JOURNAL OF CHEMICAL EDUCATION

Open Access

This article is licensed under CC-BY 4.0

pubs.acs.org/jchemeduc

Article

Development and Implementation of an AI Tutor To Scaffold Problem-Solving in Introductory Chemistry

Karen D. Wang, Joshua T. Arens,* Chinat Yu, Jordyn N. Smith, Hannah Bartels, Mercy Haub, Fun Man Fung, Nick Haber, Shima Salehi, and Jennifer Schwartz Poehlmann

Cite This: *J. Chem. Educ.* 2026, 103, 5555−5563

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Developing problem-solving competency is a central goal of undergraduate chemistry education, yet many students struggle to engage productively with complex, multistep problems. At the same time, the rapid adoption of large language model (LLM)−based tools by students in STEM courses raises concerns about over-reliance and uncritical use, potentially undermining learning. In this work, we present the design, implementation, and evaluation of STEPS (STEM Tutor for Effective Problem Solving), an AI tutor particularly designed to guide students in creating a problem-solving plan without divulging solutions. The tutor was deployed in an introductory chemistry course at an R1 university, and its impact was examined using performance data, Likert-scale survey responses, and qualitative analysis of student reflections. Students demonstrated high performance on STEPS-assisted problems and reported strong confidence in their solutions. Survey results indicated consistently high perceptions of helpfulness. Qualitative analysis revealed that students valued the tutor's guidance in organizing their thinking and checking reasoning without giving away solutions. These findings suggest that carefully designed, task-specific AI tutors can promote productive engagement with problem solving while mitigating risks associated with uncritical AI use in chemistry education.



KEYWORDS: *First-Year Undergraduate/General, Internet/Web-Based Learning, Problem Solving/Decision Making, Applications of Artificial Intelligence*

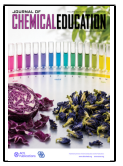
INTRODUCTION

Developing problem-solving competency is a central goal of chemistry education.[1] The significance of this educational goal is increasing given the rapid integration of artificial intelligence (AI) into professional and everyday life, particularly as AI systems demonstrate proficiency in handling well-structured cognitive tasks.[2] Consequently, the ability to solve complex, unstructured problems in effective collaboration with AI tools has become an essential competency for STEM graduates.

To date, educators have explored various instructional strategies to teach problem-solving in college chemistry, including explicit metacognitive instruction for first-year college students[3] and a systematic workflow that scaffolds students through specific phases of problem-solving in the context of an introductory chemistry course.[4] However, chemistry education still faces major challenges in helping students develop complex problem-solving competencies. Introductory chemistry courses often rely on well-structured, textbook-style problems that strip away the complexity and ambiguity inherent in real-world scenarios. Moreover, students are typically assessed on the correctness of their final answers rather than the quality of their problem-solving processes.[5] These oversimplifications stem from practical constraints: given large class sizes and fast-paced, content-heavy curricula, instructors face significant challenges in developing authentic complex problem scenarios and providing individualized scaffolding.[6] Consequently, students may develop the mindset that reaching the correct answer is more important than understanding the problem-solving process,[7] which ultimately limits their development of problem-solving competency.

It is worthwhile to examine whether and how large language models (LLMs) can help address these instructional challenges. They can help instructors create real-world, complex problems by generating case-study topics, problem statements, and mock experimental data.[8,9] During problem solving, LLMs can provide adaptive scaffolding that guides students through the problem-solving process, thereby shifting student behavior from solution-seeking to process-oriented approaches.[10,11] Furthermore, LLM-based tutoring systems provide students opportunities to develop productive human−AI collaboration,[12,13] an emerging essential competency for chemistry graduates. Realizing these benefits, however, depends on careful instructional design to specify when and how LLMs are used in chemistry education, prioritizing transparency and metacognition throughout the course, and

Received: January 10, 2026
Revised: August 22, 2026
Accepted: August 27, 2026
Published: September 23, 2026



ACS Publications

© 2026 The Authors. Published by American Chemical Society and Division of Chemical Education, Inc.

https://doi.org/10.1021/acs.jchemed.6c00039
J. Chem. Educ. 2026, 103, 5555−5563

aligning tasks and assessments with authentic complex problem solving.[14] Effective LLM integration into instruction therefore requires more than tool availability and depends on how the technology is situated within the course environment, including instructor modeling of use, clear communication of expectations for AI use, and alignment with instructional goals.[15]

Recent research indicates that college students are rapidly adopting generative AI tools such as ChatGPT as part of their academic workflows.[16] Many students report using these tools in pedagogically productive ways, including to understand complex topics and clarify concepts.[17] At the same time, emerging evidence suggests that students often engage with AI tools in ways that prioritize efficiency and task completion, sometimes relying on AI to directly generate solutions to problems without much critical engagement.[18,19] Such use can further exacerbate the disconnect between traditional chemistry coursework and complex problem-solving practices. Although LLMs such as GPT-5 and Claude-3.5 Sonnet have performed well on high-school- and college-level chemistry benchmarks consisting of well-structured textbook style questions,[20,21] they still generate incorrect solutions in a substantial portion of cases, particularly with multistep, multiconcept problems, where students struggle the most. More concerning, LLMs present incorrect solutions with the same level of confidence as correct ones, creating significant risk of misleading students and reinforcing misconceptions unless outputs are critically evaluated.[22] Students vary widely in their ability to recognize and respond to these inaccuracies,[23] highlighting the need for scaffolding to support AI literacy in chemistry problem solving.[24]

The approach of using AI for direct solutions may also risk incurring significant long-term cognitive costs: neurological evidence from a recent EEG study shows that LLM users exhibit different brain connectivity patterns and reduced cognitive engagement, compared to students working independently.[25] The limitations of existing AI tools in education become even more problematic when considered alongside extensive research demonstrating that students' productive struggles deepen their learning.[26] In order to build effective problem-solving competency, students need opportunities to practice how to get started on complex problems, make mistakes and reflect on them, and develop their own plan that can ultimately solve the problem. Furthermore, they need to practice how to mindfully and critically leverage AI tools as a learning companion, rather than a replacement, in this process. Over-reliance on AI tools to generate immediate solutions deprives students of these critical productive struggles, ultimately undermining their development of the strategies and metacognitive skills essential for complex problem solving in the real world.

There is a clear need for AI tools designed to guide students through problem solving rather than providing solutions and to support instructors in expanding the opportunities for students to engage with complex, real-world problems in introductory STEM courses. To address this gap, we developed an LLM-powered tutor named STEPS, or STEM Tutor for Effective Problem Solving. This paper describes the design, implementation, and initial evaluation of STEPS in an introductory chemistry course and addresses the following questions:

- How can an LLM-based tutor be designed through prompt engineering to scaffold chemistry problem-solving processes?
- How can this tutor be effectively integrated into an introductory chemistry course?
- What are the educational impacts of the tutor, including student perceptions of its utility, problem-solving performance, and development of AI literacy?

BACKGROUND

Problem Solving and Expert Decision Making

We have conducted extensive research to characterize problem-solving expertise in STEM domains.[27,28] Our framework breaks down the problem-solving process of experts into key phases including planning, execution, and reflection, each with a set of specific decisions that must be considered and deliberately made. In the planning phase, critical decisions include determining (1) the relevant domain knowledge for solving the problem, (2) the necessary assumptions and simplifications to make the problem more tractable, (3) the information/data needed for solving the problem, and (4) methods to collect the information. Based on this framework, we designed paper-and-pencil problem-solving templates for students in college introductory physics and chemistry courses.[29,30] We found that structured scaffolding through the template could improve student problem-solving performance when problems were appropriately complex. However, even with introductory sessions introducing the template to students, it was difficult to motivate students to work through this long, paper-based activity on their own. At the same time, our analysis of students' problem-solving processes revealed that students struggle most with selecting and adapting appropriate domain knowledge and formulating detailed solution plans.[29] These findings highlight the need for more targeted, adaptive support in the planning phase of chemistry problem solving, one that scaffolds both content-specific and metacognitive processes.[31] We therefore explored how an LLM-based tutor could scaffold the planning phase and keep students engaged with it.

METHODS

Technology Description

Existing literature on AI tutors has largely focused on general-purpose tutors with the aim of providing readily accessible resources for students to explore any topic across various disciplines.[32,33] However, this broad approach imposes an unrealistic burden on students to verify the accuracy and appropriateness of AI-generated content, especially as current AI models remain prone to generating inaccurate information. Given that students are novices in the subjects they seek to learn, this may lead to misconceptions and reinforce misunderstanding among students. Moreover, students often lack the mental models necessary to formulate productive questions when engaging with AI tutors to support their problem-solving. Without clear guidance on how to interact effectively with these tools, students may default to asking for complete solutions rather than seeking targeted support for specific aspects of their problem-solving process.[18] Researchers have explored approaches such as teaching students to prompt AI as a tutor rather than a solution provider[34] and providing AI literacy workshops covering how LLMs work, prompting, and response evaluation to support productive AI use.[35] Custom-designed AI tools represent another approach to addressing this challenge, as we demonstrate with STEPS.

https://doi.org/10.1021/acs.jchemed.6c00039
J. Chem. Educ. 2026, 103, 5555−5563

In contrast to general-purpose AI tutors, the scope of STEPS is defined to scaffold problem solving in college STEM courses by guiding students through our validated problem-solving practices framework, with a particular focus on the planning phase. We have designed STEPS to function as a thought companion, mirroring how we envision LLMs will be used in future professional careers. At the same time, because its users are students who are novices in the subject material, STEPS also serves as a domain-specific tutor to answer student questions about content. Each instance of STEPS is designed for a specific problem situated in a real-world context and requires the application and integration of multiple concepts, formulas, and skills learned in class. This targeted design ensures greater alignment with specific learning goals while reducing the likelihood of LLMs generating inaccurate or pedagogically ineffective content. By adapting LLMs to ask questions corresponding to key effective problem-solving decisions and guide students through the planning process step-by-step, we created STEPS to be a tool that instructors can integrate into student learning experiences, ensuring it complements course instruction and supports student learning in a timely, targeted, and pedagogically-informed manner.

Prompt Engineering

Our prompt-engineering process began by adapting a prompt that we had developed for a physics problem to a chemistry problem based on our validated problem-solving template.[30] The initial chemistry prompt followed the physics prompt in its overall structure, retained the same defined roles and purpose, behavioral guidelines, and adapted the sequence of questions the LLM should pose to students. The prompt defines STEPS as a chemistry problem-solving coach that guides students through a structured question sequence: identifying relevant concepts, explaining their relevance, specifying needed information and equations, and creating a step-by-step plan. It also incorporates behavioral guidelines that prevent direct solution provision, maintain engagement, and redirect students back to the planning process when they seek complete answers or stray off-task. For each implementation in the chemistry course, we first selected or developed problems that required knowledge and skills covered in that week's coursework. We adjusted the question sequences to target specific chemistry concepts, specified the correct answers for each question, and updated both questions and answers within the STEPS prompt. Before deployment, our teaching team, which included course instructors and undergraduate and graduate students, conducted extensive testing by adopting different student personas, including answer-seeking students, students providing minimal responses, and engaged students offering detailed explanations, to evaluate the tutor's behavior across varied interaction styles. Based on these tests, we refined the behavioral guidelines and response patterns to handle different student approaches effectively. STEPS operated solely based on the prompt. No additional course materials were provided to or accessed by the model. Chart 1 presents a summary of the final prompt structure. For the full prompt, see Listing S1 in the Supporting Information.

Chart 1. Summary of Prompt for the AI Tutor

Role and Purpose:

- You are a specialized chemistry problem-solving coach. Your purpose is to guide students in developing a plan for solving the problem without solving the problem for them.

Core Objectives:

- Guide students to build their own structured problem-solving plan.
- Develop students' metacognitive skills by promoting reflection on problem-solving approaches.
- Support students' development of both chemistry knowledge and problem-solving skills.
- Maintain student engagement through positive reinforcement and strategic questioning.

Behavior Guidelines:

- Never solve the problem for the student.
- Ask questions one at a time and wait for student responses.
- Keep responses short and succinct ($\leq$ 3 sentences at a time).
- Be encouraging and motivational, but do not praise incorrect answers.
- If a student gives an incorrect or incomplete answer, offer increasingly specific, targeted hints rather than providing the correct answer.

Problem: <problem statement>
Structured Guidance Sequence and Correct Answers:

1. What are relevant chemistry concepts? <correct answer>
2. Why are these concepts relevant to this type of problem? <correct answer>
3. What details (equations, constants, conversions, definitions, etc.) might you need and how will you use them? <correct answer>
4. What is your step-by-step plan for solving the problem? <correct answer>

Response Handling Guidelines:

- *If a student asks for the answer* → Remind them your role is to help develop strategy; redirect to appropriate step; offer targeted hint.
- *If a student derails the conversation* → If problem-related, acknowledge and steer back to planning; if unrelated, remind of main goal and staying on task

Implementation

We developed STEPS as a stand-alone web application to host the tutor interface and manage student interactions. For this study, the system called the OpenAI API to access the GPT-4o model, which processed the prompts and generated responses in multiturn conversations with students. The model version was fixed for the duration of the study and did not automatically update; all interactions reported here were generated using the same model configuration to ensure reproducibility. The web interface captures student inputs, sends API requests with the prompt and conversation history, and displays the model's responses in a conversational format. Figure 1 shows the web interface of the tutor. An example student−chatbot conversation is provided in the Supporting Information for illustration of system behavior (see Textbox S1). The complete, anonymized interaction dataset is publicly available via the Salehi Lab Redivis repository.[36]

STEPS was integrated into an introductory chemistry course that emphasizes foundational concepts and problem-solving skills at an R1 university in spring 2025. The four instances of STEPS implementations were spread across the 10-week academic quarter. Table 1 presents the problems used in each implementation and their corresponding target knowledge and skills. A detailed timeline for the development and deployment of STEPS activities is shown in Table S11 in the Supporting Information. Students were informed that use of the tool was expected as part of the course's learning activities, and they could choose whether their chat and survey data could be analyzed for research purposes. The study was approved by the Stanford Institutional Review Board (IRB#70070). Participation was voluntary and had no effect on course grades.

We initially introduced STEPS during class time, which allowed students to work with the tutor under the supervision of the instructor and research team. This enabled us to closely observe how students interact with the tutor and identify areas for improvement. We intended to incorporate STEPS into an activity in which students would first chat with the AI tutor independently to create a plan for solving the assigned problem and then gather in groups to discuss these individually developed problem-solving plans. However, we observed that (1) some students required more than the allotted 15 min to complete their individual chat within the class period, and (2) that it was unnatural for students to chat on their computers independently when they could discuss their problem-solving approach with classmates. Based on these observations, we changed the implementation structure so that students chatted with STEPS as part of their homework assignments to help them solve one problem in the problem set. Integrating STEPS into homework assignments allows students to receive timely, personalized feedback when they are more likely to be working alone and can benefit from scaffolding support without social pressure. Students accessed the STEPS learning experience through a Qualtrics survey. After reading the consent form and indicating whether they agreed to have their data analyzed for research purposes, students proceeded through the following steps:

https://doi.org/10.1021/acs.jchemed.6c00039
J. Chem. Educ. 2026, 103, 5555−5563

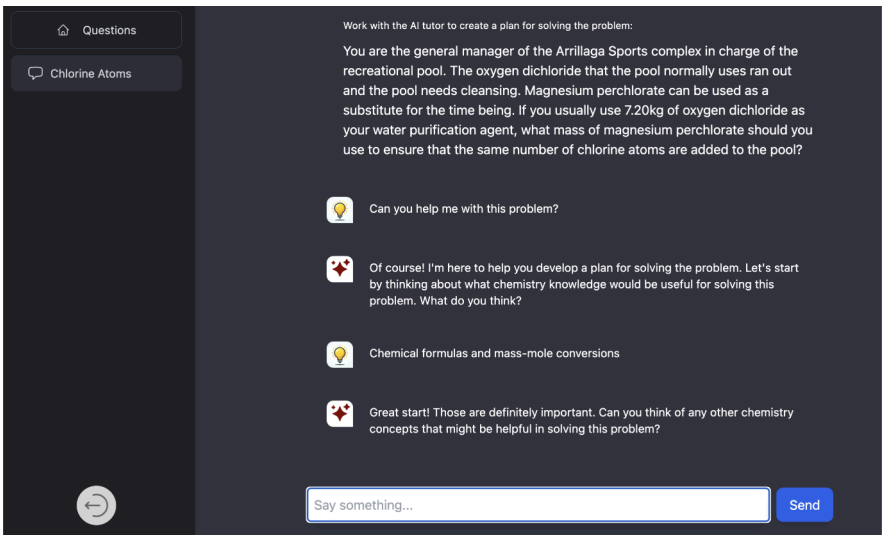


Figure 1. Screencast of the STEPS user interface.

Table 1. Problem Characteristics and Schedule in the 10-Week Quarter

Week	Problem	Underlying Chemistry Concepts and Skills	Format
Week 2	Jet Fuel	Unit conversion	in-class
Week 3	Isotope	Isotopes, atomic mass	homework
Week 5	Chlorine	Molar mass, mole ratio, chemical formulas	homework
Week 8	Barium Titration	Dilutions, acid/base titrations, stoichiometry	homework

Following a design-based research approach,[37] we iteratively revised the prompts based on student feedback from post-interaction surveys (panel 4 in Figure 2) to improve how STEPS asks questions and provides feedback. We also used this feedback to clarify instructions in subsequent problem sets on how to use the tutor and to set clearer expectations with students in the lecture preceding each problem set.

Data Sources

We collected two types of measures: learning outcomes from homework assignments and self-report data from post-interaction surveys administered after each STEPS session and a post-course survey administered at the end of the academic quarter. Table S12 in the Supporting Information provides details on when these surveys were administered. Learning outcomes were assessed through performance on the STEPS-supported problems from weekly problem sets, which teaching assistants (TAs) graded using a rubric based on a 5-point scale. Post-interaction surveys administered immediately after each STEPS session captured students' perceptions of the tool's helpfulness, their confidence in their problem-solving plans, and their likes and dislikes of the STEPS tutor through both Likert-scale items and open-ended questions. Additionally, the post-course survey included open-ended questions asking students to reflect on how their interactions with STEPS evolved over time and to provide advice to future students on working effectively with the tool. The full survey instruments, including all post-interaction and post-course items, are provided in the Supporting Information.

Data Analysis

To assess students' problem-solving outcomes, we examined the scores that students received for the problems in their homework assignments for which they had chatted with STEPS to develop a problem-solving plan. To understand student perceptions of STEPS and identify areas for tool improvement, we analyzed student responses in the post-interaction and post-course surveys. A combination of inductive and deductive coding was employed following a hybrid thematic analysis approach.[38] Themes were first derived inductively from students' open-ended responses about what students found helpful (strength) and what they identified as areas for improvement. Because STEPS was designed to replicate effective human tutoring behaviors, inductively-derived strength themes were interpreted and refined through the INSPIRE framework of effective human tutoring[39] to establish theoretical grounding for the final codes. The dimensions most salient to the AI tutoring context include Socratic engagement through asking questions, indirect guidance through offering hints rather than answers, progressive scaffolding of increasing specificity, and an encouraging and nurturing tone. Improvement themes remained primarily inductive, as challenges unique to LLM-based tutoring may not be fully anticipated by frameworks developed in the context of human tutoring. Two researchers (K.W. and J.T.A.) read student open-ended responses and documented the initial themes and codes. The codes and descriptions for each code were then discussed among all investigators. The two researchers independently coded 20% of the data to establish inter-

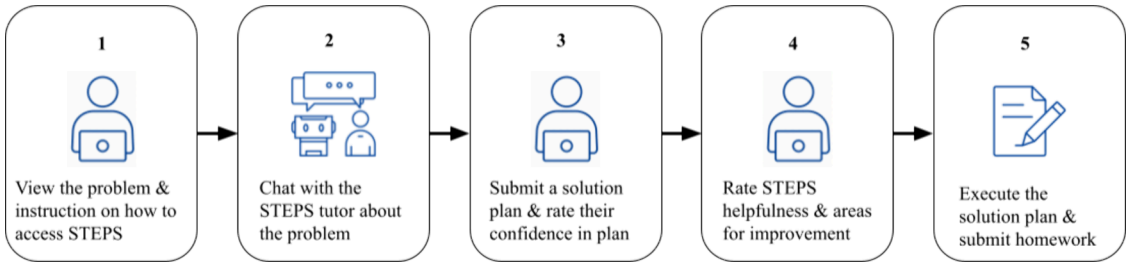


Figure 2. Procedure for STEPS learning activity from the student perspective.

https://doi.org/10.1021/acs.jchemed.6c00039
J. Chem. Educ. 2026, 103, 5555−5563

rater reliability of above 80% before dividing the remaining responses and coding them independently.

RESULTS

Student Problem-Solving Performance

We first examined student confidence in their problem-solving plans to assess the effectiveness of STEPS in scaffolding the planning stage of problem solving. Overall, students reported high confidence that their problem-solving plans developed after chatting with STEPS would lead to correct solutions, with the majority of responses falling between somewhat confident (4 on a 5-point scale) and very confident (5 on a 5-point scale). We then examined homework scores on the target problems. The Jet Fuel problem, implemented as an in-class activity, did not have solution scores as students did not submit their final solutions. For the three homework assignments, students achieved strong performance with average scores ranging from 4.59 to 4.80 out of 5.00 points (Table 2).

Table 2. Student Confidence and Solution Scores on STEPS-Assisted Problems[a]

Problem	Student confidence	Average score out of 5.00	*N*
Jet Fuel	3.92 (1.06)	–	24
Isotope	4.29 (0.72)	4.71 (0.71)	34
Chlorine	4.39 (0.82)	4.80 (0.61)	29
Barium Titration	4.26 (0.81)	4.59 (0.65)	27

[a]Note. Problems were graded by the course instructional team on a 5-point scale: with 2 points were allocated for a correct final answer and 3 points were awarded for demonstrating appropriate problem-solving work and reasoning.

Student Ratings of STEPS Helpfulness

After interacting with STEPS, students rated the tool's helpfulness in supporting their problem-solving compared to other generative AI tools like ChatGPT in three aspects: overall helpfulness, helpfulness of the AI tutor's questions, and helpfulness of the AI tutor's feedback. Responses were collected using a 5-point scale (from 1 (not helpful at all) to 5 (extremely helpful)). Figure 3 presents student ratings across the four problems. On average, students found STEPS helpful in supporting their problem-solving, with ratings showing steady increases from the first to the last implementation and approaching "Very helpful". Students also rated the AI tutor's questions and feedback comparably, with both aspects receiving similar helpfulness scores across all implementations.

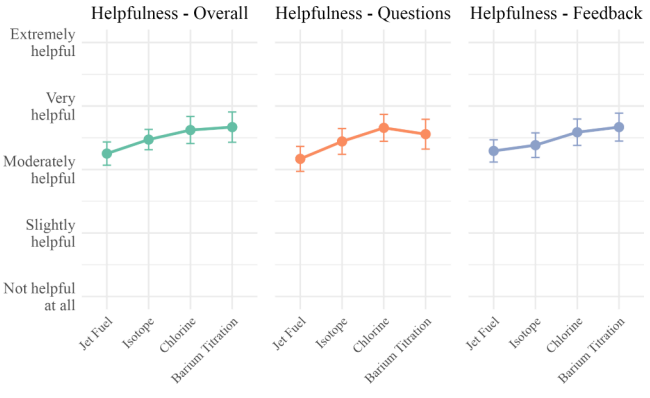


Figure 3. Student ratings of STEPS's overall helpfulness, question helpfulness, and feedback helpfulness across four problems.

Student-Perceived Strengths

When asked what aspects of the interactions with the AI tutor were helpful in supporting their problem-solving/learning, students identified several strength areas that aligned with the tool's pedagogical design, including (1) providing guidance on how to solve the problem without directly giving solutions (e.g., "I liked how it didn't give you the answers but rather helped you reach them."), (2) asking questions to promote independent problem-solving (e.g., "I liked how it kept asking me questions, because it was forcing me to actually think through the problem."), (3) providing helpful feedback and explanations (e.g., "It instilled my original thoughts and gave feedback if I was wrong, helping me narrow down my answer."), (4) maintaining an encouraging and supportive tone (e.g., "I enjoyed how encouraging it was."), and (5) custom-designed for the course and chemistry problem-solving (e.g., "I appreciate how much the AI tutor fits to the learning model proposed in class—it's very consistent in helping reinforce that approach to problem solving.") Tables S1−S4 in the Supporting Information present key themes, sample quotes from students, and frequency of each theme in student responses for the four implementations. Note that responses sum to more than the total number of respondents because multiple codes were assigned to the same response when students mentioned several themes.

Student perceptions of STEPS's strengths evolved from the first implementation to the final implementation. Figure 4 illustrates how the prevalence of the top strength codes explicitly mentioned in student comments shifted across the four implementations. One notable change was an increase in students citing the ability of STEPS to provide helpful feedback and explanations as its major strength, rising from 17% in the first implementation to 37% in the final implementation. This shift likely reflects both iterative prompt improvements and students' developing proficiency in formulating effective questions. Conversely, the emphasis on STEPS guiding problem-solving without providing direct solutions as its main strength decreased from 42% in the first implementation to 15% in the final one. Across the four implementations, 21% to 31% of students acknowledged STEPS' ability to ask questions that promote independent thinking.

Student-Perceived Areas for Improvements

When asked what aspects of the interactions with the AI tutor could be improved to support their problem solving, students identified several key areas for improvement. The most commonly cited concern was unclear responses or explanations. Students noted that STEPS sometimes failed to adequately address their questions or provide explanations at an appropriate level of detail (e.g., "I feel like sometimes the questions I asked weren't answered very well" and "When I asked for the AI tutor to write their comment in simpler terms, I was still confused. It would be helpful to improve its ability to explain in a more detailed manner"). Students also expressed frustration with STEPS's rigid problem-solving approach. Some felt constrained by the structured questioning sequence, particularly when their preferred solution approach differed from the framework STEPS was designed to support (e.g., "It wouldn't let me use my own solution path. I could be doing things in a different order without messing up the solution but

https://doi.org/10.1021/acs.jchemed.6c00039
J. Chem. Educ. 2026, 103, 5555−5563

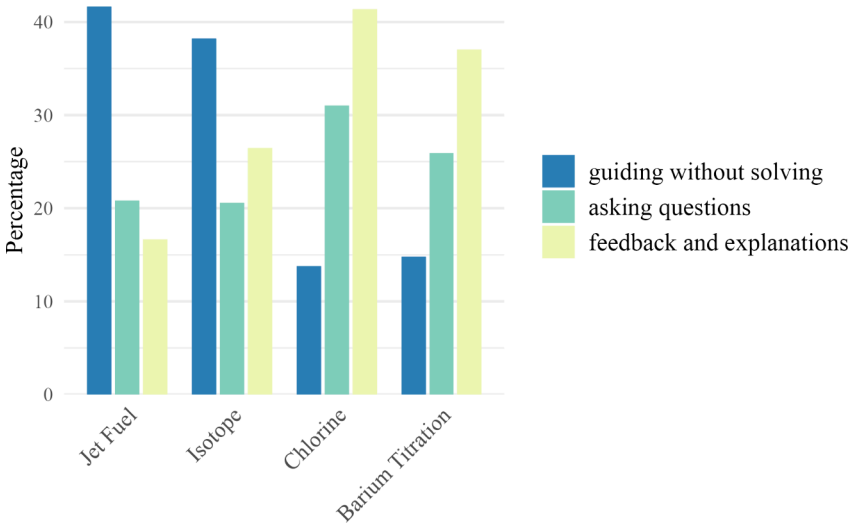


Figure 4. Percentages of coded student responses reflecting what they found most helpful about the STEPS Tutor across the four problems.

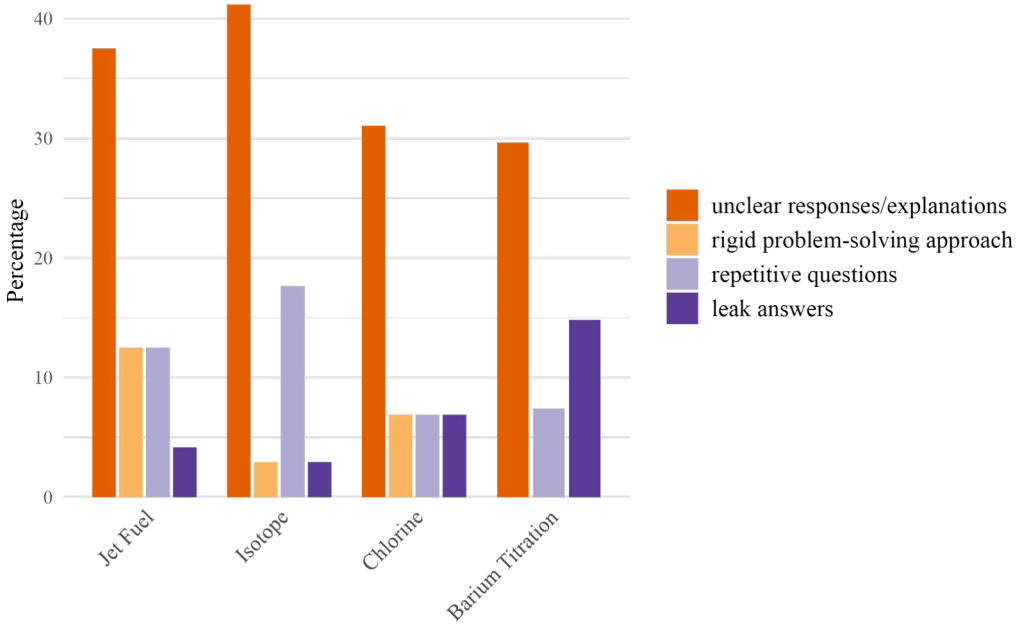


Figure 5. Percentages of coded student responses identifying areas for improvements for the STEPS Tutor across the four problems

it wouldn't let me"). Another concern was repetitive questioning, especially when students felt genuinely stuck and needed more direct guidance rather than additional prompts for independent thinking (e.g., "One frustrating aspect of the AI tutor was when I was confused about a step, it just kept prompting me to think about an aspect of the question when I really was just looking for an answer because I did not know"). Finally, some students reported that STEPS occasionally leaked answers by providing overly direct guidance when they sought clarification (e.g., "When you ask the tutor to ask more guided questions or explain that you don't know, it sometimes gives you the direct steps without letting you try on your own"). Other areas of improvement reported by students include requests for additional resources and more examples, assistance with finding and sorting given values, and concerns that the AI tutor might be too encouraging and prevent them from recognizing mistakes in their own answers. Figure 5 presents how the prevalence of the main improvement themes evolved across the four implementations. Tables S5−S8 in the Supporting Information present key themes, sample quotes from students, and frequency of each theme in student responses for the four implementations.

Based on feedback from the first implementation, we updated the prompts to reduce repetitive questioning and instructed the LLM to provide hints and answers when students provided partially correct responses rather than always requiring complete correct answers. Following this refinement, students less frequently reported repetitive questioning as an issue. This issue was raised by only 7% of students on the final STEPS implementation compared to 18% on the second and 13% on the first; however, this prompt modification also unintentionally increased instances of the AI leaking correct answers from 4% to 15%. These data demonstrate the challenge of balancing appropriate scaffolding with maintaining productive struggle in LLM-based tutoring, a balance that is hard to achieve via prompt engineering alone. In addition, students did not criticize STEPS's "rigid problem-solving approach" in the final implementation, suggesting they

https://doi.org/10.1021/acs.jchemed.6c00039
J. Chem. Educ. 2026, 103, 5555−5563

had adapted to the structured expert problem-solving framework over time through both increased use of STEPS and in-class support and emphasis of the importance of the problem-solving process.

Student-AI Interactions Across Implementations

We also found changes in how students interacted with the AI tutor across multiple implementations. Specifically, the post-survey data suggest that students developed more sophisticated AI literacy skills and learned to leverage STEPS to scaffold their problem-solving rather than directly generate solutions. When asked whether and how they changed their interaction approach from the first to the last problem, the majority of students who responded to the post-course survey (10 out of 15) reported that they improved their question-asking behavior. On one hand, students reported that they developed the agency to steer conversations with STEPS and asked more questions. One student explained, "instead of just following along with the prompts, I asked my own questions in order to gain a better understanding." On the other hand, students also learned to ask targeted questions focusing on the problem-solving process rather than seeking direct solutions. As one student noted, "I've asked it more questions about the process, because it's helping me rethink how I approach problems." Similarly, another student reported that, "I think I have gotten better at asking questions. Learning how to ask whether this be about the process and not the answer." This shift toward more self-directed, process-oriented questioning was echoed by a third student: "I asked more clear and direct questions to get more helpful answers that did not give me the answer but guided me to finding the correct answer."

Beyond asking more and better questions, students exhibited a mindset shift from using AI to directly find solutions to focusing on planning the problem-solving process, such as first determining the relevant concepts. One student articulated this shift: "I figured it's better to think about how to start the problem first and brainstorm some concepts I could utilize to solve the problem, then use the AI tutor so it could let me know if I'm on the right track." This shift represents students' growing understanding of how to leverage AI as a thought companion rather than as a solution provider. It also suggests that students recognized identifying relevant concepts for a problem as a key decision in the planning phase.

Indicators of Developing AI Literacy

When asked to provide advice for future students on how to effectively use STEPS, students consistently emphasized the importance of asking questions and taking ownership of the problem-solving process rather than seeking direct solutions. One student recommended, "Don't try to get the AI tutor to just give you the answer through each step. Instead, use it to brainstorm some of your own ideas first, and if you get stuck, ask it to guide you through where you went wrong." Students also highlighted the value of deepening understanding through follow-up questions, for example, how to prompt the AI tutor to provide better feedback: "if the tutor says you're right, but you don't fully understand then ask why you're right." These recommendations suggest that students recognized effective AI use requires deliberate cognitive engagement and prompting strategies.

DISCUSSION

This study demonstrates that effective prompt engineering for adapting LLMs as tutors requires a systematic approach that translates pedagogical frameworks into specific instructional prompts. Our adaptation of a paper-and-pencil problem-solving template, which is itself grounded in empirical research, into an interactive tutor required both structured questioning sequences aligned with expert problem-solving practices and behavioral guidelines that prevented answer leaking while providing a proper level of scaffolding to maintain student engagement. While the initial prompts were based on the problem-solving practices and behavioral guidelines, the iterative prompt refinement process, informed by teaching-team testing and feedback from students, was essential for developing pedagogically effective prompts to handle a wide range of interaction styles and prior knowledge and skills among students in the targeted class.

Our findings reveal that successful AI tool integration requires careful consideration of both pedagogical and practical factors, guided by specific learning goals (in this case, developing competency in problem-solving practices, particularly ones related to the planning stage). In our case, AI instructional integration to support teaching problem-solving was achieved through interdisciplinary collaboration between education researchers, EdTech engineers, and STEM course instructors. Because STEPS is designed around a problem-solving framework that is central to the course, the time and effort invested in its development create a reusable resource for more students in future offerings of the course. Furthermore, students recognized the value of an AI tool specifically designed for particular learning goals with instructor input and aligned with course content, contrasting it favorably with general-purpose tools like ChatGPT. The transition from in-class to homework-based implementation addressed the key usability issues of time constraints and social dynamics of interacting with AI in class, providing students with convenient access to the tutor when they were most in need of timely, personalized problem-solving support, and where the traditional human tutoring methods face resource limitations to provide such feedback. The two-step process of chatting with STEPS to develop a plan then solving the problem independently was designed to promote student accountability, as they knew they would need to submit their final solutions as part of homework assignments. Finally, the survey accompanying each implementation of the tutor facilitated continuous feedback collection, enabling iterative improvement to both the tutor and the in-class discussion of the problem-solving process based on student experiences.

The implementation of our AI tutor supported positive outcomes across multiple dimensions of student learning and engagement. Student problem-solving performance remained consistently high across all three problems. Moreover, student perceptions of the tool's helpfulness showed steady improvement from the initial to the final implementation. Most significantly, students demonstrated growth in general AI literacy, as they learned to ask more and better questions when interacting with the AI tutor and viewed AI as a learning partner instead of a solution generator. Student advice to future cohorts consistently emphasized active engagement with the problem-solving process and strategically asking AI questions to support problem-solving and deepen understanding, indicating that they have begun to develop productive AI interaction strategies. These findings align with prior work showing that productive engagement with generative AI in complex chemistry tasks depends on students' ability to structure prompts, interpret outputs, and integrate AI

https://doi.org/10.1021/acs.jchemed.6c00039
J. Chem. Educ. 2026, 103, 5555−5563

contributions into their reasoning processes, rather than outsourcing cognition to the tool.[40] However, the absence of a control condition limits our ability to attribute these outcomes directly to STEPS. Future work will incorporate quasi-experimental designs to more rigorously evaluate the impact of AI tutors on student learning and AI literacy development.

CONCLUSION AND FUTURE DIRECTIONS

Through this project, we demonstrate that a prompt-engineered AI tutor, grounded in a validated problem-solving practice framework and integrated into an introductory chemistry course, can scaffold students' problem-solving planning and support the development of productive AI interaction strategies. Besides its role as an instructional resource, STEPS also serves as a research platform that captures rich data on student−AI interactions. We are currently analyzing the chat data to understand the types of interaction patterns that emerge among students and identify which interaction patterns lead to optimal problem-solving and learning outcomes. This analysis will help refine both the tool design and instructional approach for future implementations, such as having instructors demonstrate effective ways of using the tool in class before having students use it independently. In parallel, we are developing a new version of the STEPS platform that allows instructors to (1) create and assign their own problems to their students and (2) directly monitor student engagement with the tutor. The goal is to make the tool accessible to a broader range of courses, students, and STEM disciplines.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available at https://pubs.acs.org/doi/10.1021/acs.jchemed.6c00039.

> Student-reported strengths and areas for improvement of the AI tutor across four implementations; chemistry problems used in the study; API costs; example student−chatbot conversation; full AI tutor prompt; implementation timeline and workflow; post-interaction and post-course survey instruments (PDF)

AUTHOR INFORMATION

Corresponding Author

Joshua T. Arens − *Department of Chemistry, Stanford University, Stanford, California 94305, United States;* orcid.org/0009-0006-4306-2377; Email: jarens@stanford.edu

Authors

Karen D. Wang − *School of Information, San José State University, San Jose, California 95192, United States*

Chinat Yu − *Graduate School of Education, Stanford University, Stanford, California 94305, United States*

Jordyn N. Smith − *Department of Chemistry, Stanford University, Stanford, California 94305, United States;* orcid.org/0009-0005-1522-8862

Hannah Bartels − *Department of Chemistry, Stanford University, Stanford, California 94305, United States;* orcid.org/0000-0003-2649-999X

Mercy Haub − *Department of Chemistry, Stanford University, Stanford, California 94305, United States*

Fun Man Fung − *Department of Chemistry, Stanford University, Stanford, California 94305, United States; School of Chemistry, University College Dublin, Dublin D04 V1W8, Ireland;* orcid.org/0000-0003-4106-3174

Nick Haber − *Graduate School of Education, Stanford University, Stanford, California 94305, United States*

Shima Salehi − *Graduate School of Education, Stanford University, Stanford, California 94305, United States*

Jennifer Schwartz Poehlmann − *Department of Chemistry, Stanford University, Stanford, California 94305, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.6c00039

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Carl Wieman, Eric Burkholder, Chas Blakemore, and Zhangyang Wu for their input on developing the initial LLM prompt for STEPS and for their feedback on this project. This work was supported by the Gates Foundation (Investment ID: INV-086418) and the Curriculum Transformation Project at Stanford. This work was also supported by the 2024-2025 Fulbright U.S.-ASEAN Scholarship awarded to F.M.F. for his work at Stanford as a visiting scholar.

REFERENCES

(1) Tsaparlis, G. *Problems and Problem Solving in Chemistry Education: Analysing Data, Looking for Patterns and Making Deductions*; Royal Society of Chemistry, 2021.

(2) *AI Index Steering Committee The AI Index 2025 Annual Report*, 2025 (accessed: June 2, 2025).

(3) Muteti, C. Z.; Zarraga, C.; Jacob, B. I.; Mwarumba, T. M.; Nkhata, D. B.; Mwavita, M.; Mohanty, S.; Mutambuki, J. M. I realized what I was doing was not working: the influence of explicit teaching of metacognition on students' study strategies in a general chemistry I course. *Chemistry Education Research and Practice* **2021**, *22*, 122−135.

(4) Yuriev, E.; Naidu, S.; Schembri, L. S.; Short, J. L. Scaffolding the development of problem-solving skills in chemistry: guiding novice students out of dead ends and false starts. *Chemistry Education Research and Practice* **2017**, *18*, 486−504.

(5) Hunter, R. A.; Pompano, R. R.; Tuchler, M. F. In *Active Learning in the Analytical Chemistry Curriculum*; Wenzel, T. J., Kovarik, M. L., Robinson, J. K., Eds.; ACS Symposium Series, Vol. *1409*; American Chemical Society, 2022; Chapter 15, pp 269−295.

(6) Prince, M. J.; Felder, R. M. Inductive teaching and learning methods: Definitions, comparisons, and research bases. *Journal of engineering education* **2006**, *95*, 123−138.

(7) Cardellini, L. Fostering creative problem solving in chemistry through group work. *Chemistry Education Research and Practice* **2006**, *7*, 131−140.

(8) Ramos, B.; Condotta, R. Enhancing learning and collaboration in a unit operations course: Using AI as a catalyst to create engaging problem-based learning scenarios. *Journal of Chemical Education* **2024**, *101*, 3246−3254.

(9) Chen, Z. Reliable generation of isomorphic physics problems using generative AI with prompt-chaining and tool use. *arXiv Preprints* **2025**, arXiv:2508.14755, URL: https://arxiv.org/abs/2508.14755.

(10) Gupta, A.; Reddig, J.; Calo, T.; Weitekamp, D.; MacLellan, C. J. Beyond final answers: Evaluating large language models for math tutoring. Presented at the *International Conference on Artificial Intelligence in Education*, 2025; pp 323−337.

(11) Dinucu-Jianu, D.; Macina, J.; Daheim, N.; Hakimi, I.; Gurevych, I.; Sachan, M. From Problem-Solving to Teaching Problem-Solving: Aligning LLMs with Pedagogy using Reinforcement Learning. *arXiv Preprints* **2025**, arXiv:2505.15607, URL: https://arxiv.org/abs/2505.15607.
(12) Vaccaro, M.; Almaatouq, A.; Malone, T. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* **2024**, *8*, 2293−2303.
(13) Kim, J. Integrating artificial intelligence (AI) chatbots and green chemistry principles in the synthesis of cyclohexene. *Journal of Chemical Education* **2025**, *102*, 3058−3064.
(14) Cooper, M. M. A Preliminary Set of Principles to Support Learning in the Context of Generative AI. *Journal of Chemical Education* **2026**, *103*, 36−42.
(15) Clark, T. M.; Tafini, N. Design and Implementation of a General Chemistry Course That Promotes AI Use. *Journal of Chemical Education* **2025**, *102*, 4017−4023.
(16) Freeman, J. *Student Generative AI Survey 2025*; Higher Education Policy Institute: London, U.K., 2025.
(17) Black, R. W.; Tomlinson, B. University students describe how they adopt AI for writing and research in a general education course. *Sci. Rep.* **2025**, *15*, 8799.
(18) Wang, K. D.; Wu, Z.; Tufts, L.; Wieman, C.; Salehi, S.; Haber, N. Scaffold or Crutch? Examining College Students' Use and Views of Generative AI Tools for STEM Education. *2025 IEEE Global Engineering Education Conference (EDUCON)*. 2025; pp 1−10.
(19) Hou, C.; Zhu, G.; Sudarshan, V.; Lim, F. S.; Ong, Y. S. Measuring undergraduate students' reliance on Generative AI during problem-solving: Scale development and validation. *Computers & Education* **2025**, *234*, 105329.
(20) Yu, B.; Baker, F. N.; Chen, Z.; Herb, G.; Gou, B.; Adu-Ampratwum, D.; Ning, X.; Sun, H. ChemToolAgent: The Impact of Tools on Language Agents for Chemistry Problem Solving. *arXiv Preprints* **2024**, arXiv:2411.07228, URL: https://arxiv.org/abs/2411.07228
(21) Morbidoni, C.; Mascitti, A.; Sarra, A.; Tonucci, L.; Coccia, F. Exploring the potential of using ChatGPT in chemistry education. *J. Chem. Educ.* **2026**, *103*, 286.
(22) Clark, T. M. Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *Journal of Chemical Education* **2023**, *100*, 1905−1916.
(23) Ding, L.; Li, T.; Jiang, S.; Gapud, A. Students' perceptions of using ChatGPT in a physics class as a virtual tutor. *Int. J. Educ.Technol. High. Educ.* **2023**, *20*, 63.
(24) Young, J. D.; Dawood, L.; Lewis, S. E. Chemistry students' artificial intelligence literacy through their critical reflections of chatbot responses. *Journal of Chemical Education* **2024**, *101*, 2466−2474.
(25) Kosmyna, N.; Hauptmann, E.; Yuan, Y. T.; Situ, J.; Liao, X.-H.; Beresnitzky, A. V.; Braunstein, I.; Maes, P. Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task. *arXiv Preprints* **2025**, arXiv:2506.08872.
(26) Kapur, M. Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist* **2016**, *51*, 289−299.
(27) Salehi, S. *Improving Problem-Solving through Reflection*; Stanford University, 2018.
(28) Price, A. M.; Kim, C. J.; Burkholder, E. W.; Fritz, A. V.; Wieman, C. E. A detailed characterization of the expert problem-solving process in science and engineering: Guidance for teaching and assessment. *Life Sci. Educ.* **2021**, *20*, ar43.
(29) Burkholder, E.; Miles, J.; Layden, T.; Wang, K.; Fritz, A.; Wieman, C. Template for teaching and assessment of problem solving in introductory physics. *Physical Review Physics Education Research* **2020**, *16*, 010123.
(30) Schwartz Poehlmann, J. K.; Nardo, J. E.; Rojas, M.; Salehi, S. Introducing the problem-solving template as a tool for equity: Addressing incoming preparation disparities. *Journal of Chemical Education* **2024**, *101*, 1332−1340.
(31) Vo, K.; Sarkar, M.; White, P. J.; Yuriev, E. Exploring problem-solving scaffolds in general chemistry: a systematic review of scaffolding goals and instructional approaches. *Journal of Chemical Education* **2025**, *102*, 1004−1018.
(32) Labadze, L.; Grigolia, M.; Machaidze, L. Role of AI chatbots in education: systematic literature review. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 56.
(33) Albadarin, Y.; Saqr, M.; Pope, N.; Tukiainen, M. A systematic literature review of empirical research on ChatGPT in education. *Discover Educ.* **2024**, *3*, 60.
(34) Xiao, R.; Ye, R.; Hou, X.; Wen, J.; Kumar, H.; Liut, M.; Stamper, J. Transforming GenAI Policy to Prompting Instruction: An RCT of Scalable Prompting Interventions in a CS1 Course. *arXiv Preprints* **2026**, arXiv:2602.16033, URL: https://arxiv.org/abs/2602.16033.
(35) Clerc, O.; Abdelghani, R.; Desvaux, C.; Poisson, E.; Oudeyer, P. Y.; Sauzéon, H. Teaching Students to Question the Machine: An AI Literacy Intervention Improves Students' Regulation of LLM Use in a Science Task. **2026**, arXiv:2604.01955, URL: https://arxiv.org/abs/2604.01955.
(36) Lab, S. AI Tutor Conversation Dataset. *Redivis Data Repository*. Available via the Internet at: https://stanford.redivis.com/Salehi_Lab (accessed: May 2, 2026).
(37) Collective, D.-B. R. Design-based research: An emerging paradigm for educational inquiry. *Educ. Res.* **2003**, *32*, 5−8.
(38) Xu, W.; Zammit, K. Applying thematic analysis to education: A hybrid approach to interpreting data in practitioner research. *Int. J. Qual. Methods* **2020**, *19*, 1609406920918810.
(39) Lepper, M. R.; Woolverton, M. *Improving Academic Achievement*; Elsevier, 2002; pp 135−158.
(40) Haraldsrud, A.; Odden, T. O. B. Bridging artificial intelligence and real intelligence: self-scaffolding computational modeling with generative AI in chemistry. *Journal of Chemical Education* **2025**, *102*, 4255−4266.

https://doi.org/10.1021/acs.jchemed.6c00039
J. Chem. Educ. 2026, 103, 5555−5563