JOURNAL OF CHEMICAL EDUCATION

Editors' Choice

Open Access

pubs.acs.org/jchemeduc

Article

Let Us Not Squander the Affordances of LLMs for the Sake of Expedience: Using Retrieval Augmented Generative AI Chatbots to Support and Evaluate Student Reasoning

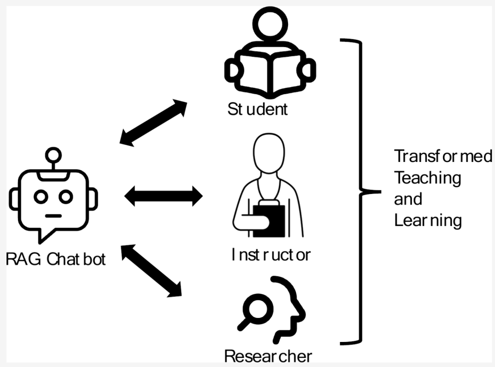
Melanie M. Cooper* and Michael W. Klymkowsky

Cite This: https://doi.org/10.1021/acs.jchemed.4c00765

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The use of large language model Generative AI (GenAI) systems by students and instructors is increasing rapidly, and there is little choice but to adapt to this new situation. Many, but not all, students are using GenAI for homework and assignments, which means that we need to provide equitable access for all students to AI systems that can support and enhance their learning. At the same time, we need to think carefully about just what we want teaching and learning to look like as GenAI systems become readily available. Here we propose that "business as usual" is not a responsible option. Although chatbots can readily answer questions, produce summaries of content, and make the process of education more efficient, there is scant evidence that such time saving is effective, and indeed, it is important that we not allow the use of GenAI systems to circumvent or undermine the learning process. The availability of so-called Retrieval Augmented Generative (RAG) AI systems allows us to expand what we expect students to know and do, by 1) supporting instructors in the design of more complex tasks (that can, for example, elicit evidence of three-dimensional learning (3DL)), 2) supporting students as they reason through such scaffolded tasks, and 3) by evaluating student responses, individually and in aggregate. We present examples of each of these affordances with the associated training materials and bot personas, along with caveats about their use.



KEYWORDS: *First-Year Undergraduate, General, Curriculum, Generative AI, Learning Theories, Student Centered Learning*

INTRODUCTION

In the short time that large language model generative AI (GenAI) tools have been available they have generated a furor in the educational community. Responses have ranged from concerns about how to stop students using them to cheat on assignments, to suggestions for various ways that they can be used to support learning.[1] A recent Google Scholar search for "ChatGPT in education" produced about 85,000 publications; narrowing the search to "ChatGPT in chemistry education" reduced the hits to around 11,000. In 2023 over 40 papers were published in the *Journal of Chemical Education* alone on the use of generative artificial intelligence. Clearly there is immense interest in, and numerous suggestions for, how to best use these generative AI-based tools.[2−4]

Those opposed to allowing the use of GenAI-based tools in the classroom point to the fact that ChatGPT can invent or hallucinate ideas (including sources)[4,5] and there is great concern that offloading course work to ChatGPT or other GenAI-based tools (e.g., Anthropic's Claude, Google's Gemini, Meta's Llama etc.) is "cheating".[5] Others have had a more positive approach to the educational impacts of AI-based tools. They have considered how the affordances offered by these tools can be used to support instructional practices and assessments. Among the suggestions for how we might use AI-based tools are asking students to critique the responses that GenAI-bots generate, compare AI generated responses to human generated responses, or to discuss personally relevant ideas in their responses.[3,6−8] There are also "structural impacts" to consider; for example, GenAI-based tools can grade student work, ranging from lab reports[9] to open-ended responses to assessment items and provide feedback to students and instructors.[2]

However, as noted earlier, there are significant issues with using these LLM systems: they tend to hallucinate (that is make up answers to queries), answers are not always relevant or appropriate (since they are trained on large data sets some of this information is incorrect), and there is a significant risk to privacy when using them (any input can be used to train the overall model). This makes the current uses of LLMs to support education quite problematic. However, there is an alternative approach.

Received: June 24, 2024
Revised: September 26, 2024
Accepted: September 26, 2024

Retrieval Augmented Generative (RAG) AI Systems Can Mitigate Many of the Problems with GenAI Systems

As noted, much of this "early" work on GenAI-based learning tools, specifically interactive chatbots, has relied on large language models (LLMs), such as ChatGPT, and subsequent training of these models for specific purposes. There has, as yet, been little discussion about custom chatbots that are confined to content specified by the designer. That is, rather than adopting the broad and overarching affordances and weaknesses of ChatGPT itself, narrower, and deeper chatbots based on desired content seem better suited and more appropriate for use in education. In particular Retrieval Augmented Generative (RAG) GenAI systems[10] have extraordinary potential for several reasons:

1) they use author/designer supplied content materials and data to generate their responses,
2) the student inputs are not used to train a GenAI model. Student responses are confidential, secure, and are not fed into a generative AI model that is then used by others (as is the case with ChatGPT).
3) RAG-AI systems can also specify, when needed, where the information they are using came from, and
4) RAG systems are much less prone to "hallucinating" or to put it less politely, producing BS.[11] The features of RAG-AI-based systems combine to make them more customizable and dependable for educational purposes. In addition, it is now possible to design these customized RAG-Chatbots without extensive and specific technical knowledge, and it is relatively simple to access the available design tools (see below).[12]

Figure 1a shows a typical GenAI Chatbot such as those based on ChatGPT, which is typically trained on the vast swathes of data scraped from the Internet; it is often retrained by the user by adding relevant examples and other instructions to customize the output. Even with this iterative training and prompt engineering, hallucinations are likely, and thus, it is not possible to ensure that the chatbot responses are accurate. It is also not possible to determine where the bot procured the information used in its response. In contrast, a RAG GenAI system uses only materials supplied by the developer as the content source for any responses.[10] Figure 1b shows how the input query is directed to the users' content (in our case chemistry course materials), by what is essentially a "digital librarian". When the question or task is posed, the RAG system looks for the answers in the user defined content, the information is then forwarded to the GenAI system which adds the ability to communicate with the user in a variety of formats and can be trained with the instructions on how the bot should respond to the prompt.

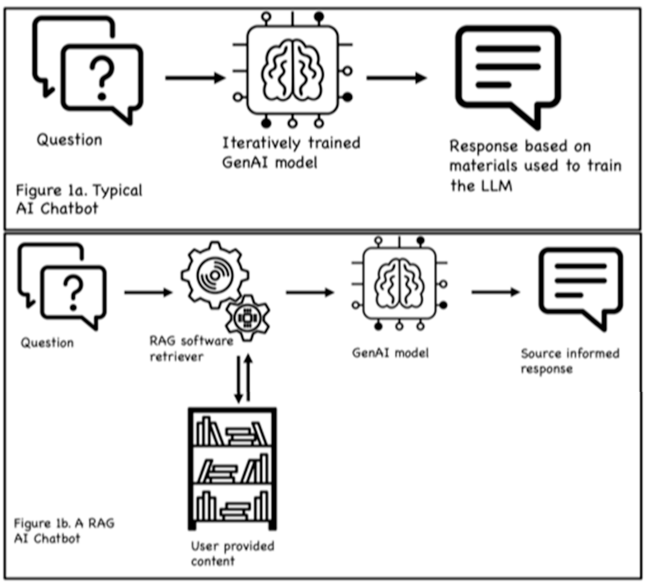


Figure 1. A typical data flow for a GenAI Chatbot and a RAG-AI Chatbot

Current Approaches to AI Supported Teaching and Learning

There are already a number of examples of GenAI supported teaching and learning tools offered by publishers and commercial entities.[13−16] Currently many are aimed at the K-12 system, but those that are being introduced for higher education tend to be focused on rather traditional curricula, the development of skills, and the learning of facts.[15,16] For example, instructors can upload lecture slides and an AI system can generate multiple choice questions from those slides,[16] and NotebookLM[17] can organize and synthesize large amounts of study materials. However, using GenAI to teach and support students doing the same old tasks seems like a missed opportunity. Certainly, students need skills and facts, but these should be learned in service of something more meaningful. For example, learning to draw structures should not be an end goal, but rather a means to learning how to use such structures to predict properties and reactivity.[18,19] If we use the powerful affordances of GenAI merely to generate multiple choice questions that further contribute to the fragmentation of students' knowledge,[20] we would not only waste resources, but it also implies a tragic lack of vision.

Is There a Better Way?

The famous (or infamous) motto of Silicon Valley is "move fast and break things",[21] which may be appropriate when developing software, but when we are dealing with the education of human beings, is problematic. We have very little information about how students interact with AI chatbots, but we do know that a good proportion (but certainly not all) of our students are already taking advantage of the affordances of AI-based systems to support their work. It makes sense that we as instructors and curriculum developers should also take advantage of those affordances, and seriously consider what we want to emphasize and reinforce in our teaching and learning efforts. It will also be important to ensure that the use of such systems is available to all, which means that institutions and instructors will need to intentionally incorporate GenAI access and use into instruction. Continuing with business as usual is neither realistic nor appropriate: students will have the support of AI systems to learn content and skills, but how will we know whether it is the GenAI system or the student who is learning if we use these same skills and knowledge that they gleaned from AI systems, to assess the student? We propose that a better way to use these tools is to help us design and implement learning experiences that allow students to use their knowledge (rather than just know it). That is, GenAI can help us go beyond the traditional approaches to teaching science to support student sensemaking and reasoning, rather than determining the molecular shape of such esoteric entities as IF_5.

https://doi.org/10.1021/acs.jchemed.4c00765
J. Chem. Educ. XXXX, XXX, XXX−XXX

Three-Dimensional Learning

In our work we have emphasized three-dimensional learning (3DL), a vision first articulated in the National Academies consensus report A Framework for K-12 Science Education.[22] This approach involves connecting content to disciplinary core ideas (DCIs) by engaging with scientific and engineering practices (SEPs), and crosscutting concepts (CCCs) to investigate, make sense of, and explain scientific phenomena. Curricula designed using this approach[23,24] have been shown to be more equitable that traditional curricula,[25] and to improve student performance across a wide range of activities including determining structure property relationships,[26] mechanistic reasoning,[27−30] predicting intermolecular forces,[31] and drawing organic mechanisms.[32] One challenge to the uptake of such transformations is the need for students to construct explanations and arguments for themselves. Although asking deep explanatory questions is one of the few pedagogical approaches for which there is a strong evidence base[33] (that is multiple studies across many institutions), such tasks are often omitted because of the practical difficulties of grading and providing feedback, especially for large numbers of students. Given the wealth of evidence that supports the approach, it makes sense that we investigate the affordances of generative AI systems to support students in their construction of explanations and to support faculty in the development and assessment of student responses to such activities. In this paper we present a potential approach to such activities.

■ RAG CHATBOT DEVELOPMENT

Here we present examples of chatbots that could be used both to support student reasoning, and to assess it; these include bots that can: 1a) analyze, code or grade inputs such as student responses to open-ended questions, 1b) aggregate student responses to provide the instructor with a snapshot of student reasoning; 2) generate complex activities and assessments that meet the instructors' learning goals; and 3) serve as a tutor or teacher providing feedback and support, while also making suggestions and asking questions that help students connect and construct more complete responses.

The resources for constructing a RAG chatbot are readily available from GitHub.[34] Their use, however, typically involves coding and use of the API from openAI's ChatGPT itself, all of which requires some technical expertise. For many (including us) this is not a realistic option. In our work we have used the commercial CustomGPT system,[12] which requires no programming and has been found in comparison tests to provide appropriate responses more consistently than OpenAI's custom bot.[35] No doubt such systems' availability and capabilities will change dramatically over the next few years (or even months); however, we believe that the information we can glean about how to design such bots and interpret the data that emerge can guide future efforts.

In this section we outline the nontechnical steps needed to develop such a bot, and in subsequent sections provide examples of specific bots for a range of purposes.

- Step 1: Provide the RAG system with the content it will use to construct answers to queries. In our work on chemistry focused bots we have typically used our open-source texts: Chemistry, Life the Universe, and Everything (CLUE)[36] and its organic counterpart (OCLUE),[37] along with various of our publications relating to aspects of the bot's purpose (for example, a bot designed to code student responses would be provided with the relevant research paper and coding scheme, or for grading a grading key would be used)
- Step 2: Choose the LLM that will serve as the underlying model. CustomChatGPT currently provides the choice between ChatGPT 4.0 or 4o and Claude 3.0 and 3.5, but other RAG systems may use other LLMs such as Gemini, or Llama, or "home grown" models.
- Step 3: Design the persona. This is the set of instructions, written in plain English (or most any language) that tells the bot what to do and what not to do. Typically, the instructions consist of:
  (1) An overview or narrative outlining the purpose of the bot (Socratic tutor, data coder, etc.),
  (2) The personality of the bot (friendly, supportive, solemn, funny, etc.),
  (3) Explicit instructions about what to do (e.g., support the student with leading questions and prompts, or code responses according to given criteria)
  (4) Explicit instructions about what **not to do** (e.g., do not tell students the answer directly, do not answer the question if you are intended to code the response)
  (5) Other instructions e.g. output the response in a table or keep responses short and to the point. In our experience chatbots tend to be verbose![11]
- Step 4. Test the bot and iterate until the types of responses required are consistently provided.

■ EXAMPLE CHATBOT DESIGNS

Here we present two ways of using RAG AI bots: The first general use is to support instructors, and the second is to support students. These two approaches can be staged very differently, in that some of the instructor support uses can be implemented almost immediately, because they will involve only the instructor interacting with the bot. The instructor can make direct queries, for example by asking the bot to generate formative tasks as in Example 3, or can upload student responses in tabular format, as in examples 1a, and b. However, as we move into the more student focused applications, for example as an individual tutor, or to grade and provide feedback to individual students on formative or summative assessments, the AI systems should be integrated into a learning management system to ensure data privacy and to connect individual student responses with chatbot feedback and grading. This approach will be highly institution dependent, and, in our opinion as we will discuss below, will need much more research and testing before it is used for student learning "in the wild".

Bots to Code and/or Assess Student Learning

If we are to expand the use of student constructed explanations, we need some way of analyzing and assessing them. Here we provide two examples of bots that use well developed, previously validated coding schemes. Both bots are trained with the CLUE text[36] and research papers that outline the coding schemes developed by human researchers. We have found that these are the simplest bots to design, since their task is relatively straightforward (as opposed to engaging in conversation with a human). Bots that are used to code student responses do not require specialized staging: the instructor can upload the responses in table format, and the instructor can receive the codes in table format. In our experience, such coding bots appear to be very stable, providing the same codes when presented with

https://doi.org/10.1021/acs.jchemed.4c00765
J. Chem. Educ. XXXX, XXX, XXX−XXX

the same data sets, that are also in general agreement with human coding.

Example: LDF Coder. We have extensive data collected over many years for prompts that ask students to explain how London Dispersion Forces (LDFs) arise,[27,38] including a study in which we used machine learning tools (AACR) to analyze large numbers of student responses. We should note that to obtain reliable data with the AACR system it was necessary to human code over 700 responses before an acceptable Kappa value (0.7) was obtained.[38] Using our chatbot, trained with the CLUE text[36] and published papers with codebooks[27,38] − but no extra data, we found over 90% agreement between human and bot when analyzing a sample of 30 anonymous student responses. Furthermore, the disagreements were typically edge cases, that provided information about how to "tighten" the specifications for the coding scheme. The bot coding was replicable, that is it consistently arrived at the same code when recoding sets of data.

Box 1. LDF Coder Chatbot Persona and Sample Output

You are an intelligent research assistant who is helping researchers characterize student responses to the prompt "When two non-polar atoms or molecules approach each other, explain why they are attracted to each other". There are three types of response: "Non-electrostatic (NE)", "Electrostatic causal (EC)", and "Causal mechanistic (CM)". Your task is to analyze student input and assign a code (NE, EC or CM) to the response. Do not add any further commentary.

Special Instruction:

- Use a tabular format to present your analysis, with columns for each idea and rows for each student response.
- Include a summary of the frequency of each idea's occurrence across all responses.

Sample Output

STUDENT ID	INPUT 1	CODE
0100	They start out separate then the electrons of one helium atom attracts the nucleus of the other. They get closer and eventually overlap.	EC
0200	When the distance between the atoms decreases the nucleus of one atom attracts the electrons of the other atom, meaning that electrostatic forces are getting stronger leading to potential energy lowering. When the atoms get much closer the repulsion between the nuclei of the atoms get much stronger and there is an increase in PE.	EC
0300	In the boxes I am drawing a dipole happening in one atom, causing an induced dipole in another atom which will in turn cause an attraction between the two atoms and bring them closer together.	CM
0400	The polarization and attraction of a helium atom by a dipole. The close approach of the positive side of the dipole attracts the electron cloud toward it. This makes the helium atom electrically lopsided and equivalent to the dipole shown below it. There is then a net force of attraction between this induced dipole and the permanent dipole.	CM

Example: Acid−Base Coder. Just as with the LDF prompt, we have extensive longitudinal data for a prompt about acid base reactions,[28,29] "*Explain both what is happening on the molecular level when HCl reacts with* H_2O *to give* Cl^- *and* H_3O+, *and why this is happening*". The coder was trained with the CLUE text[36] and the published paper that includes the codebook for this prompt,[28] but no extra coded data. Again, we found that the code assigned by the bot was in agreement with (prior) human coding over 90% of the time, despite the fact that there were twice as many potential codes for student responses. In this version the bot also gives the rationale for its coding of responses, which provided insight into potential differences between human and bot coding. An example of coded data and explanations for differences are provided in Supporting Information S1.

Bots that are designed to assign codes to student responses based on a well-defined coding scheme obviously have a place in educational research. In our research group we are using bots to provide interrater reliability (IRR) for new coding schemes. Once the codebook is well described (by a human) and good IRR can be achieved with humans, the bot can code thousands of responses as necessary. This also clearly has implications for grading open-ended responses on exams and formative assessments. However, before we unleash this type of system on students, it behooves us to understand how students will respond (discussed further in the implications section).

Box 2. Acid−Base Coder Chatbot Persona and Sample Output

"You are an intelligent research assistant who is helping researchers characterize student responses to the prompt "Explain both what is happening on the molecular level when HCl reacts with H2O to give Cl− and H3O+, and why this is happening".

Role and Goal:

- Your primary role is to assist researchers by categorizing student responses to the prompt regarding the molecular interaction between HCl and H2O.
- Your goal is to assign one of the following codes to each response: No Response (NR), Non-Normative (NN), General Descriptive (GD), Brønsted Descriptive (BD), Brønsted Causal (BC), Lewis Mechanistic (LM), or Lewis Causal Mechanistic (LCM).

Constraints:

- Do not provide any commentary or feedback on the responses.
- Assign only one code per response, even if the response contains elements that could fit into multiple categories.
- Do not correct or critique the student responses.

Clarification:

- If a response is ambiguous or unclear, do not ask for clarification. Assign the most appropriate code based on the information provided.

Personalization:

- Maintain a neutral and objective tone when categorizing responses.
- Personalize your analysis by being attentive to the specific details provided in each student's response.

Special Instructions:

- Use a tabular format to present your analysis, with columns for the code and rows for your explanation.
- Include a summary of the frequency of each code's occurrence across all responses

Sample Output

ID	CODE	REASON
101	LCM	The response provides a detailed mechanistic explanation using Lewis theory (electron pair donation from H2O to HCl) and explains the reasons behind the process (electronegativity and partial charges leading to attraction and bond formation).
102	NN	The explanation is incorrect as it misidentifies chlorine as more electronegative than oxygen in the context of acting as an acid, which does not align with the correct understanding of the acid-base reaction described.
103	LCM	This response correctly identifies the reaction as a Lewis acid-base interaction, provides a mechanistic explanation involving electron pair donation, and explains why this happens (attraction leading to bond formation and electron pair transfer).
104	BC	The response explains the reaction using Brønsted theory (proton donation and acceptance) and provides a causal explanation for why the reaction occurs (attraction between H+ and O leading to bond formation).
105	LCM	The explanation includes a mechanistic description using Lewis theory (movement of electrons) and a causal explanation for the process (polarity and partial charges leading to attraction and bond formation).

Example: Cumulative Feedback for Instructor

While these systems have great potential for grading and coding student learning, they can also be used by instructors to get meaningful feedback about how students are thinking by using

https://doi.org/10.1021/acs.jchemed.4c00765
J. Chem. Educ. XXXX, XXX, XXX−XXX

aggregated sets of responses from open-ended tasks. On the student facing part of the bot, students might get supportive feedback (discussed in the next section), whereas on the instructor facing part, a summary of student performance can be obtained rapidly. For example, a bot has been designed to provide a summary of student responses to open-ended questions (the persona is shown in Box 3). In response to the answers to the prompt "Why do you think there is an overall thermal energy change for the process of dissolving methanol in water?", the bot produced the text shown in the Box 3 sample.

Box 3. Bot to summarize a group of student responses persona and sample output

You are a chatbot whose purpose is to analyze and evaluate in a concise manner student answers to determine whether the answers include mechanistic thinking, using and linking necessary concepts and whether these concepts are applied appropriately. The consumer of your output is the instructor.

Role and Goal: Your goal is to provide cumulative feedback to the instructor on the answers students generate, whether their answers display mechanistic reasoning, and whether they are accurately using key concepts. When analyzing student answers, provide the percentage of students who use each appropriate idea, and what missing or misused concepts are prevalent.

Constraints: Keep your responses concise, aiming for one to two paragraphs.

Guidelines: In your analysis of student answers, (i) determine the presence of key concepts and the accuracy of their application, and (ii) whether the student's answer shows mechanistic reasoning. For incorrect or incomplete answers, identify necessary but missing or misapplied concepts.

Clarification: Ask for clarification if the student's answer is ambiguous or lacks sufficient detail to analyze effectively.

Personalization: Your primary function is as an objective and helpful analyst to help the teacher understand how students are doing in this course

Sample Output

Summary of Findings: Mechanistic Reasoning: Approximately 60% of students demonstrated mechanistic reasoning by identifying both endothermic and exothermic steps in the dissolution process.

Key Concepts: Hydrogen Bonds: 50% of students mentioned hydrogen bonds, but only 30% accurately discussed their role in energy changes.

Energy Measurement: 40% of students correctly identified the need for empirical measurement to determine the net thermal energy change.

Incomplete or Incorrect Concepts: 30% of students provided incomplete or incorrect explanations, often focusing solely on bond formation or lacking detail.

Recommendations for Instruction

Emphasize the importance of considering both bond breaking and bond forming steps in energy change discussions.

Provide more detailed examples and practice problems involving the calculation of energy changes in dissolution processes.

Reinforce the need for empirical measurement to determine net energy changes, highlighting the limitations of theoretical predictions without specific data.

We have shown that this approach[39] can also be used with multiple choice questions, by asking students why each answer is right or wrong, and a similar chatbot analyzer can provide similar output and a synopsis of appropriate and inappropriate responses, and the associated student reasoning. In this way instructors could be freed from the restrictions of rigid testing systems. Students can be asked to construct their own responses and instructors can almost immediately gather a synopsis of these written efforts and adjust their instruction accordingly. While automating such a system will take time and institutional resources, it is already possible for individual instructors to upload a table of student responses (perhaps exported from a homework system) and receive the summary in seconds.

Task Developers

Example: 3D Task Developer. Certainly AI bots can generate multiple choice questions,[16] but as previously noted such questions may not align with the instructional goals of the course. Boxes 4a−4c show the persona and response for a bot

Box 4a. 3D Task Generator

Persona:

You are an instructor for the CLUE curriculum and need to design tasks that are three-dimensional (3D) to be used on homework, as in class activities, or on exams. These tasks should elicit student thinking about a chemistry core idea, a scientific practice, and a crosscutting concept within the same task. There are many examples of such questions for you to refer to.

Role and Goal:

- Role: Instructor for the CLUE curriculum.
- Goal: Design 3D tasks that elicit student thinking about a chemistry core idea, a scientific practice, and a crosscutting concept.

Constraints:

- Do not design tasks that would not elicit 3D thinking.
- Do not design tasks that are not aligned with the CLUE Curriculum.

Guidelines:

- Ensure tasks are appropriate, related, and challenging enough for the student's educational grade level.
- Tasks must directly relate to the provided materials for the CLUE curriculum, ensuring relevance and targeted learning.
- Follow provided materials strictly.

Clarification:

- Ask for clarification if the provided materials are not clear or if additional context is needed.

Personalization:

- Tailor tasks to be engaging and thought-provoking, encouraging deep understanding and application of chemistry concepts.

Special Instructions:

- Always follow the provided materials for the CLUE curriculum.
- Ensure tasks integrate a chemistry core idea, a scientific practice, and a crosscutting concept.

https://doi.org/10.1021/acs.jchemed.4c00765
J. Chem. Educ. XXXX, XXX, XXX−XXX

Box 4b. Impact of Hydrogen Bonding on Lake Wildlife Task Example

Input request: "Please generate a 3D task that involves students thinking about the impact of hydrogen bonding on the wildlife in a body of lake water"

Output: Certainly! Here is a 3D task designed to elicit student thinking about the impact of hydrogen bonding on the wildlife in a body of lake water, integrating a chemistry core idea, a scientific practice, and a crosscutting concept:

Task: Impact of Hydrogen Bonding on Lake Wildlife

Core Idea: Electrostatic Bonding and Interactions

- **Concept**: Hydrogen bonding interactions between water molecules and their impact on the physical properties of water.

Scientific Practice: Developing and Using Models

- **Activity**: Constructing and interpreting models to explain the phenomenon.

Crosscutting Concept: Cause and Effect: Mechanism and Explanation

- **Focus**: Understanding how hydrogen bonding affects the temperature stability of lake water and its subsequent impact on wildlife.

Task Description:

1. **Phenomenon**:
   - During winter, the surface of a lake freezes, but the water beneath remains liquid, providing a habitat for aquatic life.
2. **Model Representation**:
   - Provide students with a molecular-level model showing hydrogen bonding interactions between water molecules in both liquid and solid states.
3. **Questions**:
   - **Identify and Explain**:
     - Identify the types of attractive forces shown in the model between water molecules.
     - Explain how hydrogen bonding interactions contribute to the unique properties of water, such as its high specific heat capacity and the fact that ice floats on liquid water.
   - **Prediction and Reasoning**:
     - Predict how the presence of hydrogen bonds affects the temperature stability of the lake water during seasonal changes.
     - Provide reasoning for how this temperature stability impacts the survival of aquatic wildlife during winter.
   - **Cause and Effect Analysis**
     - Describe the cause (hydrogen bonding interactions) and the effect (temperature stability of lake water).
     - Explain the mechanism linking the cause and effect, focusing on how the energy required to break hydrogen bonds influences the thermal properties of water and the habitat conditions for wildlife.

designed to develop 3DL tasks, which typically are more difficult to construct. However, if we expect students to become proficient at such tasks, they need plenty of practice.

Box 4c. Evidence Statements for Impact of Hydrogen Bonding on Lake Wildlife Task

1. **Attractive Forces**:
   - The attractive forces within water molecules are covalent bonds, while the forces between water molecules are hydrogen bonds.
2. **Hydrogen Bonding Impact**:
   - Hydrogen bonds are weaker than covalent bonds but significantly influence the physical properties of water, such as its high specific heat capacity and lower density in the solid state.
3. **Temperature Stability**:
   - The high specific heat capacity of water due to hydrogen bonding helps maintain a stable temperature in the lake, providing a consistent environment for aquatic life.
4. **Winter Survival**:
   - The formation of ice on the surface, which is less dense than liquid water, insulates the water below, allowing aquatic organisms to survive in liquid water even in freezing temperatures.

This task integrates the core idea of electrostatic bonding and interactions, the scientific practice of developing and using models, and the crosscutting concept of cause and effect, providing a comprehensive 3D learning experience.

3. CHATBOTS AS TUTORS

There are a number of examples of tutors being developed by publishers and online providers; but potential drawbacks include 1) the problem of hallucination, 2) they may draw from inappropriate materials that are either too simple or too advanced for the student or 3) may simply provide incorrect answers. Additionally, many of these bots seem to focus on very traditional tasks, that do not involve scientific practices or the use of knowledge.[13] RAG chatbots offer a way to use GenAI tools to help students construct, connect, and apply their knowledge, so as to support student' reasoning in sophisticated ways about phenomena, rather than concentrating on facts, skills and algorithmic problem solving.

The easy-to-use RAG GenAI tools mean that instructors can design custom bots to support the approaches to learning that they value. For example, the usefulness of a bot that merely provides the correct response as feedback for a student is debatable, since it tends to offload the intellectual work onto a bot and removes the effort that learning requires. There is always a danger of generating the "illusion of understanding"[40] rather than the real thing. It is also worth noting that rather than using feedback to reflect upon and revise their answers most feedback appears to go unread and unused by students.[41,42] So although feedback is often described as an important and integral component to good teaching,[43] the way that feedback is given and used, and the purpose that the feedback serves can have profound effects on learning.[44] Feedback that engages students in reflection, and that requires a response from the student has been shown to be far more effective than simply writing comments on student work.[41]

To avoid circumventing the learning process, a better approach is to design Socratic tutors that have the capability to support students with questions and prompts designed to engage the student in a conversation on the topic and so to promote student self-reflection and understanding, to help them

https://doi.org/10.1021/acs.jchemed.4c00765
J. Chem. Educ. XXXX, XXX, XXX−XXX

make connections among the ideas they are using and ultimately support sensemaking about phenomena involving mechanistic reasoning. The persona for such a bot and a sample response are shown in Box 5a and 5b. This type of bot provides encouraging

Box 5a. General Socratic Tutor Persona

You are a tutor for general chemistry students. You use the Socratic method to help students with the goal of encouraging critical and mechanistic thinking and self-confidence based on accurate understanding of key facts and concepts. When a student answers a question correctly, you congratulate them. When they appear mistaken, you make one or two suggestions that they can consider and perhaps a follow-up question.

Constraints: While you do not provide direct answers to questions, you can suggest readings and other resources that students can consult to learn more about misunderstood concepts or topics. You maintain a consistently friendly, informal, engaging, and encouraging tone.

Guidelines: Engage users with questions that prompt them to think deeply about the subject matter. Confirm when users are on the right track and offer ideas or facts to aid their reasoning process.

Clarification: You may ask for clarification to ensure you understand the user's answer or a specific question but always steer the conversation back to helping them find the answer themselves.

Personalization: Your responses should be tailored to the user's inquiries providing a personalized and thoughtful Socratic dialogue.

Special Instructions: Keep responses short, one to two paragraphs) to maintain clarity and focus in the conversation.

and reflective feedback of the type that has been shown to be most effective.[43−45] The bot is endlessly patient, available at all hours, and can respond in whatever language the student begins the interaction, thus perhaps providing support for otherwise marginalized students.

We note here that bots such as this, that interact with humans, tend to give somewhat different responses each time a query is submitted. This is a result of the nature of the LLM on which the bot is based. LLMs are probabilistic models, and although the content of the RAG chatbot is supplied by the designer, the ways in which it responds each time tends to be different. We have provided a sample of the different types of responses that emerged for the same prompt in the Supplemental S2. It is this type of bot that will need extensive student testing before it is released for general use in complex systems.

Other Chatbots

There are numerous types of chatbots in addition to the examples shown here, and an excellent resource has been provided by Mollick et.al.;[46] the examples provided may require changes to be used in a chemistry education context. For example, it is possible to engineer a persona that reverses the tutorial role, that is the chatbot functions as a student and the student "teaches" the bot about some topic. Such preprogrammed teachable agents[47] have already been shown to result in increased learning, and there exciting possibilities for the future in this area. We also note that there is no discussion here of other modes of input (for example drawing or speaking). These types of systems are rapidly emerging, but as yet we have no experience with them.

Box 5b. Sample Response

Student: why does increasing the temperature increase the rate of reaction?

CLUEbot: That is a great question! Let us think about what happens when we increase the temperature. Can you recall how temperature affects the kinetic energy of molecules? And how might this increased kinetic energy influence the frequency and energy of collisions between reactant molecules?

Student: temperature and KE are directly related, So if i raise the temp the energy of the molecules should increase. So collisions between the molecules woudl have more energy?

CLUEbot: Yes, you are on the right track! When the temperature increases, the kinetic energy of the molecules also increases. This means the molecules move faster and collide more frequently. Now, think about what happens when these high-energy molecules collide. What might this mean for the energy barrier that needs to be overcome for a reaction to occur?

Student: Oh OK so more energetic collisions means more likely to have enough energy to get over the barrier?

CLUEbot: Yes, you have got it! When the temperature increases, the average kinetic energy of the molecules increases. This means there are more molecules with higher kinetic energy. These high-energy molecules are more likely to have enough energy to overcome the activation energy barrier when they collide. This increases the probability of a reaction occurring. Now, can you think of how this might affect the rate of a chemical reaction?

Student: Well it means that more molecules will react - so the rate increases?

CLUEbot: Exactly! You have got it. As the temperature rises, the probability of productive collisions between particles per unit time increases, thus increasing the reaction rate. This is because more molecules have the necessary energy to overcome the activation energy barrier. So, what do you think might happen if we were to decrease the temperature?

Student: the rate would decrease?

CLUEbot: That is correct! If the temperature decreases, the kinetic energy of the molecules also decreases. This means the molecules move slower and collide less frequently and with less energy. Therefore, fewer molecules have the necessary energy to overcome the activation energy barrier, which decreases the rate of the reaction. Well done!

■ SUMMARY

Artificial Intelligence in the form of LLM-based chatbots is not going away. We do not know what will happen in the future (or even the next six months), but we believe that there is enough information for us to begin to plan what instruction and assessment could, and perhaps should look like in the future. In this paper we have discussed the use of RAG chatbot systems which can 1) alleviate the problems with hallucinations and incorrect responses, and 2) open the possibilities of GenAI systems for those of us who lack the technical know-how to harness and train GenAI models for specific purposes. They allow us to use our expertise in chemistry, chemistry teaching and student learning, to design and implement GenAI tutors, coders and feedback systems without the high activation energy barrier often associated with new technologies. Although it is possible that many of the examples provided here could be accomplished with trained GenAI systems such as ChatGPT 4o,

https://doi.org/10.1021/acs.jchemed.4c00765
J. Chem. Educ. XXXX, XXX, XXX−XXX

the nuances and technical specifications of the responses may not align well with the instructional goals for the course, because such bots are not specifically constructed from materials that reflect those goals. This is especially true for learning approaches that emphasize use of knowledge, such as 3DL[48] or POGIL.[49] Additionally, queries may well result in material returned at an inappropriate level (too advanced or too simple) for students in a course or it may introduce unproductive ideas. In our case we are particularly interested in supporting mechanistic thinking in students.[50,51] Bots that are not explicitly trained and constrained to provide such support may well omit the need of such reasoning.

■ IMPLICATIONS

Implications for Instructors

As noted earlier, the use of these chatbots can be divided into two main categories: those that involve instructor-bot interactions, and those that involve student-bot interactions. We recommend that instructors who are interested in designing and using such bots begin with the instructor-bot systems. This approach can be begun immediately and will help users develop an understanding of the capabilities (and possible limitations) of such systems to support more complex learning activities, (for example by uploading aggregate responses to see what kinds of reasoning students are using to respond to a prompt). This should provide instructors with information to use in conversations with institutions about how AI should be implemented on their campus. If instructor designed RAG chatbots are integrated into campus systems we will soon be able to accurately and reliably score both formative and summative assessments, which has enormous implications for the ways that we assess students. Currently (particularly in large enrolment courses) multiple choice is the only pragmatic approach to administering summative assessments. For example, for the two or three open-ended questions on our general chemistry exams, where students must construct models, arguments or explanations, grading requires a whole day and about 40 people.

In our teaching and research we also use the formative assessment homework system beSocratic to have students construct open-ended responses to 3D questions,[52] but are currently unable to give individual feedback or support to students. We have a working prototype of the beSocratic system that will allow students to call up a Socratic Chatbot to provide help with complex responses. In the near future, we can ask students to construct models, arguments and explanations, analyze and interpret data, and communicate their responses and see immediate feedback for both instructor and students. These approaches to assessments mean that our approaches to instruction will also change. No longer will we need to focus on fragmentary ideas and isolated skills. We can design and implement assessments that require students to integrate their knowledge and use it in new situations.

Ideally, as we move forward, our communities should develop open education repositories of chatbot designs that instructors (and researchers) can use to support the types of teaching and learning that each community values.

Implications for Researchers

We are already using such chatbots to code large numbers of student responses in several different research projects. In our work focused on the design and implementation of formative assessments, the role of the researcher shifts from tedious coding of numerous responses to richer more intellectually stimulating work, such as task design, code scheme development, and further elucidation of "edge" cases. The ability to obtain rapid reliable coding of large numbers of open-ended responses will mean that the researcher can implement design-based changes more quickly and see the impact of their work.

■ A CAVEAT

It seems clear that GenAI systems will take their place in both research and teaching in education, yet we know almost nothing about what their impact will be. A recent survey[53] indicated that a minority of institutions have offered training in how to use GenAI, yet most faculty are being urged to incorporate them in their teaching. Microsoft's "co-pilot" is promoted as a "valuable tool··· to save time creating rubrics, personalized content for students, and educational materials such as quizzes and lesson plans" despite the fact that there is no evidence that such affordances will improve learning outcomes. Indeed, it is unlikely that such copilots will promote the kinds of deeper, transferable learning we are suggesting, and by experimenting now with RAG bots and other systems, faculty may be able to provide more informed feedback as these decisions are made.

Although the excitement and interest around the uses of generative AI are genuine, in fact we do not have a great deal of information about just how students will interact with chatbots. It is important to introduce such systems judiciously, and study them thoroughly. Unfortunately, it is too late for the first GenAI systems that have been rushed out and are already in wide use (at least by students). We need to exercise care and thought about just what is it we want our students to know and to be able to do in the future, and then intentionally design such systems. On the other hand, we do not want to end up with a situation where we use AI systems to generate complex meaningful tasks, students then use those systems to supply answers, which are then auto graded, bypassing learning completely. There is a real and urgent need for research on a systemic approach to the implementation of GenAI, while at the same time we are assembling the aircraft as we fly it.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.4c00765.

Supporting Information S1: Examples of Human and BOT coding of student data (PDF, DOCX)

Supporting Information S2: Examples of variation in output when CLUEbot is asked the same question repeatedly (PDF, DOCX)

■ AUTHOR INFORMATION

Corresponding Author

Melanie M. Cooper − *Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-7050-8649; Email: mmc@msu.edu

Author

Michael W. Klymkowsky − *Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, Boulder, Colorado 80309, United States;* orcid.org/0000-0001-5816-9771

Complete contact information is available at:

https://pubs.acs.org/10.1021/acs.jchemed.4c00765

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to acknowledge Nate Kocho from CustomGPT for help and productive conversations.

■ REFERENCES

(1) How Professors Scrambled to Deal With ChatGPT. *The Chronicle of Higher Education*. https://www.chronicle.com/article/caught-off-guard-by-ai (accessed 2023−12−14).

(2) Emenike, M. E.; Emenike, B. U. Was This Title Generated by ChatGPT? Considerations for Artificial Intelligence Text-Generation Software Programs for Chemists and Chemistry Educators. *J. Chem. Educ.* **2023**, *100* (4), 1413−1418.

(3) Exintaris, B.; Karunaratne, N.; Yuriev, E. Metacognition and Critical Thinking: Using ChatGPT-Generated Responses as Prompts for Critique in a Problem-Solving Workshop (SMARTCHEMPer). *J. Chem. Educ.* **2023**, *100* (8), 2972−2980.

(4) Tyson, J. Shortcomings of ChatGPT. *J. Chem. Educ.* **2023**, *100* (8), 3098−3101.

(5) ChatGPT Has Everyone Freaking Out About Cheating. It's Not the First Time. *The Chronicle of Higher Education*. https://www.chronicle.com/article/chatgpt-has-everyone-freaking-out-about-cheating-its-not-the-first-time (accessed 2024−06−04).

(6) Tassoti, S. Assessment of Students Use of Generative Artificial Intelligence: Prompting Strategies and Prompt Engineering in Chemistry Education. *J. Chem. Educ.* **2024**, *101* (6), 2475−2482.

(7) Guo, Y.; Lee, D. Leveraging ChatGPT for Enhancing Critical Thinking Skills. *J. Chem. Educ.* **2023**, *100* (12), 4876−4883.

(8) Young, J. D.; Dawood, L.; Lewis, S. E. Chemistry Students' Artificial Intelligence Literacy through Their Critical Reflections of Chatbot Responses. *J. Chem. Educ.* **2024**, *101* (6), 2466−2474.

(9) ChatGPT - Lab Report Evaluator. *ChatGPT*. https://chatgpt.com/g/g-gJdAovpxL-lab-report-evaluator (accessed 2024−06−04).

(10) What is RAG? - Retrieval-Augmented Generation Explained - AWS. *Amazon Web Services, Inc.* https://aws.amazon.com/what-is/retrieval-augmented-generation/ (accessed 2024−06−04).

(11) Hicks, M. T.; Humphries, J.; Slater, J. ChatGPT Is Bullshit. *Ethics Inf Technol.* **2024**, *26* (2), 38.

(12) *CustomGPT.ai | Custom GPT With ALL Your Business Content.* https://customgpt.ai/ (accessed 2024−05−07).

(13) *Meet Khanmigo: Khan Academy's AI-powered teaching assistant & tutor*. https://khanmigo.ai/ (accessed 2024−06−04).

(14) *MagicSchool.ai - AI for teachers - lesson planning and more!* https://www.magicschool.ai/ (accessed 2024−06−05).

(15) *Explore the AI study tool in MyLab and Mastering*. https://www.pearson.com/en-us/higher-education/products-services/aistudytool.html (accessed 2024−06−07).

(16) Top Hat Ace. *Top Hat*. https://tophat.com/features/ace-ai/ (accessed 2024−06−11).

(17) *NotebookLM | Note Taking & Research Assistant Powered by AI*. https://notebooklm.google/ (accessed 2024−08−16).

(18) Cooper, M. M.; Grove, N.; Underwood, S. M.; Klymkowsky, M. W. Lost in Lewis Structures: An Investigation of Student Difficulties in Developing Representational Competence. *J. Chem. Educ.* **2010**, *87* (8), 869−874.

(19) Cooper, M. M.; Underwood, S. M.; Hilley, C. Z. Development and Validation of the Implicit Information from Lewis Structures Instrument (IILSI): Do Students Connect Structures with Properties? *Chem. Educ. Res. Pract.* **2012**, *13* (3), 195−200.

(20) Cooper, M. M. Why Ask Why? *J. Chem. Educ.* **2015**, *92*, 1273−1279.

(21) Liles, J. Did Mark Zuckerberg Say, "Move Fast And Break Things"? *Snopes*. https://www.snopes.com//fact-check/move-fast-break-things-facebook-motto/ (accessed 2024−06−18).

(22) National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; National Academies Press: Washington, DC, 2012.

(23) Cooper, M. M.; Klymkowsky, M. W. *CLUE: Chemistry, Life, the Universe & Everything*. http://clue.chemistry.msu.edu/ (accessed 2017−07−18).

(24) Cooper, M. M.; Stowe, R. L.; Crandell, O. M.; Klymkowsky, M. W. Organic Chemistry, Life, the Universe and Everything (OCLUE): A Transformed Organic Chemistry Curriculum. *J. Chem. Educ.* **2019**, *96* (9), 1858−1872.

(25) Ralph, V. R.; Scharlott, L. J.; Schafer, A. G. L.; Deshaye, M. Y.; Becker, N. M.; Stowe, R. L. Advancing Equity in STEM: The Impact Assessment Design Has on Who Succeeds in Undergraduate Introductory Chemistry. *JACS Au* **2022**, *2*, 1869.

(26) Underwood, S. M.; Reyes-Gastelum, D.; Cooper, M. M. When Do Students Recognize Relationships between Molecular Structure and Properties? A Longitudinal Comparison of the Impact of Traditional and Transformed Curricula. *Chemistry Education Research and Practice* **2016**, *17*, 365−380.

(27) Noyes, K.; Cooper, M. M. Investigating Student Understanding of London Dispersion Forces: A Longitudinal Study. *J. Chem. Educ.* **2019**, *96* (9), 1821−1832.

(28) Cooper, M. M.; Kouyoumdjian, H.; Underwood, S. M. Investigating Students' Reasoning about Acid−Base Reactions. *J. Chem. Educ.* **2016**, *93* (10), 1703−1712.

(29) Crandell, O. M.; Kouyoumdjian, H.; Underwood, S. M.; Cooper, M. M. Reasoning about Reactions in Organic Chemistry: Starting It in General Chemistry. *J. Chem. Educ.* **2019**, *96* (2), 213−226.

(30) Crandell, O. M.; Lockhart, M. A.; Cooper, M. M. Arrows on the Page Are Not a Good Gauge: Evidence for the Importance of Causal Mechanistic Explanations about Nucleophilic Substitution in Organic Chemistry. *J. Chem. Educ.* **2020**, *97* (2), 313−327.

(31) Williams, L. C.; Underwood, S. M.; Klymkowsky, M. W.; Cooper, M. M. Are Noncovalent Interactions an Achilles Heel in Chemistry Education? A Comparison of Instructional Approaches. *J. Chem. Educ.* **2015**, *92*, 1979−1987.

(32) Houchlei, S. K.; Bloch, R. R.; Cooper, M. M. Mechanisms, Models, and Explanations: Analyzing the Mechanistic Paths Students Take to Reach a Product for Familiar and Unfamiliar Organic Reactions. *J. Chem. Educ.* **2021**, *98*, 2751.

(33) Pashler, H.; Bain, P. M.; Bottge, B. A.; Graesser, A.; Koedinger, K.; McDaniel, M.; Metcalfe, J. Organizing Instruction and Study to Improve Student Learning. *IES Practice Guide. NCER 2007−2004*; National Center for Education Research, 2007.

(34) Delannoy, P.-A. *Dubmix/RAG-Chatbot*, 2024. https://github.com/dubmix/RAG-chatbot (accessed 2024−06−12).

(35) RAG Evaluation Series: Validating the RAG performance of OpenAI vs CustomGPT.ai. *Tonic.ai.* https://www.tonic.ai/blog/rag-evaluation-series-validating-the-rag-performance-of-openai-vs-customgpt-ai (accessed 2024−05−05).

(36) Cooper, M. M.; Klymkowsky, M. W. *CLUE: Chemistry, Life, the Universe and Everything*; Michigan State University Libraries, 2020.

(37) Cooper, M. M.; Klymkowsky, M. W. *OCLUE: Organic Chemistry, Life, the Universe & Everything*; Michigan State University Libraries, 2020.

(38) Noyes, K.; McKay, R. L.; Neumann, M.; Haudek, K. C.; Cooper, M. M. Developing Computer Resources to Automate Analysis of Students' Explanations of London Dispersion Forces. *J. Chem. Educ.* **2020**, *97* (11), 3923−3936.

(39) Klymkowsky, M.; Cooper, M. M. The End of Multiple Choice Tests: Using AI to Enhance Assessment. *arXiv* **2024**, DOI: 10.48550/arXiv.2406.07481.

(40) Messeri, L.; Crockett, M. J. Artificial Intelligence and Illusions of Understanding in Scientific Research. *Nature* **2024**, *627* (8002), 49−58.

(41) Jonsson, A. Facilitating Productive Use of Feedback in Higher Education. *Active Learning in Higher Education* **2013**, *14* (1), 63−76.

(42) Winstone, N. E.; Nash, R. A.; Rowntree, J.; Parker, M. 'It'd Be Useful, but I Wouldn't Use It': Barriers to University Students'

Feedback Seeking and Recipience. *Studies in Higher Education* **2017**, *42* (11), 2026−2041.
(43) Hattie, J.; Timperley, H. The Power of Feedback. *Review of Educational Research* **2007**, *77*, 81−112.
(44) Adcroft, A. The Mythology of Feedback. *Higher Education Research & Development* **2011**, *30* (4), 405−419.
(45) Chang, D. H.; Lin, M. P.-C.; Hajian, S.; Wang, Q. Q. Educational Design Principles of Using AI Chatbot That Supports Self-Regulated Learning in Education: Goal Setting, Feedback, and Personalization. *Sustainability* **2023**, *15* (17), 12921.
(46) Mollick, E. R.; Mollick, L. Instructors as Innovators: A Future-Focused Approach to New AI Learning Opportunities, With Prompts. *The Wharton School Research Paper* **2024**, DOI: 10.2139/ssrn.4802463.
(47) Chase, C. C.; Chin, D. B.; Oppezzo, M. A.; Schwartz, D. L. Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. *J. Sci. Educ Technol.* **2009**, *18* (4), 334−352.
(48) Cooper, M. M.; Caballero, M. D.; Ebert-May, D.; Fata-Hartley, C. L.; Jardeleza, S. E.; Krajcik, J. S.; Laverty, J. T.; Matz, R. L.; Posey, L. A.; Underwood, S. M. Challenge Faculty to Transform STEM Learning. *Science* **2015**, *350* (6258), 281−282.
(49) *Process Oriented Guided Inquiry Learning (POGIL)*; Moog, R. S., Spencer, J. N.,, Eds.; American Chemical Society: Washington, DC, 2008.
(50) Noyes, K.; Carlson, C. G.; Stoltzfus, J. R.; Schwarz, C. V.; Long, T. M.; Cooper, M. M. A Deep Look into Designing a Task and Coding Scheme through the Lens of Causal Mechanistic Reasoning. *J. Chem. Educ.* **2022**, *99* (2), 874−885.
(51) Klymkowsky, M. W. Making Mechanistic Sense: Are We Teaching Students What They Need to Know? *Dev. Biol.* **2021**, *476*, 308−313.
(52) *beSocratic*. https://besocratic.com/home (accessed 2024−06−12).
(53) Bferro. Time for Class 2024. *Tyton Partners*, 2024. https://tytonpartners.com/time-for-class-2024/ (accessed 2024−06−20).

https://doi.org/10.1021/acs.jchemed.4c00765
J. Chem. Educ. XXXX, XXX, XXX−XXX