cases where they were confused by AI agent's feedback message and had little idea what to answer next. P8 felt that sometimes "[The AI agent] was wording it in a weird way that I did not understand it. I feel that if it was more specific, it would have been more helpful." P2 exemplifies this feeling with the following lengthy interaction:

Question	What's the difference between gene and chromosomal mutations?
Participant	Gene mutations involve abnormal changes to mRNA and amino acids while chromosomal mutations affect DNA and proteins.
AI Agent	[COMMENT] You've touched on important aspects of gene and chromosomal mutations! However, there's one crucial detail missing about gene mutations. [HINT] Remember to review the video around 1:56, focusing on what changes in a single gene.
Participant	A genetic mutation in a single gene can alter the proteins that are produced, which affect an organism's physical traits.
AI Agent	[COMMENT] Excellent progress on gene mutations! Now, let's delve deeper into chromosomal mutations. [ANSWER] What you missed: Gene mutations are changes in a single gene, DNA base, or nucleic acid. This is crucial to understanding how these can lead to alterations in an organism. [QUESTION] How do chromosomal mutations differ in terms of the scale of changes they cause to DNA within a chromosome?
Participant	Chromosomal mutations differ from genetic mutations because the genetic material in chromosomes is passed to offspring.
AI Agent	[COMMENT] That's a good start in understanding the inheritance aspect of mutations. But let's also think about the physical scale of DNA changes involved in chromosomal mutations. [HINT] Think about what happens to larger pieces of DNA within a chromosome. Review the video content at 3:13.
Participant	Chromosomes are made up of DNA and proteins. Humans have 46 chromosomes, half from the mother and half from the father. When these mutations are passed down, they will affect future offspring.
AI Agent	[COMMENT] Great job outlining the basics of inheritance! Remember, chromosomal mutations are about changes in larger DNA segments. [ANSWER] What you missed: Chromosomal mutations involve changes in large pieces of DNA within a chromosome. These changes can significantly impact an organism because they affect larger segments of genetic material. [END] (concluding statements omitted)

The participant was considered by the AI agent as failed to satisfy both grading components of this question, but the participant remained at a loss even after seeing the revealed answers and performing multiple video rewinds. They said, "I do think that there was a point where I felt [the AI agent] was trying to get me somewhere... but I wasn't really sure exactly where specifically. Like around here (underlined response), I was sort of rephrasing what it said in the video at that point, still not necessarily understanding how to answer the question." This was a prime example of the AI agent failing to guide the participant back on track towards the intended answer, delivering reduced educational value while costing the participant significantly more time to complete than usual.

6.4.5 (–) *AI agent fails to acknowledge participants' answers that should have been correct.* Another common complaint that we received from half of all participants (P1, P2, P8, P9, P12, P13, P14) was that the AI agent was not accepting their correct answers because they were not exactly the same as the grading components.

P14 observed, "*The AI system essentially wants a very specific answer, regardless of whether my answer was similar or just [phrased] in different words. [...] I noticed that in one of the questions, [...] the final answer was exactly what I had written down initially.*" P14's attempts that were not recognized by the AI agent are given below:

Question	What's the difference between gene and chromosomal mutations?
Attempt 1	(Intermediate follow-up questions omitted) The chromosomal mutations are on a larger scale while the genes are within the chromosomes
Attempt 2	Chromosomal mutations changes the DNA and Protein while gene mutations change the DNA bases
Component 1	Gene mutations are changes in a single gene / DNA base / nucleic acid
Component 2	Chromosomal mutations involve changes in large pieces of DNA within a chromosome

P14 continued to argue, "*I feel like that [mine] should have been the correct answer, and my past correct answers had to be restated in my final answer. [...] The component] is not exactly my answer, but it's implied in my answer.*"

The same issue happened to P12 in a different question:

Question	In what way does atmospheric CO2 contribute to global warming?
Participant	CO2 contributes to the greenhouse effect in the sense that it traps infrared heat emitted by the sun in our own atmosphere.
AI Agent	[COMMENT] Good start on identifying the role of CO2 in the greenhouse effect! However, there's a small error. CO2 traps infrared heat, but not from the sun directly. [QUESTION] Can you think about where the infrared heat that CO2 traps actually comes from? Hint: Consider the Earth's role in this process.
Participant	<u>The sun emits the light, it makes contact with the earth, and the earth releases it.</u>
AI Agent	[COMMENT] Nice work refining your response! You're very close to the key idea. Keep focusing on the interaction between Earth and heat. [HINT] Review the video content at 1:33 to see how CO2 is involved in trapping heat emitted by the planet.
Participant	<u>It's heat that was absorbed and sequentially released by the Earth.</u>
AI Agent	[COMMENT] Excellent! You've accurately described how CO2 traps heat that is emitted by the Earth after it absorbs sunlight. This detailed understanding shows a clear grasp of the greenhouse effect. [END] (concluding statements omitted)

P12 explained, "*It feels like [the AI agent] is looking for like a particular word or phrase, so that sometimes I would almost answer the question, only if I had worded it differently.*" In this case, P12 believed that his two underlined responses were identical in content: somehow, the AI agent rejected it the first time, but accepted it the second time. This had added to the participant's confusion and impatience.

6.4.6 (–) *Participants are frustrated by dialogs that keep going on without a clear goal.* Ambiguous AI agent feedback and unclear answer expectations are factors that often trap participants in a lengthy dialog that does not terminate until they reach the maximum number of turns allowed by the *ScienceChat* implementation. This turns out to be the greatest source of frustration during Chat Mode interaction (P3, P5, P6, P8, P10, P13, P14). P10 said, "*When [the system] was caught in this loop, it was almost like I was wanting to put less effort into the questions, because I just wanted to move on to the next task.*"

Though they did not explicitly bring it up in the interview, P1 had gone through this exact process of Q&A fatigue in the following 6-turn dialog. Notice how the answer first became longer in an attempt to fulfill as many grading components as possible, before collapsing down into a single word:

Question	What's the difference between gene and chromosomal mutations?
Attempt 1	(Intermediate follow-up questions omitted) The difference between them is that they change and can have mutation is different ways. (15 words)
Attempt 2	gene mutations can impact the way ones proteins are formed or even eye color pigment. (15 words)
Attempt 3	Changes occur in chromes can be one of 4 can impacts humans build up and influences and it can cause it to not get DNA from sperm and egg cells accurately compared to where gene mutation impacts ways proteins are formed or things such as pigment and eye color. (49 words)
Attempt 4	It is much more impactful with changes in chromosomes when compared to genes. (13 words)
Attempt 5	I dont know (3 words)
Attempt 6	ok (1 word)

To prevent this kind of potential frustration, P13 advocated for a "give up" functionality on open-ended questions, which "*would be better than just keep going until you get everything right.*"

6.4.7 (–) *Participants appreciate an opportunity to ask AI agent questions during their session, but did not do so.* When asked, half of all participants (P2, P3, P4, P7, P8, P12, P13) replied that they would like to ask the AI agent some questions for clarification (P2, P7), concept definition (P7), a hint on missing components (P8), or further exploration into a specific topic (P2). However, participants may have been too preoccupied with drafting answers to the open-ended questions that they had little mental capacity to think about

anything else (P3). This could have explained the unusually low utilization frequency of the Q&A agent (Table 3), with the only few calls being a result of participants' more inquisitive answer statements rather than conscious questioning (e.g. P3: "*no how would that work?*"; P5: "*Is it proteins?*").

6.4.8 (~) Participants express mixed preferences between Chat Mode and Quiz Mode. When asked about their preference between Chat Mode and Quiz Mode for their future science video learning, participants' responses are given in Table 4. (Participants were allowed to vote for "It depends" or "Mix of both" after making a binary selection between Chat Mode and Quiz Mode.)

The majority of participants prefer Chat Mode over Quiz Mode, because they feel they could learn more through verbalizing answers (P2, P11, P13), the feedback is more dynamic and customized (P3), the dialogs are easier to review (P5), and their knowledge retention will be longer (P7, P14). However, some participants still prefer Quiz Mode for taking significantly less time to complete (P6), being more straightforward (P8) and involving less manual input (P10).

Participants who think "It depends" have highlighted decisive factors like how abstract / conceptual the video topic is (P1, P7) and the level of participant's prior knowledge (P6) or interest (P9, P12). Participants who support "Mix of both" think it will produce a better balance of overall difficulty (P4) and be more similar to their usual learning experience at school (P14).

Table 4: Participant preferences between Chat Mode and Quiz Mode for science video learning.

Preference	Frequency	
It depends	5	29.4%
Chat Mode	7	41.2%
Quiz Mode	3	17.6%
Mix of both	2	11.8%
Total	17	100.0%

7 Discussion

7.1 Which to Choose: Chat-Based or Quiz-Based Learning?

Based on interviews and performance analysis of high school students, we observe that there is not a definitive better solution between chats and quizzes in facilitating science video learning. We did not observe significant difference on learning gains between the two conditions. Qualitative evidence collected during the learning activity and the follow-up interviews suggested that many students preferred to use chat tutor, and did gain new knowledge through back-and-forth conversations with the chat agent. Moreover, we observed qualitative evidence that for students who had in particularly lower pre-test scores, they seemed to benefit more from chat-based tutoring.

Our evidence suggests that chat-based learning is particularly useful for

- Internalizing learned knowledge through the construction of arguments (P2, P11, P13)
- Helping students grasp concepts that they struggle with (P6)
- Enhancing students' knowledge retention (P2, P6, P7, P11, P14)
- Keeping students focused throughout the learning process (P4, P5, P6, P9)
- Producing helpful reference materials when reviewing for tests and exams (P7, P12)

while quiz-based learning is more suitable for

- Introductory exploration / surface understanding of a science topic (P1, P4, P6, P11, P13)
- Reinforcement of knowledge recognition (rather than recall) in a straightforward manner (P1, P3, P10)
- Quick knowledge check-in's that can be done within a short amount of time (P4, P6, P7, P11, P12).

Since we observed a significant pre- to post-test gains in both conditions, it suggests that when teachers are limited on time, a carefully designed LLM pipeline could deliver pedagogically beneficial quiz questions which are ready for immediate use, e.g., on platforms that are already popular among teachers such as EdPuzzle [3]. Given that using AI-powered chat tutors requires more computational resources, it needs further evidence to advocate for their large-scale use.

7.2 Considerations for Designing LLM-Powered Chat Tutoring Agents

The implementation of AI agent in *ScienceChat* has exposed several pitfalls that could inform future designs.

7.2.1 Balancing between divergent and grounded discussions. During our preliminary testing, we found that out-of-the-box LLMs (like ChatGPT) had a strong tendency of going out of scope of the learning material when interacting with the user as a tutor. This is because the LLM knows virtually "anything" in the corpus that it was trained on, which far exceeds the subset of information that is accessible to high school students.

To suppress the LLM (GPT-4o in our case) from eliciting knowledge beyond what was intended by the teachers, we performed heavy prompt engineering and utilized RAG to impose constraints on its knowledge base. While effective, this had brought about the side effect of being overly strict on student evaluation, characterized by follow-up questions that ask for unnecessarily trivial details from the video (P3, P4, P7, P8, P12, P14).

For instance, P7 had the following interaction with the AI agent on the consequences of a volcanic eruption. In this case, the AI agent did not mark P7's answer as correct ("*it dropped the global climate by blocking out some of the sun*"; "*over time the climate cooled*") until they have explicitly mentioned that global temperature dropped by "*0.5 degrees celsius*". But in reality, the 0.5 °C might not have been a detail worthy of repeated searches back into the video and two additional turns of dialog. Although P7 did not complain about this event, they mentioned that they actually "*would prefer if [the AI agent] went beyond the video*" to "*further explain a concept that the video doesn't explain that much, maybe with more details*", so they can "*know more about the content*".

Question	How did the eruption of Mount Pinatubo influence the global climate?
Participant	it dropped the global climate by blocking out some of the sun
AI Agent	[COMMENT] Great start! You correctly mentioned that the eruption blocked sunlight. [QUESTION] Can you explore what exactly was released into the atmosphere during Mount Pinatubo's explosion that caused this effect?
Participant	sulfuric acid
AI Agent	[COMMENT] You're on the right track with mentioning sulfuric acid, as it played a significant role. [HINT] Think about what the eruption sent into the air in large quantities, around the timestamp 2:47 in the video.
Participant	sulfur dioxide
AI Agent	[COMMENT] Well done on identifying sulfur dioxide! That was an important component. [ANSWER] What you missed: The explosion ejected millions of tons of particles and gases into the atmosphere. [HINT] Consider how the presence of these particles and gases might have affected the Earth's climate over time, especially around timestamp 3:05 in the video.
Participant	over time the climate cooled
AI Agent	[COMMENT] Excellent observation! You've noted that the global climate cooled, which is a crucial effect. [HINT] Keep investigating this cooling effect as there any specific measurement of how much the Earth's temperature decreased after the eruption? You might want to review around timestamp 3:10 in the video.
Participant	(Participant rewinds the video manually to look for answers)
AI Agent	0.5 degrees celsius [COMMENT] Excellent job! You've effectively captured the significant aspects of the event, both the sulfur dioxide release and the global temperature drop. [END] (concluding statements omitted)

Therefore, it will require more investigation and experiment to find out the ideal balance between "staying relevant to the video" and "introducing more diverse and in-depth perspectives" for AI agents.

7.2.2 Preventing Unintended Revelation of Correct Answer. Despite our deliberate prompt engineering efforts to prevent the AI agent from giving away expected components unless instructed to, it would still occasionally include expected answers right inside the feedback message. Instead of giving the students a shortcut to completion, this actually led to greater confusion. For example, P14 noticed that *"In the intermediate feedback on some of the questions, the feedback itself was the right answer... I guess that was a hint, but I never mentioned a hint, nor is it asking you to go to this section (timestamp) like in some other hints."* The feedback in question is underlined below:

Question	What organisms are affected by genetic mutations?
Participant	Fungi, plants, animals, protists, bacteria, archae, and viruses
AI Agent	[COMMENT] Great list! You've mentioned a variety of organisms. <u>However, remember that mutations occur in anything with RNA or DNA.</u> [QUESTION] Can you think about what all these organisms you listed have in common at a genetic level? What do they all possess that enables mutations to happen?
Participant	RNA and DNA
AI Agent	[COMMENT] Well done! You've hit the key point by noting that all these organisms possess RNA or DNA, which allows for mutations. Excellent job identifying the common factor crucial for genetic mutations. [END] (concluding statements omitted)

This example highlights the challenge for AI agent to avoid both (1) overly conservative hints that are barely any thought-provoking at all, and (2) hints that give away the final answer directly without any thought process left to the student.

8 Limitation

Due to time limitation per study session, we dropped a third learning task in which participants would watch a science video without working on any problems at all. This added condition may reveal how significant the effects of learning with tutoring interventions actually are compared to without them. In this study, the student sample size is limited (14 participants) and we did not formally evaluate the effectiveness of Editor Mode interfaces with high school teachers. Further investigations are needed with larger and more comprehensive population coverage from the high school communities to reflect the overall usability of *ScienceChat*.

9 Conclusion

This paper aims to address the question of "to chat or to quiz" when providing generative AI-based tutoring for high school students to learn science videos. We carefully implemented a platform *ScienceChat* which provides two modes of tutoring, namely Chat Mode and Quiz Mode. Students would watch a science video and engage in back-and-forth conversations with an AI agent in Chat Mode, and answer multiple-choice questions in Quiz Mode. A within-subject experiment with 14 high school students showed that students had significant learning gains in both conditions, though we did not observe a difference between the two conditions. Through qualitative analysis of learner's dialogues and post-session interviews, we found strong evidence of student learning in Chat Mode. Many students also applauded the chat experience in enhancing knowledge retention, keeping them focused, and supporting them construct arguments in their own words. We suggest it requires further experiments to understand when and how LLM-powered chat-based tutoring can be safe, reliable and pedagogically beneficial. We recommend future LLM-powered chat tutoring system to consider the trade-off between being grounded in the conversation versus being more creative and divergent-thinking.

References

- [1] [n. d.]. Assistants API. <https://platform.openai.com/docs/assistants/overview>. Accessed July 15, 2024.
- [2] [n. d.]. Django. <https://www.djangoproject.com/>
- [3] [n. d.]. Edpuzzle. <https://edpuzzle.com/>. Accessed Sep 12, 2024.
- [4] [n. d.]. File Search. <https://platform.openai.com/docs/assistants/tools/file-search>
- [5] [n. d.]. Next Generation Science Standards. <https://www.nextgenscience.org/>. Accessed July 15, 2024.
- [6] [n. d.]. React: The library for web and native user interfaces. <https://react.dev/>
- [7] Amoeba Sisters. 2019. Mutations (Updated). <https://www.youtube.com/watch?v=vl6Vlf2thvI>. Accessed Sept. 9, 2024.
- [8] John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1995. Cognitive tutors: Lessons learned. *The journal of the learning sciences* 4, 2 (1995), 167–207.
- [9] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051* (2022).
- [10] Chenxi Dong. 2023. How to build an AI tutor that can adapt to any course and provide accurate answers using large language model and retrieval-augmented generation. *arXiv preprint arXiv:2311.17696* (2023).
- [11] Andy Extance. 2023. ChatGPT has entered the classroom: how LLMs could transform education. *Nature* 623, 7987 (2023), 474–477.
- [12] Arthur C Graesser, Sidney D Mello, Xianguan Hu, Zhiqiang Cai, Andrew Olney, and Brent Morgan. 2012. AutoTutor. In *Applied natural language processing: Identification, investigation and resolution*. IGI Global, 169–187.
- [13] Arthur C Graesser, Sidney D Mello, and Natalie Person. 2009. Meta-knowledge in tutoring. In *Handbook of metacognition in education*. Routledge, 361–382.
- [14] Arthur C. Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M. Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers* 36, 2 (01 May 2004), 180–192. <https://doi.org/10.3758/BF03195563>
- [15] Tony Harland. 2003. Vygotsky's zone of proximal development and problem-based learning: Linking a theoretical concept with practice through action research. *Teaching in higher education* 8, 2 (2003), 263–272.
- [16] Sebastian Hobert. 2019. Say Hello to 'Coding Tutor'! Design and Evaluation of a Chatbot-based Learning System Supporting Students to Learn to Program.
- [17] Sibel Kazak, Rupert Wegerif, and Taro Fujita. 2015. Combining scaffolding for content and scaffolding for dialogue to support conceptual breakthroughs in understanding probability. *ZDM* 47 (2015), 1269–1283.
- [18] Kenneth R Koedinger, John R Anderson, William H Hadley, and Mary A Mark. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education* 8 (1997), 30–43.
- [19] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.
- [20] James A Kulik and John D Fletcher. 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research* 86, 1 (2016), 42–78.
- [21] James A. Kulik and J. D. Fletcher. 2016. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research* 86, 1 (2016), 42–78. <https://doi.org/10.3102/0034654315581420> arXiv:<https://doi.org/10.3102/0034654315581420>
- [22] Kurzgesagt – In a Nutshell. 2020. Geoen지니어링: A Horrible Idea We Might Have to Do. <https://www.youtube.com/watch?v=dSu5sXmsur4>
- [23] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. Dapie: Interactive step-by-step explanatory dialogues to answer children's why and how questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22.
- [24] Jonathan Leo and Kelly Puzio. 2016. Flipped instruction in a high school science classroom. *Journal of Science Education and Technology* 25 (2016), 775–781.
- [25] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [26] Mark Liffiton, Brad E Sheese, Jaromir Savelka, and Paul Denny. 2023. Codehelp: Using large language models with guardrails for scalable support in programming classes. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*. 1–11.
- [27] Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [28] Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. GPTEach: Interactive TA Training with GPT Based Students. (2023).
- [29] Elham Mousavinasab, Nahid Zarifasanaei, Sharareh R. Niakan Kalhori, Mahnaz Rakhshan, Leila Keikha, and Marjan Ghazi Saeedi. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments* 29, 1 (2021), 142–163.
- [30] Pedro J. Muñoz-Merino, Manuel Fernández Molina, Mario Muñoz-Organero, and Carlos Delgado Kloos. 2012. An adaptive and innovative question-driven competition-based intelligent tutoring system for learning. *Expert Systems with Applications* 39, 8 (2012), 6932–6948. <https://doi.org/10.1016/j.eswa.2012.01.020>
- [31] Huy A Nguyen, Hayden Stec, Xinying Hou, Sarah Di, and Bruce M McLaren. 2023. Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European Conference on Technology Enhanced Learning*. Springer, 278–293.
- [32] Andre Nickow, Philip Oreopoulos, and Vincent Quan. 2020. *The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence*. Working Paper 27476. National Bureau of Economic Research. <https://doi.org/10.3386/w27476>
- [33] Stellan Ohlsson. 2016. Constraint-based modeling: from cognitive theory to computer tutoring—and back again. *International Journal of Artificial Intelligence in Education* 26 (2016), 457–473.
- [34] Graziella Orrù, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. 2023. Human-like problem-solving abilities in large language models using ChatGPT. *Frontiers in Artificial Intelligence* 6 (2023), 1199350.
- [35] Zachary A Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871* (2023).
- [36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [37] Alan Ramirez-Noriega, Reyes Juárez-Ramírez, and Yobani Martínez-Ramírez. 2017. Evaluation module based on Bayesian networks to Intelligent Tutoring Systems. *International Journal of Information Management* 37, 1, Part A (2017), 1488–1498. <https://doi.org/10.1016/j.ijinfomgt.2016.05.007>
- [38] Genghu Shi, Anne M. Lippert, Keith Shubeck, Ying Fang, Su Chen, Philip Pavlik, Daphne Greenberg, and Arthur C. Graesser. 2018. Exploring an intelligent tutoring system as a conversation-based assessment tool for reading comprehension. *Behaviormetrika* 45, 2 (01 Oct 2018), 615–633. <https://doi.org/10.1007/s41237-018-0065-9>
- [39] Alyson Simpson. 2016. Designing pedagogic strategies for dialogic learning in higher education. *Technology, Pedagogy and Education* 25, 2 (2016), 135–151.
- [40] John Stamper, Ruiwei Xiao, and Xinying Hou. 2024. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education*. Springer, 32–43.
- [41] Siriwan Suebnukarn and Peter Haddawy. 2004. A collaborative intelligent tutoring system for medical problem-based learning. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira, Portugal) (IUI '04). Association for Computing Machinery, New York, NY, USA, 14–21. <https://doi.org/10.1145/964442.964447>
- [42] Janneke Van de Pol, Monique Volman, and Jos Beishuizen. 2010. Scaffolding in teacher–student interaction: A decade of research. *Educational psychology review* 22 (2010), 271–296.
- [43] KURT VanLEHN. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist* 46, 4 (2011), 197–221. <https://doi.org/10.1080/00461520.2011.611369> arXiv:<https://doi.org/10.1080/00461520.2011.611369>
- [44] Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2174–2199.
- [45] Wayne Ward, Ronald Cole, Daniel Bolanos, Cindy Buchenroth-Martin, Edward Svirsky, Sarel Van Vuuren, Timothy Weston, Jing Zheng, and Lee Becker. 2011. My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing (TSLP)* 7, 4 (2011), 1–29.
- [46] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682 [cs.CL] <https://arxiv.org/abs/2206.07682>
- [47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837.
- [48] Ursula Wingate. 2019. 'Can you talk me through your argument'? Features of dialogic interaction in academic writing tutorials. *Journal of English for Academic Purposes* 38 (2019), 25–35.
- [49] Vanessa L Wyss, Diane Heulskamp, and Cathy J Siebert. 2012. Increasing middle school student interest in STEM careers with videos of scientists. *International journal of environmental and science education* 7, 4 (2012), 501–522.

A ScienceChat System Prompts

A.1 Assistant Persona

Function	The system instruction for <i>video assistant</i> and <i>grading component assistant</i> , which defines their behavior. This prompt is the global "prefix" for other task-specific assistant prompts.
Parameters	{m_title}: The video title {m_standard}: The video learning standards

You are a middle school teacher teaching Grades 9-12 who want to assign a learning media video titled "{m_title}" to your students for an upcoming science class discussion session. The video transcript is provided as a file search attachment and please review it in its entirety.

If the user provided video learning standards

The NGSS learning standard(s) of this video are given below.

[[LEARNING STANDARD BEGIN]]

{m_standard}

[[LEARNING STANDARD END]]

End if

A.2 Substance Check

Function	Checks whether the student response has shown any effort in answering the question.
Model	gpt-4o-mini (chat completion; embedded full transcript context)
Parameters	{m_transcript}: The video transcript {m_question}: The open-ended question stem {m_student_response}: The student-submitted response

[[TRANSCRIPT BEGIN]]
{m_transcript}
[[TRANSCRIPT END]]

To facilitate active learning, collaborative thinking and science communication, you have created the following open-ended question for students to discuss in class based on the video contents:

* Open-Ended Question: {m_question}

Based on your understanding of the video transcript and the open-ended question, please decide whether the following student response shows any effort towards answering the question.

* Student Response: {m_student_response}

Your decision must be rendered as one single lower-case letter.

- + If the student response shows at least some amount of effort, please respond with "t" (for "true"). It is perfectly acceptable if:
 - The student's answer is not 100% correct or comprehensive and still has space for improvement.
 - Based on prior dialog, the student raises a new question instead of answering the existing one.
- + If the student response does not show any effort (like passively saying "I don't know"), please respond with "f" (for "false").

Your lesson plan requires that you respond only with "t" or "f", without quotation marks, new lines or additional reasoning for your decision. Do not include anything else.

A.3 Question Check

Function	Checks whether the student is asking a question (instead of giving a statement).
Model	gpt-4o-mini (chat completion; embedded full transcript context)
Parameters	{m_transcript}: The video transcript {m_question}: The open-ended question stem {m_student_response}: The student-submitted response

[[TRANSCRIPT BEGIN]]
{m_transcript}
[[TRANSCRIPT END]]

To facilitate active learning, collaborative thinking and science communication, you have created the following open-ended question for students to discuss in class based on the video contents:

* Open-Ended Question: {m_question}

Based on your understanding of the video transcript and the open-ended question, please decide whether the following student response is giving an answer to the original question or asking a new question. Sometimes, students may ask you back for clarifications or about certain concepts they are unsure about before they make an attempt to answer.

* Student Response: {m_student_response}

Your decision must be rendered as one single lower-case letter.

- + If the student is explicitly asking a question instead of giving an answer, please respond with "t" (for "true").
- + If the student response is not explicitly asking a question, please respond with "f" (for "false").
- + If the student is asking but their question is too similar to the original open-ended question (such that if you, the teacher, answers it directly, you will give away the correct answer), please also respond with "f".

Your lesson plan requires that you respond only with "t" or "f", without quotation marks, new lines or additional reasoning for your decision. Do not include anything else.

A.4 Rubric Evaluation

Function	Checks whether the student have by far fulfilled a given grading component.
Model	gpt-4o (chat completion; embedded grading component context)
Parameters	{m_rubric_context}: The context of current grading component (extracted from video transcript) {m_question}: The open-ended question stem {m_rubric}: The current grading component {r_dialog}: The dialog history {m_student_response}: The student-submitted response

[[TRANSCRIPT EXCERPT BEGIN]]
{m_rubric_context}
[[TRANSCRIPT EXCERPT END]]

To facilitate active learning, collaborative thinking and science communication, you have created the following open-ended question for students to discuss in class based on the video contents:

* Open-Ended Question: {m_question}

Based on your understanding of the video transcript and the open-ended question, and taking into account the discussion history, please decide whether the student has mentioned the following grading component up to their latest response (if there are terms enclosed in square brackets with vertical bars in between, it means that they are equivalent and mentioning any one of them is sufficient).

* Grading Component: {m_rubric}

On the first line, provide a brief and concise explanation for whether or not you think the student has fulfilled the grading component "{m_rubric}".

On the second line, based on your argument above, render your decision as one single lower-case letter (without quotation marks).

- + If the student has already mentioned the grading component, please use "t" (for "true").
- + If the student has not yet mentioned the grading component and you will need to ask more about it later, please use "f" (for "false").

Your lesson plan requires you to follow the above rules strictly and only contain two lines in your response. Do not include anything else.

The discussion history is given below (do not use the tag [TEACHER] or [STUDENT] in your response). In addition to all the above instructions, as a bottom line, your response should keep the conversation going easy and smooth for the student, like a relaxed chat rather than a rigid and concise exam problem. You should avoid pushing for an answer, being overly critical or making abrupt jumps in context. If necessary, you can use a sentence or two to transition from what the student has just said.

{r_dialog}

[STUDENT] {m_student_response}

A.5 Hint Agent

Function	Generates an easier hint question based on a given rubric (if the student response failed the substance check).
Model	gpt-4o (knowledge base retrieval; <i>grading component assistant</i> with grading component context)
Parameters	{m_question}: The open-ended question stem {m_rubric}: The current grading component {m_done_rubric}: The grading component(s) that student has fulfilled {m_expired_rubric}: The grading component that student did not fulfill after maximum attempts {response_length_limit=60}: The expected upper limit of agent output length (in words) {r_dialog}: The dialog history {m_student_response}: The student-submitted response

To facilitate active learning, collaborative thinking and science communication, you have created the following open-ended question for students to discuss in class based on the video contents:

* Open-Ended Question: {m_question}

By this time, the student has fulfilled the following grading components:

* Grading Component: {m_done_rubric}

If the student has failed the last component

Through your discussion with the student, you realized that they are still having some difficulties satisfying the following grading component after a hint. So you decide to reveal this part of the answer to them and explicitly pinpoint what exactly they were missing based on the video contents, then move on to the next unfulfilled grading component.

* Grading Component: {m_expired_rubric}

You must use "[ANSWER] ... [ANSWER]" (without quotation marks) to enclose the revelation and explanation part of your response.

After doing the explanation, you want to lead the student to focus on the following grading component.

* Grading Component: {m_rubric}

With reference to what the student has said, comment on their response and current progress, then provide the student with an approachable and thought-provoking hint to help them pinpoint what exactly they were missing.

Otherwise

Through your discussion with the student, you have identified that they may be having some difficulties satisfying the following grading component.

* Grading Component: {m_rubric}

(({m_rubric.minute}:{m_rubric.second}))

With reference to what the student has said, comment on their response and current progress, provide them with an approachable and thought-provoking hint to help them pinpoint what exactly they were missing, and remind them to review the video content at the rubric's timestamp shown above (in the exact same format, without parentheses).

End if

You must use "[COMMENT] ... [COMMENT]" (without quotation marks) to enclose the comment part of your response. You must use "[HINT] ... [HINT]" (without quotation marks) to enclose the hint and timestamp part of your response.

Your hint content should target only the grading component above and should refrain from diverging to other topics. You must not reveal the final correct answer (or any key concept or argument from the correct answer) to the open-ended question in any way, either directly or indirectly. In other words, what you respond must not be identical or closely similar to the grading component.

Your lesson plan requires that you do not write down anything other than the content of your response, on a single line and without quotation marks. Your response must not exceed {response_length_limit} words, but it must still meet all standards above. Please draft in this exact format.

The discussion history is given below (do not use the tag [TEACHER] or [STUDENT] in your response). In addition to all the above instructions, as a bottom line, your response should keep the conversation going easy and smooth for the student, like a relaxed chat rather than a rigid and concise exam problem. You should avoid pushing

for an answer, being overly critical or making abrupt jumps in context. If necessary, you can use a sentence or two to transition from what the student has just said.

```
{r_dialog}
[STUDENT] {m_student_response}
```

A.6 Q&A Agent

Function	Answers the student's question (if the student response passed both substance and question checks).
Model	gpt-4o (knowledge base retrieval; <i>video assistant</i> with full transcript context)
Parameters	{m_question}: The open-ended question stem {response_length_limit=60}: The expected upper limit of agent output length (in words) {r_dialog}: The dialog history {m_student_response}: The student-submitted response

To facilitate active learning, collaborative thinking and science communication, you have created the following open-ended question for students to discuss in class based on the video contents:

* Open-Ended Question: {m_question}

Through your discussion with one of the students, you have identified that they may be having some difficulties approaching this open-ended question. Hence, they are asking you back about something they want to know about or for certain clarifications on the question.

Please analyze what might be the student's confusion (you may or may not mention it explicitly in your response) and respond to their question in a way understandable for middle school students.

You must use "[EXPLAIN] ... [EXPLAIN]" (without quotation marks) to enclose the Q&A part of your response.

Your response should focus entirely on what the student has asked and refrain from diverging to other topics. You must not reveal the final correct answer (or any key concept or argument from the correct answer) to the open-ended question in any way, either directly or indirectly. In other words, what you respond must not be identical or closely similar to the grading component.

Your lesson plan requires that you do not write down anything other than the content of your response, on a single line and without quotation marks. Your response must not exceed {response_length_limit} words, but it must still meet all standards above. Please draft in this exact format.

The discussion history is given below (do not use the tag [TEACHER] or [STUDENT] in your response). In addition to all the above instructions, as a bottom line, your response should keep the conversation going easy and smooth for the student, like a relaxed chat rather than a rigid and concise exam problem. You should avoid pushing for an answer, being overly critical or making abrupt jumps in context. If necessary, you can use a sentence or two to transition from what the student has just said.

```
{r_dialog}
[STUDENT] {m_student_response}
```

A.7 Rubric Agent

Function	Generates a follow-up question based on a given rubric.
Model	gpt-4o (knowledge base retrieval; <i>grading component assistant</i> with grading component context)
Parameters	{m_question}: The open-ended question stem {m_rubric}: The current grading component {m_done_rubric}: The grading component(s) that student has fulfilled {m_expired_rubric}: The grading component that student did not fulfill after maximum attempts {response_length_limit=60}: The expected upper limit of agent output length (in words) {r_dialog}: The dialog history {m_student_response}: The student-submitted response

To facilitate active learning, collaborative thinking and science communication, you have created the following open-ended question for students to discuss in class based on the video contents:

* Open-Ended Question: {m_question}

By this time, the student has fulfilled the following grading components:

* Grading Component: {m_done_rubric}

If the student has failed the last component

Through your discussion with the student, you realized that they still have not satisfied the following grading component after quite a few tries. So you decide to reveal this part of the answer to them and explicitly pinpoint what exactly they were missing based on the video contents, then move on to the next unfulfilled grading component.

* Grading Component: {m_expired_rubric}

You must use "[ANSWER] ... [ANSWER]" (without quotation marks) to enclose the revelation and explanation part of your response.

After doing the explanation, you want to lead the student to focus on the following grading component.

* Grading Component: {m_rubric}

Otherwise

Through your discussion with the student, you have identified that their response so far has not satisfied (either insufficient, incorrect or missing) the following grading component.

* Grading Component: {m_rubric}

End if

Please comment on the student's response and current progress, then ask them a follow-up question based on this grading component to guide them to pinpoint what they were missing and amend their answer. The follow-up question must be explicit (i.e. it ends with a question mark) and there can only be one of it to avoid confusion.

You must use "[COMMENT] ... [COMMENT]" (without quotation marks) to enclose the comment part of your response. You must use "[QUESTION] ... [QUESTION]" (without quotation marks) to enclose the question part of your response.

Your follow-up question should target only the grading component above and should refrain from diverging to other topics.

You must not reveal the final correct answer to the open-ended question or any grading component in any way, either directly or indirectly. In other words, your response must not be identical or closely similar to any grading component.

Your lesson plan requires that you do not write down anything other than the content of your response, on a single line and without quotation marks. Your response must not exceed `{response_length_limit}` words, but it must still meet all standards above. Please draft in this exact format.

The discussion history is given below (do not use the tag `[TEACHER]` or `[STUDENT]` in your response). In addition to all the above instructions, as a bottom line, your response should keep the conversation going easy and smooth for the student, like a relaxed chat rather than a rigid and concise exam problem. You should avoid pushing for an answer, being overly critical or making abrupt jumps in context. If necessary, you can use a sentence or two to transition from what the student has just said.

```
{r_dialog}
[STUDENT] {m_student_response}
```

A.8 End of Dialog

Function	Generates a concluding statement to the dialog.
Model	gpt-4o (knowledge base retrieval; <i>video assistant</i> with full transcript context)
Parameters	<code>{m_question}</code> : The open-ended question stem <code>{m_rubric}</code> : The current grading component <code>{m_done_rubric}</code> : The grading component(s) that student has fulfilled <code>{m_expired_rubric}</code> : The grading component that student did not fulfill after maximum attempts <code>{response_length_limit=60}</code> : The expected upper limit of agent output length (in words) <code>{r_dialog}</code> : The dialog history <code>{m_student_response}</code> : The student-submitted response

To facilitate active learning, collaborative thinking and science communication, you have created the following open-ended question for students to discuss in class based on the video contents:

* Open-Ended Question: `{m_question}`

By this time, the student has fulfilled the following grading components:

* Grading Component: `{m_done_rubric}`

If the student has failed the last component

Through your discussion with the student, you realized that they still have not satisfied the following grading component after quite a few tries. So you decide to reveal this part of the answer to them and explicitly pinpoint what exactly they were missing based on the video contents, then conclude the discussion.

* Grading Component: `{m_expired_rubric}`

You must use `"[ANSWER] ... [ANSWER]"` (without quotation marks) to enclose the revelation and explanation part of your response.

Otherwise

Through your discussion with the student, you have identified that they have satisfied all grading components for this question.

End if

Please comment on the student's response and overall performance, then conclude the discussion of this question. Do not ask the student any new questions at this point.

You must use `"[COMMENT] ... [COMMENT]"` (without quotation marks) to enclose the comment part of your response. You must use `"[END] ... [END]"` (without quotation marks) to enclose the conclusion part of your response.

Your follow-up question should target only the grading component above and should refrain from diverging to other topics. You must not reveal the final correct answer to the open-ended question or any grading component in any way, either directly or indirectly. In other words, your response must not be identical or closely similar to any grading component.

Your lesson plan requires that you do not write down anything other than the content of your response, on a single line and without quotation marks. Your response must not exceed `{response_length_limit}` words, but it must still meet all standards above. Please draft in this exact format.

The discussion history is given below (do not use the tag `[TEACHER]` or `[STUDENT]` in your response). In addition to all the above instructions, as a bottom line, your response should keep the conversation going easy and smooth for the student, like a relaxed chat rather than a rigid and concise exam problem. You should avoid pushing for an answer, being overly critical or making abrupt jumps in context. If necessary, you can use a sentence or two to transition from what the student has just said.

```
{r_dialog}
[STUDENT] {m_student_response}
```

A.9 Context Extraction

Function	Finds which part(s) of the source material is relevant to the target material.
Model	gpt-4o [knowledge base retrieval; <i>context assistant</i> with full transcript context (as source material) and temperature=0]
Parameter	<code>{m_target}</code> : The target material

You are a professional context analyst working in the education domain. Given a source material and a target material, your job is to find section(s) of the source material that is the most relevant to the target material, if any. The results you find in the source material must be reported verbatim in your response.

The source material is provided (as a file search attachment) and please review it in its entirety. More detailed instructions and the target material will be provided to you later.

With reference to the source material available to you, please find all relevant context of the following target material.

* Target Material: `{m_target}`

Please refer to the following requirements when formulating your response.

- + If you find at least one section of the source material that is directly relevant to the target material:
 - Report all segment(s) of the source material that you find verbatim. Do not modify or summarize in any way.
 - If there are multiple continuous segments, add a single line break between them.

- Do not add numbering or leading titles to distinguish between segments.
- Do not add inline citation (i.e. the source file name and line number in brackets).
- + If you find that nothing in the source material is directly relevant to the target material:
 - Please respond with a single letter "x" (for "unknown"), without quotation marks, new lines or additional reasoning for your decision.

Your are required to follow the above rules strictly. Do not include anything else.

A.10 Converting Open-Ended Question into Multiple Choice Question

Function	Generates a multiple choice question based on an open-ended question.
Model	gpt-4o (knowledge base retrieval; <i>video assistant</i> with full transcript context)
Parameters	<code>{m_question}</code> : The open-ended question stem <code>{m_rubric}</code> : All grading components of the open-ended question

To facilitate active learning, collaborative thinking and science communication, you have created the following open-ended question for students to discuss in class based on the video contents:

* Open-Ended Question: `{m_question}`

The question has the following grading components (if there are terms enclosed in square brackets with vertical bars in between, it means that they are equivalent and mentioning any one of them is sufficient):

* Grading Component: `{m_rubric}`

Although this question is able to fulfill your teaching needs, you realize that not all students can receive instant feedback from you due to the question's open-ended nature. Therefore, you also want to write a corresponding multiple choice question for this open-ended question to provide students with instant, helpful feedback.

Your multiple choice question should help students further grasp the core knowledge covered by this open-ended question, but in a less involved and more approachable way. You want to avoid asking for simple information recall from the video; instead, you want the question stem and options to provoke students' higher-level thinking and deepen their comprehension of relevant knowledge and concept.

To avoid confusion, you need to make sure that the incorrect options are "strictly incorrect" according to the video content, as opposed to statements that are unmentioned, partially wrong or only less ideal than the correct one. Meanwhile, you must try your best to formulate these "strictly incorrect" options as distractors to promote your students' critical thinking.

Please carefully review the video transcript, NGSS standards (if any) and the learning goals above. Based on your understanding, please present your multiple choice question in the following format:

- + On the first line, write down at which time in the video the students should be asked this question.

- For example, 3 minutes 5 seconds should be written as "3:05" (without quotation marks).
- The question should only be asked after all necessary information have been discussed in the video. Please approximate an ideal time based on the timestamps embedded in the video transcript (VTT / SRT format) to the best of your ability. If appropriate, you may also use the original question time (`{m_question.minute}`;`{m_question.second}`).
- + On the second line, write down the question stem of this multiple choice.
- + On each subsequent line, write down one concise option for this multiple choice. Append each of the following segments together (with one whitespace in between) on one single line:
 - First, indicate whether this option is correct or not. Use "(t)" for correct and "(f)" for incorrect (without quotation marks).
 - Next, enclose the content of this option in square brackets "[...]" (without quotation marks).
 - Lastly, enclose the student feedback for this option in curly brackets "{...}" (without quotation marks). You should explain why exactly this option is right or wrong in no more than 30 words. Remember to keep your wording accessible for all levels of understanding.
 - It is up to you to decide how many options and correct options you need, but the total number of options must not exceed 4.

Your lesson plan requires you to follow the above rules strictly. Do not include anything else.