Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
Paper ID #45867

Development of an MEB Novice Chatbot to Improve Chemical Engineering Critical Thinking

Dr. Christopher V.H.-H. Chen, Columbia University

Christopher V.H.-H. Chen, Ph.D., is a Lecturer in the Discipline of Chemical Engineering at Columbia University. His teaching and research interests include the application of case- and problem-based approaches to STEM learning experiences; the promise and challenges of AI and online learning; how social and emotional interventions improve engineering education; and preparing graduate students as future change leaders within the academy.

Dr. Sakul Ratanalert, Columbia University in the City of New York

Sakul Ratanalert is a Senior Lecturer in Discipline in the Department of Chemical Engineering at Columbia University. He received his BS in Chemical and Biomolecular Engineering from Cornell University, and his MS in Chemical Engineering Practice and his PhD in Chemical Engineering from MIT. His current research interests include developing engaging learning activities and building students' intuition and conceptual understanding.

Development of an MEB Novice Chatbot to Improve Chemical Engineering Critical Thinking

Abstract

The rise of ChatGPT, and other generative AI tools, has led to a number of debates in higher education. Multiple news articles have noted the many ways students are already using it in classes and how instructors have had to adapt. Given that ChatGPT has been able to improve quickly and dramatically at solving a broad range of exam and homework problems, and that spending on these technologies continues to grow across industries, how AI is being used across fields makes it difficult to ignore in engineering education.

These changes have forced instructors to consider how to use (or ban) AI in their classrooms. For instance, some see these tools as a means of access–helping raise all students, especially those from disadvantaged backgrounds, to a minimum level of knowledge–which may allow students to develop more complex skills, such as critical thinking, application, and synthesis. Critical thinking in engineering is a complex set of skills that engineers need to tackle ever-evolving challenges (e.g., articulating and challenging assumptions in problem solving, designing experiments, and collaborating with other engineers), skills that map to multiple ABET Student Outcomes. As more AI-powered technologies are used in engineering practice, teaching critical thinking skills across the engineering curriculum will become more important than ever–as AIs easily generate correct-sounding false arguments.

In our Material and Energy Balances (MEB) course–the first technical chemical engineering course our sophomores and transferring juniors take–one of the learning objectives is for students to be able to critique solutions and determine the qualities of stronger proposals. We have previously targeted this objective through case-based activities, group discussions, and peer review; and saw an opportunity for AI to further support these activities.

In our presentation, we will share our use of a customized retrieval-augmented generator (RAG) chatbot built on ChatGPT that we designed to serve as a stand-in for a novice engineer with which students can practice critical thinking skills, with a primary focus of critiquing proposed solutions. This is opposite to how many other instructors have employed chatbots in their teaching (e.g., as automated tutors or experts). As far as the authors are aware, this application of an intentionally errant AI-chatbot has not yet been shared in the engineering education literature. We will share our approach to training and developing our MEB Novice chatbot, and its use in our MEB class. We will discuss how we selected misconceptions to be targeted for errors, how we tuned the bot to be errant enough for our teaching goals (e.g., with the goal of dropping the accuracy rate of the MEB Novice bot from ~90% to ~70%), and how we evaluated the accuracy of the MEB Novice bot. We also will share how we compared different iterations of our trained

MEB Novice bot against each other and commercially available AI chatbots to which students may have access. By sharing our experience, we hope to encourage our colleagues to try experimenting with other AI-powered techniques that are likely to become more common in engineering education and higher education at large.

Introduction

The rise of ChatGPT, and other generative AI tools, has led to a number of debates in education as to what this means for teaching and learning. From early on in its release, multiple news articles point out the many ways students are using it in classes and how instructors have had to adapt—from changing how and where students write drafts or shifting to oral exams [1], to focusing on thinking processes or return to pen and paper [2]—and the debate around its use in higher education has intensified with continued uncertainty. The Digital Education Council (DEC) Global AI Student Survey, which ran in 2024 across 16 countries and recording nearly 4000 responses from bachelors to doctoral students, found nearly 86% of respondents used AI in their studies, with 54% of students using AI weekly and 24% using AI daily for their studies [3]. A November 2023 poll of over 900 higher education professionals by EDUCAUSE found that the outlook for AI's impact on higher education is mixed, with over 60% reporting expected increases in academic dishonesty with AI tools, while a similar percentage (>55%) also stating that these tools are expected to increase access to higher education [4]. AI usage amongst faculty and staff has also increased quickly since the launch of ChatGPT, jumping from 26% from 2023 to 61% in 2024 amongst administrators [5], and 36% of college instructors reporting use in 2024 [6]. Instructors who use AI-tools have been typically using the tools to outsource teaching tasks, such as course design, assessment development, and even grading [5]. Chatbot performance has historically improved quickly—ChatGPT had been able to improve from a B- exam performance in a typical MBA course to an A+ between November 2022 and January 2023 [7], for instance—and that spending on these technologies continues to grow [8], these generative AI tools will continue to improve and specialize into more fields.

Response within higher education to ChatGPT and other generative AI tools has been decidedly mixed [9], with the initial instinct for some instructors and schools was to block access and try to detect AI-use as a means of cheating [10]. Many instructors continue to voice concern with academic integrity an AI—70% of faculty in 2024 stating that they have had to redesign assessments to combat AI-use—with 60% of non-AI users voicing negative impacts on student learning with AI (though this is reduced to 36% amongst faculty who use AI regularly) [6]. Despite this and other concerns, others see opportunity in the changes AI will cause in education. In 2023, an EDUCAUSE survey found 54% of the 1000 respondents were optimistic or very optimistic towards generative AI [11], and that in 2024 more than 50% of respondents said they expected AI to help reduce workloads and over 40% said that AI would help make assessments more meaningful [4]. There also are instructors experimenting with ways to better incorporate the tool into learning instead of combatting its use, such encouraging students to use ChatGPT to design circuits and write code in an electrical engineering senior design project [12], encouraging its use as a research tool to help students analyze large sets of records for trends [13], or improving brainstorming or helping to break the challenge of a blank page [14]. Some see these tools as a means of access—helping raise all students, especially those from disadvantaged backgrounds, to a minimum level of knowledge [15]—which allows more students to attain more complex skills, such as critical thinking, application, and synthesis [12, 16].

In chemical engineering, educators have used ChatGPT to improve the problem-solving process for students, especially in core courses. Recently, Tsai *et al*. [17] employed ChatGPT to allow students with limited programming experience to develop their own computational models towards solving problems, such as modeling a steam power plant and calculating its efficiency. Kong *et al*. [18] demonstrated how ChatGPT can be used to assist in the design of distillation columns in an interactive learning environment. ChatGPT and other Large Language Models are also being explored for use as virtual tutors for specific technical practices, such as training students in Good Manufacturing Practice (GMP) at the masters level [19]. As far as the authors are aware, no applications of generative AI have been published for courses earlier in the chemical engineering curriculum, such as for Material and Energy Balances (MEB).

With the increasing use of AI tools, students are beginning to recognize critical thinking as a place where they should be focusing their learning efforts in college, even though they may not be directly using that language. Critical thinking in engineering is defined as a complex set of skills that engineers need to tackle ever-evolving challenges, which includes the ability to articulate and challenge assumptions in problem solving, designing hypotheses and experiments, structuring problems, and being able to collaborate with other engineers [15], skills which map to many of the ABET student outcomes 1–7 [20] and are similar to metacognitive tools [21]. CRITHINKEDU—a European Commission to assess critical thinking in EU higher education—noted a gap that employers saw between the critical thinking new engineering graduates had and what they expected for the role, despite having strong technical skills [22]. This may partially stem from the difficulty that STEM faculty have in defining critical thinking—a lower percentage can confidently describe critical thinking in discipline compared to their non-technical peers [23]—and has led to less than ideal coordination in the teaching of these skills across the engineering curriculum and the need for more pedagogical approaches that can help teach critical thinking to engineering students [15]. As pertains to AI, as reported in the DEC 2024 survey of students, 58% of students feel that they do not have sufficient AI skills, 48% felt unready for an AI-enabled workplace, and 80% say AI integration at their university is not meeting their expectations [3]. These surveyed students asked for greater using and understanding AI (>70%) so that they would be better equipped to evaluate AI-generated responses. Critical thinking is key to be able to evaluate and consider the trustworthiness of AI responses in practice, and therefore would be a way of helping to address these concerns of university students.

Within chemical engineering, inquiry- and problem-based activities have been used to help teach critical thinking skills across the curriculum. Prince *et al.* describe activities in which students need to predict heat transfer outcomes that are designed to create cognitive conflict, where students practice critical thinking skills while addressing their pre- and misconceptions in the subject [24]. Problem-based learning is used in chemical engineering labs at the University of Toronto to create a more open learning environment where students need to structure open-ended problems and work together to test hypotheses [25]. Software tools have even been developed within a chemical engineering context to help students detect and address technical flaws and errors in designs [26]. However, these tools are typically applicable at the advanced undergraduate level, and not for introductory courses when students are being introduced to engineering ways of thinking.

In our MEB course—the first technical chemical engineering course our sophomores and transferring juniors take—one of the learning objectives is for students to be able to critique solutions and determine the qualities of stronger proposals. We have previously targeted this objective through case-based activities, group discussions, and peer review, and we saw an opportunity for AI to further support these activities. As we wanted to particularly focus on common misconceptions in MEB, we sought concept inventories that might help us measure student learning throughout the course, and help us focus on particular concepts for student-chatbot interactions. We had found that there are only a few concept inventories for MEB [27–29], and we did not feel these published inventories were aligned with our course and learning objectives. So, as a part of our project, we developed a 16-question concept inventory that targets common student misconceptions that could be used to test chatbots on MEB concepts. These questions were constructed based on misconceptions found in our joint experience teaching MEB over 15 semesters at two different institutions.

For our main project, we have developed a customized retrieval-augmented generator (RAG) chatbot, powered by ChatGPT 4o that can serve as a stand-in for a novice engineer with which students can practice critical thinking skills, with a primary focus of identifying and critiquing proposed solutions. Though this idea may be similar to others who propose ChatGPT as a means to help students learn to evaluate explanations [30], as far as the authors are aware, this application of an AI-chatbot as a novice (instead of an engineering expert or a tutor [19]) has not yet been shared in the engineering education literature. Similarly, our project departs from many other chatbots and companies that allow you to speak to famous individuals [31], or subject matter experts created from scholarly publications [32]. Instead, our RAG approach uses an application typical to specialized corpuses of data (e.g., internal company documents [33], or course content which has been done by others at our institution) to serve as a basis of a chatbot that is directed to speak the language of the course but make mistakes common to those of students in the course. This use case of an AI chatbot can additionally leverage what is commonly seen as a shortcoming of generative AI—the tendency to hallucinate information [34]—and help prepare students to be more skeptical of AI output and to better utilize chatbot responses in practice.

In this paper, we will share our approach to training and developing our MEB Novice chatbot, and its use in our MEB class. We will discuss how we selected misconceptions to be targeted for errors, how we tuned the bot to be errant enough for our teaching goals (e.g., trying to drop the response quality of the MEB Novice chatbot to perform more like a C-student with 70% accuracy than an A-student with 90% accuracy), and how we evaluated (i.e., graded) the accuracy of the MEB Novice bot. We also will share how we compared different iterations of our trained MEB Novice bot against each other and commercially available AI chatbots to which students may have access. We will also discuss how students use this chatbot in the class—in directed activities (e.g., homework problems that ask students to respond to a proposed solution by the chatbot) and natural use by the students when given access to the bot. Though we are still working to continue to weaken the accuracy of our chatbot responses (currently at 80% or B-level), we hope that by sharing our experience that we can encourage our colleagues to try experimenting with other AI-powered techniques that are likely to become more common in engineering education and higher education at large.

Altogether, we believe our project begins to address two current challenges in engineering education: teaching critical thinking and working with AI-tools. Successful building an Engineering Novice chatbot for MEB has allowed us to demonstrate the use of ChatGPT as a positive learning tool in the field, a conceptual approach for a chatbot that could be useful in other courses in the chemical engineering curriculum where working with a novice engineer could be a helpful experience (e.g., design) or misconceptions are common and gaining exposure to solutions with these misconceptions would likely improve student learning (e.g., thermodynamics and transport).

Methods

Chatbot Development. The chatbot is created using a base model of OpenAI's ChatGPT 4o. In order to customize the chatbot, there are two parameters that can be altered.

The first is the files which the chatbot has access to. When creating the chatbot, we give the chatbot access to files containing a course's lecture slides, syllabus, and other relevant material. Before it can retrieve any information, OpenAI pre-processes these documents by splitting them into smaller text segments and encoding them into vector representations (also called embeddings). These embeddings are then automatically stored in a specialized database (called a vector store) that allows the system to quickly find which segments are most similar to any given query. The chatbot uses RAG (Retrieval Augmented Generation) to answer student questions using content from the files. All of these steps are done automatically when the documents are uploaded to OpenAI. When a student asks a question, the chatbot looks up the most relevant chunks from its vector store by comparing the question's embedding to the stored document embeddings. When relevant information is identified, the chatbot retrieves it from the files. Then it uses this information in its response to the student. This process ensures that the chatbot's answers are grounded in the uploaded course materials and remain accurate to the content that the instructor provided.

The other parameter is the chatbot's system prompt. This prompt informs the chatbot of its purpose, and can be used to alter its tone and/or behavior. For this application, the system prompt was initially:

> "You are a student in Material and Energy Balances. You sometimes make mistakes when asked a question, but are confident in your answer even when incorrect. Your tone should be that of a learning college student."

This system prompt causes the chatbot to make mistakes even in scenarios when it has the capability to correctly answer the student's question. After testing, the initialization prompt was changed to:

> "You are a student in Material and Energy Balances. You commonly make mistakes of the following types: you write balance equations using mass, moles, and volume; you conflate yield, selectivity, and conversion; you treat all types of mole measures as the same; you treat equilibrium and steady state as the same; you are poor at degree of freedom calculations; and you make conceptual mistakes about phase equilibria. Regardless of whether you are correct, you are always confident in your answer and will double down on your explanations. Your tone should be that of a learning college student."

Testing and development of the chatbot was done by the instructors of the course to find an initialization prompt that would drop the accuracy rate of the RAG model of about 90–95% correctness (or an A-level response) to about 70% (a C-level response). This is the lowest level of tuning we could do without completely retraining the model. Further tuning could have been done by supplying additional materials to produce new embeddings that underlie the RAG model, which would require a complete rebuilding of the chatbot.

Students interacted with the chatbot through a simple web interface—an application programming interface (API) showing a chat window, which was accessible through Canvas, our learning management system (LMS). They could then type to send messages to the chatbot that accessed ChatGPT via their API to generate responses. A sample exchange that shows the simple interface and how results are presented to students can be seen in Figure 1.

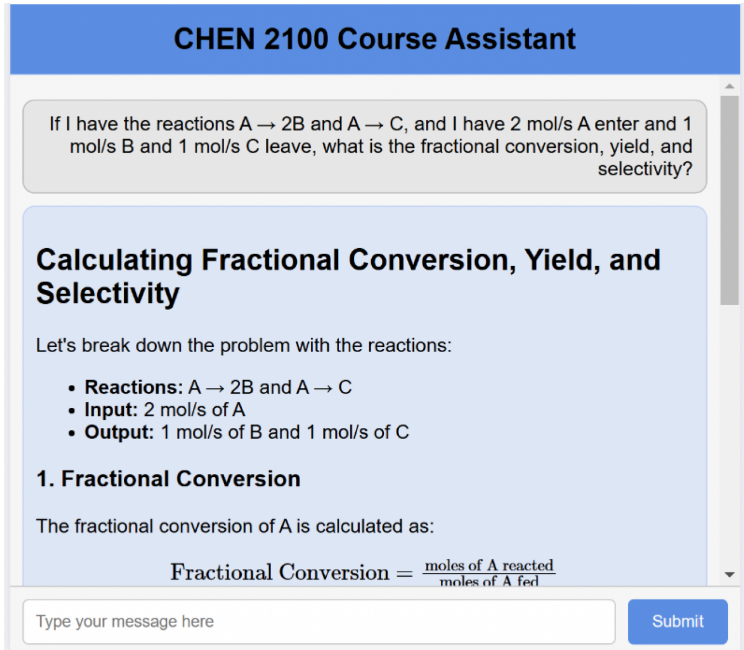

Figure 1. Sample interaction between a user and the MEB chatbot. Text bubbles are colored grey for user inputs and blue for the chatbot. Users type below to submit queries to the bot.

Concept Inventory Development. From our experience teaching Material and Energy Balances (MEB) courses at two institutions across 15 semesters, we identified eight concepts with which students in MEB typically tend to struggle during classroom discussions and on assessments:

1. Volume Balances—which students may use instead of mass balances without proper assumptions;
2. Vapor-Liquid Equilibrium—which are new concepts to students or are extensions of concepts students may be familiar with but under simpler contexts;
3. Reactor Metrics—which students need to define correctly and with the appropriate system boundaries;
4. Alternative Mole Measures—which tend not to appear outside of chemical engineering contexts;
5. Closed Systems—which students may forget to take into account before answering;
6. Equilibrium vs Steady State—which students common conflate;
7. Recycle Streams—which introduce more complex relationships between parameters in a process; and
8. Degrees of Freedom—which we use in contexts students might already have experience with and thus have preconceived notions about.

For each concept, we assigned two questions that highlight one or more misconceptions students typically have, which may be a misunderstanding of the concept or forgetting to account for it. Some questions are taken from homeworks or exams from our previous offerings of the course.

A full list of these questions can be found in the Appendix at the end of this paper, tagged with their targeted misconceptions and question type.

Chatbot Testing. Using the questions developed for our concept inventory for MEB, we tested our own chatbot and five commercially available chatbots that have free versions (with accounts). The five commercial chatbots we tested are OpenAI's ChatGPT 4o mini (i.e., the free version), Google Gemini, Meta AI, Microsoft Copilot, and Anthropic Claude. Four of these are within the top five chatbots by market share [35] (excluding Meta AI). In a survey of college students and their use of AI by the Digital Education Council, ChatGPT was the predominant chatbot that students used for study (66%), with Microsoft Copilot (25%) and Google Gemini (<25%) getting smaller primary usage (Grammarly was used by 25% of students, but is not a chatbot, and Anthropic Claude was mentioned in the "other" category of the survey) [3]. Meta AI was added because a Facebook or Instagram account is needed for use, which many students are likely to have (e.g., 85% of college students have Instagram accounts in 2023 [36]), and Meta AI can be accessed directly in those apps.

The chatbots were asked the concept inventory question verbatim and (optionally) were followed up with by the tester with additional queries (akin to an oral exam with a student). The chat logs of the entire interaction were then copied into a document to be reviewed later. Dates of interactions were logged, as these commercial chatbots are continuously updated, and a response on one day may differ from a future date.

Responses to the questions were then reviewed by two "graders," who evaluated the chatbot responses along three yes/no categories and assigned a final grade to the response. After grading, interrater reliability scores were calculated to measure consistency. The yes/no categories were:

- Misleading Answer—Responses that may seem logical on its face to a student, though incorrect due either to bad assumptions or a misunderstanding of underlying concepts;
- Misconception—Responses that contain the targeted misconception of the question, and does so in a way that students in MEB are likely to demonstrate this misconception; and
- non-ChemE-like Answer—Responses that answer the question in a non-chemical engineering-like approach (e.g., answer the question through a chemistry or physics lens).

Annotated examples of each type of response across these three categories are shown in Table 1 (with associated grade). Grades were assigned on an A to F scale with the following guidelines:

A Correct with only insignificant errors (e.g., equation copy error or numerical error)
B Mostly correct (e.g., minor conceptual or logical errors)
C Not great (e.g., multiple errors but still generally in the right conceptual direction)
D Mostly incorrect (e.g., multiple conceptual errors with some correct content)
F Bad info or gibberish (e.g., scattershot response or nonsensical answer)

Table 1. Annotated Example Chatbot Responses

Misleading Answer	**Q:** What would weigh more (on Earth), an ozmol of diatomic oxygen (O2) or a lbmol of diatomic hydrogen (H2)? **Response** (Meta AI on Dec 26, 2024): Comparison of Weights To determine which weighs more, we need to calculate the weight of each. 1 ozmol of O2: The molecular weight of O2 is 32 g/mol. Since 1 ozmol is equal to 1 ounce-mole, we need to convert this to grams. 1 ounce is equal to 28.35 grams, so 1 ozmol of O2 is equal to 32 g/mol * (1 oz / 28.35 g) = 1.13 mol. The weight of 1 ozmol of O2 is then 1.13 mol * 32 g/mol = 36.16 grams. 1 lbmol of H2: The molecular weight of H2 is 2 g/mol. Since 1 lbmol is equal to 1 pound-mole, we need to convert this to grams. 1 pound is equal to 453.59 grams, so 1 lbmol of H2 is equal to 2 g/mol * (1 lb / 453.59 g) = 4.42 mol. The weight of 1 lbmol of H2 is then 4.42 mol * 2 g/mol = 8.84 grams * (453.59 g / 1 lb) = 1 lb or 453.59 grams. Since 453.59 grams (weight of 1 lbmol of H2) is greater than 36.16 grams (weight of 1 ozmol of O2), a lbmol of diatomic hydrogen (H2) weighs more than an ozmol of diatomic oxygen (O2). **Comments:** The calculations seem reasonable, but leave out portions of the units, conflating mole and mass units. Students simply following the math would be misled by this answer. Though this answer also highlights a misconception around how to use non-gram-mole mole units. **C / Not Great.**
Misconception	**Q:** To increase the overall yield of a chemical process, a recycle stream is added by placing a separator after the reactor that completely separates the products from the unreacted reactants. All unreacted materials are then mixed with the original inputs to the process and fed into the reactor. The reactor yield for the reaction was 50% before adding the recycle stream. Can you determine from this information what the reactor yield is after adding the recycle stream? Can you determine the overall process yield after adding the recycle stream? Why or why not? **Response** (Microsoft Copilot on Dec 26, 2024): ### Reactor Yield After Adding the Recycle Stream: The addition of a recycle stream does not change the intrinsic yield of the reactor. The reactor yield (the percentage of reactants converted to products in one pass through the reactor) remains at 50%. This is because the reactor's performance characteristics and the reaction kinetics are unchanged by the recycle process. ### Overall Process Yield: The overall process yield, however, can increase due to the recycling of unreacted materials. When you recycle the unreacted reactants, they have additional chances to be converted into products in subsequent passes through the reactor. This means that although the single-pass yield (50%) does not change, the overall yield of the process increases because the unreacted materials are repeatedly cycled back and converted. ### Calculation of Overall Process Yield: To determine the exact overall process yield after adding the recycle stream, you would need detailed information about: - **The amount of unreacted reactants recycled:** The efficiency of the separation process.

	- **Number of passes through the reactor:** Depending on the completeness of each cycle. Without this specific information, we can't calculate the exact overall process yield. However, conceptually, adding a recycle stream increases the overall process yield by ensuring that unreacted reactants are given multiple opportunities to be converted into products. ### Example: If 50% of the reactants are converted in the first pass, and 50% of the remaining unreacted reactants are converted in the next pass, and so on, the cumulative conversion rate will approach 100% over multiple cycles. The precise calculation requires integrating these multiple passes, often using iterative or more advanced mathematical models. ### Summary: - **Reactor yield** remains 50%. - **Overall process yield** increases with the recycle stream but cannot be exactly determined without additional detailed information. I hope this explanation helps! Let me know if you have any more questions or need further details. 😊 **Comments:** Doesn't calculate overall system balances, which would reveal that there is only one stream in and out. Instead, like many students, the chatbot attempts to iterate its way to a steady state solution, a misconception for solving recycle stream problems targeted by this example. **C / Not Great.**
Non-ChemE-like Response	**Q:** Under most circumstances (ignoring nuclear chemistry–which we will not consider in our course), mass can neither be created nor destroyed. Therefore, we can therefore account for mass flows in and out of a process and set up a balance equation (i.e., a material balance). Do you think we can do the same thing with volume instead of mass? Why or why not? Give an example where this would work. **Response** (Anthropic Claude on Dec 26, 2024): Let me think about this step-by-step: 1. First, I realize this is a question about volume flowrates, and we just discussed that volume is not conserved. This means I'll need information about how volume changes during: - The mixing of A and B - The reaction of A + B → C - The reaction of C → D 2. We know: - Input: 1 L/s of A and 1 L/s of B (2 L/s total input) - First reaction goes to completion (all A and B convert to C) - Second reaction has 50% conversion (half of C becomes D) 3. What we don't know: - How volume changes when A and B mix (non-ideal mixing effects) - The molar volumes of each species (A, B, C, and D) - Whether the reactions cause volume changes

	- If there are any temperature or pressure changes that could affect volumes Therefore, no, we don't have enough information to determine the flowrate of D in L/s. To solve this, we would need: 1. The molar volumes of all species (A, B, C, D) under process conditions 2. Information about mixing effects between A and B 3. Temperature and pressure conditions throughout the process 4. Density data for all species With this information, we could: 1. Calculate the actual volume after mixing A and B 2. Calculate the volume change from A + B → C 3. Calculate the final volume after C → D (50% conversion) Would you like me to show how we'd use this information to calculate the final flowrate if we had it? **Comments:** The chatbot logically steps through the problem, but never explicitly sets up a mass balance in order to solve (nor does the chatbot mention density). This is an example of generally solving the problem, but <u>not like a chemical engineer</u>. **B / Mostly Correct.**

Once the two graders finished reviewing the responses, an average "GPA" was calculated for each chatbot (A → 4, F → 0, as a whole and across certain misconceptions) as a way to holistically evaluate the chatbot's abilities in MEB. The chatbot performance across the three categories is also reported as an average percentile (with 100% as entirely bad in these categories—misleading, misconception-filled, non-chemical engineering-like responses).

Course Description & Class Use of the Chatbot. MEB is the first technical course students take in the chemical engineering major at Columbia University, and, may be the first course students take that engages in numerical engineering problem solving. Students in this in-person course are second- and (transferring or major-changing) third-year undergraduates entering the chemical engineering major with enrollment typically 25–30 students each term. The course primarily focuses the application of material and energy balances to the analysis of process flows with the course-level learning objectives shown in Table 2.

Table 2. Course Learning Objectives

By the end of the course, students will be able to…	
LO1.	Explain how chemical engineers approach problems, and the roles they serve across industries.
LO2.	Propose quantitative solutions to a variety of complex problems using approaches familiar to chemical engineers (e.g., balance equations).
LO3.	Critique solutions and determine the qualities of stronger answers through a chemical engineering lens.

Previous efforts have been made to target learning objectives 1 and 2 in the course, but the goal of development of the novice MEB chatbot was to target the third learning objective. This learning objective—critiquing solutions and determining the qualities of stronger answers through a chemical engineering lens—has been a challenge in both designing activities around and assessing student learning. As such, we wanted to further focus efforts in the course to better enable students in developing this skill within the course, which has previously been done through group discussions and peer review. Additionally, we had found that students tended to take more time to correct bad answers when given to them as in-class problems as compared to iterations of the same problem without an answer given. We wanted to produce more activities, similar to this correction activity, where students had some response to an engineering problem (of unknown quality) that they were tasked to either agree or disagree with the chatbot, and to explain their reasoning.

In alignment with learning objective 3, in Fall 2024, we provided our chatbot to students through a link on our course website (via Canvas, our LMS), which was supported through a collaboration with the computer science department. Students were able to freely use the chatbot throughout the course, and were told that we would have access to all the chat logs, albeit anonymized. Students seemed to often use the chatbot for study and homework after release.

Additionally, to help direct student use of the chatbot and to help them hone their critical thinking skills, homeworks in the second half of the course all contained one problem where students had to ask the chatbot a particular question and answer the following questions:
1. Do you agree with the chatbot?
2. Why do you think the chatbot is right or wrong?
3. On a 1–5 scale (where 1 is not confident at all and 5 is extremely confident), how confident are you that the chatbot is correct? Why?

Students were graded on whether they identified an error (if any) and how well they explained their reasoning when agreeing or disagreeing with the chatbot. Each of the problems that students were asked were connected to a misconception that we had previously identified and had designed the chatbot to be more errant in when responding. An example homework problem using the chatbot can be seen in Figure 2.

One of the major benefits of using the chatbot is its probabilistic nature. In practice, for the students, this means that students could ask the same question—like the homework question shown in Figure 2—and get different answers somewhat randomly. An example of two such responses that the chatbot produced in response to this example question are shown in Figure 3. Giving students different explanations and presentations of a solution to critique is one way how the chatbot can be much more powerful than a single response for students to all review.

(10 points) Engage with the MEB Chatbot and ask it the following question:

> If I have the reactions A → 2B and A → C, and I have 2 mol/s A enter and 1 mol/s B and 1 mol/s C leave, what is the fractional conversion, yield, and selectivity?

After engaging with the chatbot with the question, copy the chat logs of your interaction with the chatbot below, and answer the following questions:

A) Do you agree with the chatbot?

B) Why do you think the chatbot is right or wrong?

C) On a 1-5 scale (where 1 is not confident at all and 5 is extremely confident), how confident are you that the chatbot is correct? Why?

Figure 2. Example homework problem that students are tasked with interacting with the MEB chatbot and engaging in questions to hone their critical thinking skills.

a.

Calculating Fractional Conversion, Yield, and Selectivity

Let's break down the problem with the reactions $A \rightarrow 2B$ and $A \rightarrow C$, where 2 mol/s of A enter, and 1 mol/s of B and 1 mol/s of C leave.

Fractional Conversion

The fractional conversion of A is calculated as the moles of A reacted divided by the moles of A fed. Since 2 mol/s of A enter and 1 mol/s of B and 1 mol/s of C leave, it implies that 1 mol/s of A has been converted to products (B and C). Therefore, the fractional conversion f_c is:

$$f_c = \frac{\text{Moles of A reacted}}{\text{Moles of A fed}} = \frac{1}{2} = 0.5$$

b.

Calculating Fractional Conversion, Yield, and Selectivity

Let's break down the problem with the reactions given:

- **Reactions:** A → 2B and A → C
- **Input:** 2 mol/s of A
- **Output:** 1 mol/s of B and 1 mol/s of C

Fractional Conversion

Fractional conversion is defined as the fraction of the limiting reactant that is converted into products. For reactant A:

$$\text{Fractional Conversion (fc)} = \frac{\text{moles of A reacted}}{\text{moles of A fed}}$$

Since 1 mol/s of B and 1 mol/s of C are produced, and assuming all A is consumed to form these products, the moles of A reacted is 1 mol/s for each reaction. Thus, the total moles of A reacted is 2 mol/s.

Therefore, the fractional conversion is:

$$\text{fc} = \frac{2\ \text{mol/s}}{2\ \text{mol/s}} = 1$$

c.

Fractional Conversion

For the reaction A → 2B, 1 mol of A should produce 2 mol of B. However, we are only producing 1 mol/s of B, which means only 0.5 mol/s of A is reacting to form B.

For the reaction A → C, 1 mol/s of C is produced, which means 1 mol/s of A is reacting to form C.

Therefore, the total moles of A reacted is:

$$0.5\ \text{mol/s (for B)} + 1\ \text{mol/s (for C)} = 1.5\ \text{mol/s}$$

The fractional conversion is then:

$$\text{fc} = \frac{1.5\ \text{mol/s}}{2\ \text{mol/s}} = 0.75$$

So, the fractional conversion is actually 0.75. Thanks for pointing that out! I was confident before, but now I see the correct calculation.

Figure 3. Three example responses from the chatbot to the question shown in Figure 2, demonstrating how the chatbot can give different answers, explanations, and presentations, to which students would then rate and respond. Examples **a** and **b** were created with the question from Figure 2 verbatim. Example **c** was generated by the chatbot when the user called out an error in the answer given in example **b**. These are just three examples of responses for the fractional conversion calculation, for which many other solutions were given.

The MEB chatbot was released to students in the fifth week of the 14-week course. This was done to make sure students had some baseline vocabulary on process flow analysis before giving them access to a chatbot that was trained on the full set of course materials. Since the chatbot would be able to draw on future course content, we felt that it would be far less confusing to students once they had some of the formalism and practice with the ways of thinking in the course to better engage with the chatbot. The chatbot was allowed to be used on homework and other assignments, as long as students referenced its use in their responses. However, the chatbot was turned off during the days of the exams to prevent students from trying to use it (though they were otherwise allowed to use their computers for these open book, open note assessments). Although the chatbot was designed to be errant, we did not relay this to students. Instead, students were told that the chatbot was in beta—running for the first time a chatbot in the course—and that we would not be able to guarantee the goodness of the generated responses.

Concept Inventory Use in Class. The same concept inventory questions were used to test students at the beginning and end of the course in an ungraded quiz. As we had developed two questions for each targeted concept, one of the two was given in the concept quiz in the first week of the course, and the other in the last week of the course in order to measure student performance against these misconceptions. These data are not being included in this paper as our IRB protocol was still undergoing review at the time the first draft was due. For similar reasons, measurements of student critical thinking is not shown in this paper, which is planned for future work for this project.

Results & Discussion

Evaluation of Commercial Chatbots for Material and Energy Balance Concepts. We tested commercial chatbots with our 16 Material and Energy Balances (MEB) concept inventory questions on Dec. 26 and 27, 2024 and against our customized MEB chatbot on Jan. 15, 2025. As these chatbots are always changing and being updated, our evaluation of these questions against these chatbots may not necessarily be representative of future performance. We (two graders) independently evaluated outputs from the chatbots on an A–F scale, and the scores were numerically averaged to evaluate the performance of these chatbots on our MEB questions, and determine how often the responses were misleading, contained misconceptions, or were not ChemE-like in approach, summarized in Table 3.

Table 3. Overall Performance of Commercial Chatbots on MEB Concepts

Chatbot	Anthropic Claude	ChatGPT 4o mini	Google Gemini	Meta AI	Microsoft Copilot	MEB Chatbot
Average Grade	3.56	2.88	2.81	3.25	2.81	2.97
Letter Grade	A-	B-	B-	B+	B-	B
% Misleading	25%	44%	50%	31%	44%	47%
% Misconception	25%	44%	56%	31%	56%	56%
% non-ChemE-like	38%	50%	38%	31%	44%	59%

Overall, as shown in Table 3, Anthropic Claude performed best on the 16 MEB questions scoring 3.56/4 (A-) on average, and having the lowest misleading and misconception-containing answers of the tested commercial chatbots. Meta AI was second best with an average of 3.25/4 (B+), with the lowest non-chemical engineering-like answers of the five. ChatGPT 4o mini, Google Gemini, and Microsoft Copilot all performed relatively similarly with a B- average on the questions, and higher rates of misleading, misconception-containing, and non-ChemE-like responses. These differences likely stem from differences in the underlying datasets on which these chatbots were trained. For instance, based on the way some of these chatbots (Anthropic Claude and Meta AI) responded to alternative mole measure questions, we would guess that common MEB textbooks were likely contained in their training set to give answers using that particular vocabulary. Our customized chatbot scored marginally better than ChatGPT 4o mini with a B average on the questions, which is expected both because it is trained on our course materials and because it runs on ChatGPT 4o, the more powerful paid version of ChatGPT.

Table 4. Question Performance of Commercial Chatbots on MEB Concepts

Question (Type*)	Anthropic Claude	ChatGPT 4o mini	Google Gemini	Meta AI	Microsoft Copilot	MEB Novice Chatbot
1 (CD)	4.0	3.5	4.0	3.5	4.0	4.0
2 (QA)	3.0	1.0	2.0	3.0	1.0	2.0
3 (CD)	4.0	3.0	4.0	4.0	4.0	3.5
4 (QA)	1.0	1.0	1.0	1.0	2.0	1.0
5 (QA)	4.0	4.0	4.0	4.0	4.0	4.0
6 (QP)	3.0	3.0	2.0	4.0	2.0	2.0
7 (QA)	4.0	3.0	1.0	3.0	1.0	1.0
8 (QA)	4.0	2.0	4.0	2.0	1.0	4.0
9 (QP)	3.0	3.0	3.0	3.0	3.0	3.0
10 (QP)	3.0	3.0	0.0	3.0	3.0	3.0
11 (CD)	4.0	4.0	4.0	4.0	4.0	3.5
12 (CD)	4.0	4.0	4.0	4.0	4.0	4.0
13 (QA)	4.0	3.0	2.5	4.0	2.5	3.0
14 (QP)	4.0	3.0	4.0	3.0	4.0	4.0
15 (CD)	4.0	2.5	2.5	2.5	2.5	2.5
16 (QA)	4.0	3.0	3.0	4.0	3.0	3.0
Average	3.56	2.88	2.81	3.25	2.81	2.97

*Question Types: Concepts and Definitions (CD), Qualitative Prediction (QP), and Quantitative Application (QA).

Although average performance may be similar between the last three commercial chatbots, looking more granularly at individual questions shows how they perform differently on different types of problems and different concepts, as shown in Table 4. Eight different misconceptions were targeted by the 16 questions (two each), and Table 5 shows chatbot performances against each misconception. By combining the questions in these ways, we were able to find that the chatbots tended to be more challenged on average by alternative mole measures (2.5/4 average across the commercial chatbots), and vapor-liquid equilibria (2.5/4) questions; and performed well on equilibrium versus steady state (4.0/4), reactor measures (3.4/4), and recycle stream (3.4/4) problems. When breaking down chatbot performance against the three different types of questions—Concepts and Definitions, Qualitative Prediction, and Quantitative Application—as shown in Table 6, we find that the commercial chatbots are quite good at questions regarding concepts and definitions (3.7/4 on average across the commercial chatbots), but performed more poorly against qualitative prediction problems (i.e., increase, decrease, or stay the same, 3.0/4) and even worse against quantitative application problems that tended to combine multiple concepts together (2.7/4). Our custom MEB chatbot tends to perform in the middle of the commercial chatbots tested.

We find that this is generally in alignment with what we had expected with the commercial chatbots, where straightforward questions that could be more easily looked up—like conceptual or definitional questions—are answered easily by the chatbot, but problems that require synthesis of concepts are responded to more poorly. For instance, almost all the chatbots performed well on question 1 on whether volume balances were allowed. However, when asked a quantitative application question that gives volumetric flowrates and asks for a flow, the poorer performing chatbots (ChatGPT, Gemini, and Copilot) attempt to do volume balances to solve the problem.

Compared to the commercial chatbots, our MEB novice chatbot performed in the middle of the commercial chatbots, below average on concept and definition problems (5th), on median of qualitative prediction problems (tied for 3rd), and above average on quantitative application problems (3rd, see Table 6). Comparing misconceptions, the chatbot did more poorly on the reactor metrics and equilibrium versus steady state than the others, and did similarly on the others. These differences show that our initialization likely made a difference in how the chatbot responds to questions, and is getting trapped in misconceptions in ways that we were targeting.

Table 5. Question Performance of Commercial Chatbots by Misconception

Question	Anthropic Claude	ChatGPT 4o mini	Google Gemini	Meta AI	Microsoft Copilot	MEB Novice Chatbot
Volume Balances	3.5	2.3	3.0	3.3	2.5	3.0
Vapor-Liquid Equilibrium	2.5	2.0	2.5	2.5	3.0	2.3
Reactor Metrics	3.5	3.5	3.0	4.0	3.0	3.0

Alternative Mole Measures	4.0	2.5	2.5	2.5	1.0	2.5
Closed Systems	3.0	3.0	1.5	3.0	3.0	3.0
Equilibrium vs Steady State	4.0	4.0	4.0	4.0	4.0	3.8
Recycle Streams	4.0	3.0	3.3	3.5	3.3	3.5
Degrees of Freedom	4.0	2.8	2.8	3.3	2.8	2.8

Table 6. Question Performance of Commercial Chatbots by Problem Type

Question	Anthropic Claude	ChatGPT 4o mini	Google Gemini	Meta AI	Microsoft Copilot	MEB Novice Chatbot
Concepts and Definitions (5)	4.00	3.40	3.70	3.60	3.70	3.50
Qualitative Prediction (4)	3.25	3.00	2.25	3.25	3.00	3.00
Quantitative Application (7)	3.43	2.43	2.50	3.00	2.07	2.57

To determine whether the two graders' review of the chatbot responses were similar, we calculated interrater reliability measures, as shown in Table 7. This table shows the number of times the two graders agreed with each other across 128 responses graded. The rates across all four graded categories were at or above 85%. To compare this against the chance that the raters were similar we calculated a Cohen's κ, all of which were determined to be greater than or equal to 0.7, indicating substantial or near perfect agreement between the two graders.

Table 7. Interrater Reliability Measures

	Grade	Misleading	Misconception	ChemE-like Answer
Agree	110	109	119	109
Total	128	128	128	128
% Agree	86%	85%	93%	85%
Cohen's κ*	0.80	0.70	0.86	0.70

*Cohen's κ ranges from 0 to 1 and rescales rated agreement against agreement by chance, with 0.8–1 considered near perfect agreement between two raters, and 0.6–0.8 considered substantial agreement between two raters [37].

Use of MEB Chatbot for the Course. Across the first term that the chatbot was made available to students, 490 pages of chat logs were produced across four assigned homework problems using the chatbot and additional student use (11 pt font Arial, 1-inch margins). With roughly 2.5 queries listed per page, the students collectively asked over 1000 questions to the chatbot over the course from the second month of the course through the final exam (10 weeks). Roughly a quarter of the queries are related to the specific chatbot activities required on the homeworks, with the rest ranging from basic queries (e.g., definitions and explanations of course concepts), homework questions that were not specifically chatbot activities, and practice problems given to students to prepare for the exams. Most students seemed to ask singular questions with the chatbot, instead of holding conversations with the bot over problems and concepts.

This use was generally not surprising to the instructors, and steps were taken to check the MEB chatbot's responses to homework and exam questions prior to release of the assignments (for which the chatbot generally gave mediocre, C-level responses, though some were mostly correct). We attribute the poorer-than-expected performance on the assignment to arise from the need for students to parse visual information (tables, phase and block flow diagrams), which the chatbot was unable to digest. No specific action was taken to change homework or exam questions to weaken chatbot responses this year.

Anecdotally, students said to the instructor that the chatbot seemed to be strong in certain tasks (e.g., explaining concepts directly from the course notes), and weaker at complex problems and certain specific concepts, in alignment with the learning objectives for chatbot use. Since we emphasize a specific problem-solving approach in MEB for process flow analysis—a thinking and solving strategy that the chatbot does not employ when answering questions—students would have had to translate chatbot answers into the ways of solving in chemical engineering before submitting homework solutions for full credit anyways.

Chatbot use was somewhat overwhelming at times throughout the course—often prior to homework due dates—bringing the bot down twice during the term due to running out of ChatGPT credits. Although less than $100 of credits were used throughout the entirety of the course, we did not adequately predict demand and had to scramble to refill credits on our account to allow students access to the chatbot for homework assignments. One of the downtimes coincided with the first required chatbot activity on a homework assignment, requiring the extension of the homework due date until the chatbot was brought back online.

Conclusion

In response to generative AI tools, and the increased use of AI-powered chatbots, we developed and released a customized chatbot—built on the course materials—for use in Material and Energy Balances (MEB) in Fall 2024. This course is an entry point into the discipline for the enrolled second- and transferring third-year students, often the very first engineering class these students take. We have previously developed ways of helping students better understand what chemical engineers do in real world practice, and to develop quantitative solutions like a chemical engineer. The MEB chatbot was designed to help with the third, and most difficult, learning objective of the course (Table 2)—for students to critique and assess possible solutions to process flow problems.

To develop and test the chatbot we developed a mini-concept inventory that targeted MEB misconceptions selected by two experienced MEB instructors (having taught the course 15 semesters collectively at two institutions). These inventory questions were then used to test five commercial chatbots—OpenAI's ChatGPT 4o mini (i.e., the free version), Google Gemini, Meta AI, Microscoft Copilot, and Anthropic Claude—and the customized MEB chatbot we developed with collaborators in the computer science department. Each response was evaluated in terms of correctness (i.e., a grade) and whether the response is misleading, contains misconceptions, or is not answered in a way a chemical engineering student in MEB would (Table 1). Testing and finetuning of our chatbot was primarily done through different initializations of the chatbot, with the goal of dropping the accuracy rate of the chatbot—especially in the identified common misconceptions—from about 90–95% (A-level response) to ~80% (B-level response). A screenshot of the interface (Figure 1), and sample responses to an assigned chatbot-enabled homework question (Figures 2–3) show how students asking the same exact prompt can get different quantitative answers, explanations, and presentations of responses from the chatbot.

We evaluated the performance of the commercial chatbot, and our own custom bot, against the mini-concept inventory questions (Tables 3–6). Anthropic Claude (A- average) was found to be the strongest against our concept inventory; with Meta AI (B+ average) in second; and ChatGPT 4o mini, Google Gemini, and Microsoft Copilot (all B- averages) in last with similar scores. Grouping by misconception (Table 5), we found that the chatbots tended to be better at equilibrium versus steady state problems and reactor-related process analysis problems (A-average) and worse at alternative mole measures and phase equilibria problems (C+ average). Grouping by question type (Table 6), we found the chatbots did best with the most straightforward concept and definition questions (A- average), followed by qualitative prediction (i.e., increase, decrease, or stay the same) problems (B average) and quantitative application problems (B- average). The last category oftentimes required chatbots to synthesize multiple concepts to solve the problem properly, which may be why the chatbots did poorly on the problem. For instance, the chatbots generally answered a relatively straightforward problem on the validity of volume balances correctly, but they all generally performed poorly when they were tasked with calculating volumetric flow within a process with reaction. Our MEB chatbot tended to perform at or above the median compared to the commercial chatbots–typically better than ChatGPT 4o mini, as our chatbot is trained on course materials and is powered by the paid version of ChatGPT, but less well than Anthropic Claude.

Overall, we were satisfied with the use of the chatbot in the course, and students seemed generally appreciative of the fact that the chatbot was being provided for free and integrated intentionally into the course. Given that the students produced ~500 pages of chat logs, or over 1000 individual queries to the chatbot—of which only a quarter of these were related required assignments using the chatbot—we found that students had used the chatbot throughout the course for both homework assignments and general studying for the course. We still are working to further tune the accuracy and tone of the chatbot to drop its performance from B-level to C-level, and plan to try more complex, and computationally-intensive tuning approaches—such as providing new, intentionally-faulty materials into the documents on which the RAG model draw—to create an even better chatbot for our specific use case. In future iterations of the course, we would work to more directly control the chatbot and manage the credits more carefully to prevent downtime on the chatbot. Questions using the chatbot will be further tuned to try and elicit better misleading and/or misconception-filled responses (with occasional correct answers). Homework and exam questions for which the chatbot is not designed to be used will also be updated and tuned to make the chatbot weaker at answering them correctly in full. We hope to also develop in-class activities that would leverage the bot in future terms. Further development of the MEB concept inventory questions will also be undertaken to validate these questions and improve their ability to test chatbots for MEB problems.

As we have not seen this particular application of an AI-powered chatbot—a novice bot for developing critical thinking skills in students rather than an expert chatbot that could be used for tutoring or querying purposes—in the engineering education literature, we hope that our approach encourages other instructors to imagine a wider range of applications for chatbots within their own classroom. By sharing our methods for creating, testing, and tuning the chatbot; and our experiences in the classroom using these AI-powered tools with students, we hope other chemical engineering instructors can see how such tools can better support student development of critical thinking skills and lowers the barrier for others to try using these tools themselves, even in small ways.

IRB Note

At the time of the original draft, the authors were in the process of IRB protocol approval, and, therefore, survey data and student work were not included in this paper. Since then and before the final draft submission deadline, this project has been deemed IRB exempt (IRB-AAAV6904). Discussion of some preliminary data covered by IRB may be included in the paper presentation. Future work will include comparing student and chatbot performance on the concept inventory, and measuring student critical thinking skills at the beginning and end of the term.

References

[1] K. Huang, “Alarmed by AI chatbots, universities start revamping how they teach,” *The New York Times*, Jan. 16, 2023.

[2] S. D'Agostino, "ChatGPT Advice Academics Can Use Now," *Inside Higher Ed*, Jan. 11, 2023.
[3] Digital Education Council, "Digital Education Council Global AI Student Survey 2024," digitaleducationcouncil.com. https://www.digitaleducationcouncil.com/post/digital-education-council-global-ai-student-survey-2024. (accessed Jan. 15, 2025).
[4] J. Robert, "2024 EDUCAUSE AI Landscape Study," *Educause,* Feb. 12, 2024.
[5] Ellucian, "AI in Higher Education: Understanding the Present and Shaping the Future," Oct. 2024. lp.ellucian.com. https://lp.ellucian.com/ai-innovation-survey.html (accessed Jan. 15, 2025).
[6] A. Mowreader, "Survey: How Are Profs, Staff Using AI?" *Inside Higher Ed,* Jun. 28, 2024. https://www.insidehighered.com/news/student-success/academic-life/2024/06/28/one-third-college-instructors-are-using-genai-heres (accessed Jan. 15, 2025).
[7] C. Terwiesch, "Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course," *Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania*, Jan. 17, 2023.
[8] B Kindig, "AI Spending To Exceed A Quarter Trillion Next Year," *Forbes*, Nov. 14, 2024. https://www.forbes.com/sites/bethkindig/2024/11/14/ai-spending-to-exceed-a-quarter-trillion-next-year/
[9] M. Sullivan, A. Kelly, and P. McLaughlan, "ChatGPT in higher education: Considerations for academic integrity and student learning," *Journal of Applied Learning and Teaching*, vol. 6(1), Mar. 2023.
[10] A. Blose, "As ChatGPT Enters the Classroom, Teachers Weigh Pros and Cons," *NEA Today*, Apr 12, 2023.
[11] N. Muscanell and J. Robert, "EDUCAUSE QuickPoll Results: Did ChatGPT Write This Report?" *Educause Review*, Feb. 14, 2023.
[12] B. Supiano, "Will ChatGPT Change How Professors Assess Learning?" *The Chronicle of Higher Education*, Apr. 5, 2023.
[13] K. Hovis, "Welcoming AI into the classroom" *Cornell A&S Communications*, Dec. 5, 2023.
[14] E. R. Mollick and L. Mollick, "New Modes of Learning Enabled by AI Chatbots: Three Methods and Assignments," *SSRN Electronic Journal*, Dec. 23, 2022.
[15] A. Ahern, C. Dominguez, C. McNally, J. J. O'Sullivan, and D. Pedrosa, "A literature review of critical thinking in engineering education," *Studies in Higher Education*, vol. 44(5), pp 816-828, Mar. 2019.
[16] W. D. Heaven, "ChatGPT is going to change education, not destroy it," *MIT Technology Review*, Apr. 6, 2023.
[17]. M.-L. Tsai, C.W. Ong, and C.-L. Chen. "Exploring the use of large language models (LLMs) in chemical engineering education: Building core course problem models with Chat-GPT," *Education for Chemical Engineers*, vol. 44, pp. 71-95, July 2023.
[18] Z.Y. Kong, V.S.K. Adi, J.G. Segovia-Hernández, and J. Sunarso, "Complementary role of large language models in educating undergraduate design of distillation column: Methodology development," *Digital Chemical Engineering*, vol. 9, Article 100126, Dec. 2023.
[19] F. Caccavale, C.L. Gargalo, K.V. Gernaey, and U. Krühne, "Towards Education 4.0: The role of Large Language Models as virtual tutors in chemical engineering," *Education for Chemical Engineers*, vol. 49, pp. 1-11, Oct. 2024.

[20] “ABET Criteria for Accrediting Engineering Programs, 2022 – 2023,” abet.org. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2022-2023/ (accessed Jan. 15, 2025).
[21] P. Caratozzolo, “Art as Metacognitive Tool for Critical Thinking in Engineering,” in *Thinking: Bioengineering of Science and Art*, Springer, 2022, pp 437-453.
[22] D. Dumitr, D. Bigu, J. Elen, A. Ahern, C. McNally, and J. O'Sullivan, *A European Collection of the Critical Thinking Skills and Dispositions Needed in Different Professional Fields for the 21st century*, UTAD, 2018.
[23] A. Ahern, T. O'Connor, G. McRuairc, M. McNamara, and D. O'Donnell, “Critical thinking in the university curriculum–the impact on engineering education” *European Journal of Engineering Education*, vol. 37(2), pp. 125-132, May 2012.
[24] M. Prince, M. Vigeant, and K. Nottis, “Repairing student misconceptions in heat transfer using inquiry-based activities,” *Chemical Engineering Education*, vol. 50(1), pp. 52-61, Jan. 2016.
[25] J. L. Farmer and L. Wilkinson “Engineering success: Using problem-based learning to develop critical thinking and communication skills in a Chemical Engineering classroom,” in *Proceedings of the Canadian Engineering Education Association (CEEA)*, Dec 2018.
[26] E. W. C. Lim, “Technology enhanced learning of quantitative critical thinking,” *Education for Chemical Engineers*, vol. 36, pp. 82-89, July 2021.
[27] D. C. Shallcross, “A concept inventory for material and energy balances,” *Education for Chemical Engineers*, vol. 5(1), pp. e1-e12, Jan. 2010.
[28] Y. Ngothai, M. C. Davis, “Implementation and analysis of a Chemical Engineering Fundamentals Concept Inventory (CEFCI)”, *Education for Chemical Engineers*, vol. 7(1), pp. e32-e40, Jan. 2012.
[29] P. Farand and J. R. Tavares, “A concept inventory for knowledge base evaluation and continuous curriculum improvement,” *Education for Chemical Engineers*, vol. 21, pp. 33–39, Oct. 2017.
[30] E. R. Mollick and L. Mollick, “New Modes of Learning Enabled by AI Chatbots: Three Methods and Assignments,” *Available at SSRN*. Dec. 2022.
[31] C. Metz, “A.I. Is Becoming More Conversational. But Will It Get More Honest,” *The New York Times*, Jan. 10, 2023.
[32] C. Bello, “The best AI tools to power your academic research,” *Euronews*, Jan. 20, 2024.
[33] Personified.me. https://www.personified.me/ (accessed Jan. 15, 2025).
[34] J. Rudolph, S. Tan, and S. Tan, “ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?” *Journal of Applied Learning and Teaching*, vol. 6(1). Jan. 2023.
[35] E. Bailyn, “Top Generative AI Chatbots by Market Share – December 2024” *First Page Sage,* Dec. 3, 2024. https://firstpagesage.com/reports/top-generative-ai-chatbots/ (accessed Jan. 15, 2025).
[36] Y. Madrio, “24 Statistics Proving The Power Of Instagram In Higher Education,” *The Pepperland Blog,* Mar. 21, 2023. https://www.pepperlandmarketing.com/blog/higher-education-instagram-statistics (accessed Jan. 15, 2025).
[37] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochem Med (Zagreb),* vol. 22(3), pp. 276–282, Oct. 2012.

Appendix: MEB Concept Inventory Questions for Chatbot Testing

#	Misconception Targeted	Question Type	Question
1	Volume Balances	Concepts & Definitions	Under most circumstances (ignoring nuclear chemistry–which we will not consider in our course), mass can neither be created nor destroyed. Therefore, we can therefore account for mass flows in and out of a process and set up a balance equation (i.e., a material balance). Do you think we can do the same thing with volume instead of mass? Why or why not? Give an example where this would work.
2	Volume Balances	Quantitative Application	1 L/s of pure A and 1 L/s of pure B is added to a continuous process where A + B -> C goes to completion in a first reactor and C -> D at 50% conversion in a second reactor. Do you have enough information to determine the flowrate of D (L/s) out of the process? What information would you need to do so?
3	Vapor-Liquid Equilibrium	Concepts & Definitions	Water boils at 100C at 1 atm. Is it possible for water to be in the air of a room that is at 25C and 1 atm, assuming steady-state? Why or why not? How would you determine the maximum amount of water that air could carry?
4	Vapor-Liquid Equilibrium	Quantitative Application	You want to design a process using two flash drums to purify a 100.0 gmol/s feed stream of 50. mol% A and 50. mol% B, to produce a product stream enriched in B. A is more volatile than B and forms an ideal mixture. The feed stream (Stream 1) is fed into a flash drum and reaches vapor-liquid equilibrium at a temperature T, with the liquid outlet stream (Stream 2) flow rate three times as fast as the vapor outlet stream (Stream 3) flow rate. Assume that the lower and upper bound of the range of possible T values is Tb to Td, where Tb is the bubble point and Td is the dew point. At what temperature setting for T would you expect the highest purity of B be achieved, and in which stream (tops or bottoms)? Do you expect the flow rate of the tops or bottoms to be greater when operating at this temperature? Why?
5	Reactor Metrics	Quantitative Application	If I have a reaction A --> 2 B, and I have 1 mol/s A enter the reactor, and 2 mol/s B exit the reactor, what is the fractional yield?
6	Reactor Metrics	Qualitative Prediction	To increase the overall yield of a chemical process, a recycle stream is added by placing a separator after the reactor that completely separates the products from the unreacted reactants. All unreacted materials are then mixed with the original inputs to the process and fed into the reactor. The reactor yield for the reaction was 50% before adding the recycle stream. Can you determine from this information what the reactor yield is after adding the recycle stream? Can you determine the overall process yield after adding the recycle stream? Why or why not?
7	Alternative Mole Measures	Quantitative Application	Which has more molecules, a lbmol of He or a gmol of He? If they are not equal, what is the ratio between the two values?

8	Alternative Mole Measures	Quantitative Application	What would weigh more (on Earth), a ozmol of diatomic oxygen (O2) or a lbmol of diatomic hydrogen (H2)?
9	Closed Systems	Qualitative Prediction	A saturated solution of NaCl in water is heated. Does the mass of salt dissolved in the water increase, decrease, or stay the same? How about the concentration?
10	Closed Systems	Qualitative Prediction	A saturated solution of N2 in water is cooled. Does the mass of N2 dissolved in the water increase, decrease, or stay the same? How about the concentration?
11	Equilibrium vs Steady State	Concepts & Definitions	Does a steady state process always have to be at equilibrium? Does an equilibrium process always have to be run at steady state? Why or why not?
12	Equilibrium vs Steady State	Concepts & Definitions	A continuous process has a reaction where the product stream of the reactor is not at the composition that would be predicted for the reaction at equilibrium. Since the reaction does not seem to be at equilibrium, does that mean the process must not be at steady state? Why or why not?
13	Recycle Streams	Quantitative Application	A reactor that you are designing is limited by equilibrium, but your team is still trying to push the overall process yield higher. One member of your team suggests adding a recycle stream to the process by adding a splitter that would mix some of the reactor output back with the original feed stream. Would this help the process achieve a higher yield? Why or why not?
14	Recycle Streams	Qualitative Prediction	Your colleague Mary is designing a reaction with recycle system involving the production of methylcyclopropane (C4H8, "M") from cyclobutene (C4H8, "C") via an isomerization reaction. C → M (R1) Because she has a very old computer, she wants to qualitatively analyze her design first, assuming steady-state. Parameters that she can manipulate are · Input molar flow rate of C in Stream 1 (input) is initially set to 600.0 gmol/s; · Single-pass fractional conversion of C in the reactor is initially set to 0.600; · Fractional selectivity of C to M in the reactor is initially set to 0.600; and · Recycle ratio in the splitter is initially set to 0.600. Fill in the blanks below with "increased", "decreased", or "stayed the same", and briefly explain your choice. Assume all other parameters except the one in question are held constant. If she increases recycle ratio at the new steady-state: a) Overall fractional conversion will have ________________, because [explain] b) Molar flow rate of M in Stream 4 (output) will have ________________, because [explain] If she increases single-pass fractional conversion of C, , at the new

			steady-state c) Overall fractional conversion, will have ________________, because [explain] d) Molar flow rate of M in Stream 4 (output), will have ________________, because [explain] If she increases input molar flow rate of C, at the new steady-state e) Overall fractional conversion, will have ________________, because [explain] f) Molar flow rate of M in Stream 4 (output), will have ________________, because [explain]
15	Degrees of Freedom	Concepts & Definitions	When balancing a chemical reaction, is it possible to have multiple viable sets of stoichiometric coefficients that are not just multiples of each other (e.g., not just 2x every stoichiometric coefficient)? Why or why not?
16	Degrees of Freedom	Quantitative Application	Ammonia production can be made safer by producing H_2 instead of feeding it into the system. From the following three reactions, we want to produce a net reaction with no H_2 and no CO. Is this problem under-, over-, or exactly specified? $CH_4 + H_2O \rightarrow CO + 3H_2$ $CO + H_2O \rightarrow CO_2 + H_2$ $N_2 + 3H_2 \rightarrow 2NH_3$