



Volume 23 (2026), pp. 91-109
*American Journal of STEM Education:
Issues and Perspectives*
eISSN 3065-1190 | Print ISSN: 3069-0072
<https://doi.org/10.32674/ycyhab69>

Analyzing GPT-5 Performance in General Chemistry: Best Practices for Assignment Design

Emilie V. Ernst,
Albert Einstein College of Medicine, United States
<https://orcid.org/0009-0008-8299-0832>

Jennifer R. DeRosa
University of Pennsylvania, United States
<https://orcid.org/0009-0005-5164-6543>

Brian G. Ernst
Wentworth Institute of Technology, United States
<https://orcid.org/0000-0002-7900-9360>

ABSTRACT

The ability of large language models (LLMs) to outperform humans on many intellectual tasks has created a difficult landscape for assessing student work. One strategy is designing assignments to be specifically difficult for AI. However, LLMs have been shown to perform well on general chemistry problems across topics, and new models continue to make AI-proofing assignments more difficult. In this work, we evaluate the recently released GPT-5 Thinking model across a set of 1736 questions from the Openstax general chemistry textbook. While GPT-5 achieves near-perfect accuracy overall, categorizing questions by the types of cognitive skills required elucidates specific weaknesses. Notably, GPT-5 tends to make major errors when asked to create figures. These findings inform best practices for AI-resistant assignment design.

Keywords: Artificial intelligence (AI), ChatGPT, chemistry, general chemistry, higher education, large language model (LLM)

INTRODUCTION

The recent advances of Large Language Models (LLMs) and seemingly ubiquitous availability of LLM-powered chatbots to the general public over the last 4 years introduce a novel challenge for educators. Gone are the days where educators can assign a problem set to their students and have reasonable certainty that the students have applied and synthesized what they have learned in class to complete the assignment. For many traditional assignments, a student may simply prompt their favorite chatbot with the homework question as assigned and receive a comprehensive solution with adequate work shown. With modern LLMs outperforming humans on a wide range of intellectual tasks, detecting and enforcing rules prohibiting the use of these tools poses an insurmountable burden on educators to ensure that students' work is their own. The perils that modern education potentially faces if these AI-powered tools are overused by their students is threefold: (1) Students miss an opportunity to reinforce the material covered in class through repetition and practice which out-of-class assignments used to provide. (2) We fail in our crucial goal as educators if students do not achieve mastery of course material, knowledge that is often necessary for success in later courses and careers. Previously, such content mastery was ensured as a product of assignment completion. (3) Finally, students are robbed of the opportunity to develop high-level critical thinking/problem solving skills, and thus self-confidence, which come from grappling with a problem that they do not immediately know how to solve and working through it themselves to reach the solution.

In this age of widespread AI-tool availability, a divergence in educational philosophy has emerged: one camp largely sees AI as a threat to academic integrity which subverts students' intellectual development, while the other largely embraces AI as a means to drastically increase student productivity. Both of these philosophies share the opinion that AI is having a transformative impact on education; and it is the responsibility of educators to redesign courses to meet this changing academic landscape. One strategy that can appeal to both camps is designing AI-proof assignments where the questions asked are those which LLMs are known to get incorrect; those who wish to embrace AI use still want students to have a solid understanding of its limitations, and for them to learn skills that AI

cannot yet do trivially. From a constructive alignment perspective, this means not only changing the assignments to be AI-proof, but also shifting the course learning outcomes towards skills and modes of reasoning cannot be outsourced to AI (Biggs, 1996). To this end, designing AI-proof assignments in chemistry has been a difficult task to date. Developing adequate coursework in the age of AI is further complicated by an ever-moving target presented as newer versions of these tools become available. Upon the release of ChatGPT-5 in August of 2025, Sam Altman touted its capabilities as akin to “having a team of PhD-level experts in your pocket” (Genovese, 2025).

It is crucial for there to be comprehensive assessments of the capabilities and performance of these tools. Extensive research that displays the vast knowledge LLMs possess in general chemistry topics, spanning elemental properties, bonding theories, thermochemistry, and much more, has been published (*vide infra*). In contrast, in this work we focus on evaluating AI’s abilities broken down by the cognitive skills used in these problems. Note that a thorough analysis of the performance of these AI tools using a large library of problems that require a range of cognitive skills (discussed further in the Research Method section of this manuscript). Isolating cognitive skills that remain weak in the latest AI models, particularly GPT-5, outlines a new road map for assignment design in the modern era. The results and subsequent discussion presented here are a crucial step in shifting our assessed student work toward problems that AI struggles with as an effective approach, whether our goal as educators is to limit AI use entirely or have students work collaboratively with AI.

LITERATURE REVIEW

The education literature responded quickly to the widespread introduction of LLMs in late 2022, with ‘AI in education’ emerging as a prolific subfield overnight. The first articles focusing on chemistry education specifically investigated how effective LLMs were at solving chemistry problems (Clark, 2023; Clark et al., 2023; Daher et al., 2023; Emenike & Emenike, 2023; Fergus et al., 2023; Hallal et al., 2023; Leon & Vidhani, 2023; Tyson, 2023; Watts et al., 2023). Initial results were unimpressive, with ChatGPT struggling to perform calculations (Clark, 2023; Clark et al., 2023; Daher et al., 2023; Tyson, 2023), complicated tasks and deep reasoning (Daher et al., 2023; Fergus et al., 2023; Hallal et al., 2023), as well as diagrammatic and spatial reasoning (Fergus et al., 2023; Guo et al., 2023; Mirza et al., 2025; Watts et al., 2023), while performing overall better on memorization/recall type questions (Daher et al., 2023). The problem of ChatGPT hallucinating references was also noted in several articles (Tyson, 2023; West et al., 2023). Within only a year, significant progress in the chemistry capabilities of widely available LLMs were made (Alasadi & Baiz, 2024; Boiko et al., 2023; Bran et al., 2024; Kim et al., 2024; Runcie et al., 2026;

Schrier, 2024; Zhang et al., 2024). How effective LLMs are at completing various forms of academic work is a central question in education research; it determines which types of coursework can now be completed simply by querying a tool like ChatGPT, without meaningful student engagement.

Despite concerns over students using LLMs to bypass coursework, many authors saw the emergence of such AI tools with optimism. A number of articles discuss LLMs being used as private tutors (Fergus et al., 2023; Humphry & Fuller, 2023; Kılınç, 2023; Lopez-Gazpio, 2025; Ruff, Engen, et al., 2024; Ruff, Franz, and West, 2024); others suggested AI can handle routine tasks while students spend their efforts on activities thought to promote deeper learning: checking, interpreting, and refining the outputs (Guo & Lee, 2023; Humphry & Fuller, 2023; Jones et al., 2024; West et al., 2023; Yik & Dood, 2024; Yuriev et al., 2024). Others highlight additional benefits, including enabling students to employ tools like programming or data visualization - skills which may have previously been out of reach to incorporate formal training in courses where this isn't the central focus (Hasrod et al., 2024; Lear, 2024).

Evaluating the capabilities and limitations of prominent AI tools continues to be a major area of research, with identifying any weaknesses as a cornerstone (Alampara et al., 2025; Lopez-Perez et al., 2026; McLeod et al., 2026; Pereira et al., 2026; Sharlin et al., 2026; Uçar et al., 2025). Furthermore, the answer to this question is constantly evolving, with new AI models rapidly expanding the scope of coursework they can do. This work continues in our research with an investigation of the performance of the newly released GPT-5 Thinking and GPT-5.1 Thinking models on a very large (1736 question) set of general chemistry questions. We analyze how ChatGPT performs on questions categorized by the required cognitive skills. Best practices for designing assignments that cannot be trivially performed by current state-of-the-art LLMs are discussed.

RESEARCH METHOD

This study seeks to determine the ability of ChatGPT's recently launched GPT-5 Thinking model (referred to as GPT-5 subsequently, for simplicity) to correctly answer problems commonly seen in an introductory chemistry course. For this study, we test GPT-5 on a large bank of 1736 questions selected from the peer-reviewed open-source textbook "Chemistry: Atoms First 2e" published through OpenStax and the University of Connecticut and UConn Undergraduate Student Government Association. We include all of the end-of-chapter questions from chapters 1-9, covering topics standard for an introductory first-semester college general chemistry course. The chapter 1-9 titles are listed below:

1. Essential ideas
2. Atoms, molecules, and ions

3. Electronic structure and periodic properties of elements
4. Chemical bonding and molecular geometry
5. Advanced theories of bonding
6. Composition of substances and solutions
7. Stoichiometry of chemical reactions
8. Gases
9. Thermochemistry

GPT-5 was prompted individually for each exercise, with multi-part exercises presented in a single prompt. Any relevant images or information needed to answer the questions were included. Importantly, to simulate naive student use, ChatGPT was not given hints, guidance, or corrections of any kind.

Our aim in this work is to focus on determining the specific kinds of questions that GPT-5 excels at and struggles with. While one approach is to categorize by content (e.g., stoichiometry vs. thermochemistry), we instead look at which cognitive skills the questions solicit. As such, we constructed a set of tags that indicate the type of cognitive skills required in the question: calculation, explanation, creativity, creating a figure, reading a figure, or simulation. Questions could be tagged with more than one of the above cognitive skills if needed. Furthermore, some questions did not use any of the specific cognitive skills defined in this study. Listed below are each of the cognitive skill tags and what is required for a question to receive that tag, as well as an example question for each:

1. *Calculation*: The question requires a calculation beyond trivial arithmetic (ex., balancing reactions does not earn this tag). Could be symbolic or arithmetic.

Example: Chapter 2 Question 50- “The approximate minimum daily dietary requirement of the amino acid leucine, $\text{C}_6\text{H}_{13}\text{NO}_2$, is 1.1 g. What is this requirement in moles?”

2. *Explanation*: The question specifically requests an explanation. Example: Chapter 5 Question 21- “The bond energy of a C-C single bond averages $347 \text{ kJ}\cdot\text{mol}^{-1}$; that of a C-C triple bond averages $839 \text{ kJ}\cdot\text{mol}^{-1}$. Explain why the triple bond is not three times as strong as a single bond.”

3. *Creative*: The question requests open-ended creative input. Example: Chapter 1 Question 22- “Matter is everywhere around us. Make a list by name of fifteen different kinds of matter that you encounter every day. Your list should include (and label at least one example of each) the following: a solid, a liquid, a gas, an element, a compound, a homogeneous mixture, a heterogeneous mixture, and a pure substance.”

4. *Creating a Figure*: The question requires a picture or diagram to be drawn. Example: Chapter 4 Question 61- “Toothpastes containing sodium hydrogen carbonate (sodium bicarbonate) and hydrogen peroxide are widely used. Write

Lewis structures for the hydrogen carbonate ion and hydrogen peroxide molecule, with resonance forms where appropriate.”

5. *Reading a Figure:* The question contains an image which must be analyzed. Example: Chapter 9 Question 99- “Use the bond energy to calculate an approximate value of ΔH for the following reaction. Which is the more stable form of FNO_2 ?”

Figure 1

Image Associated with Chapter 9 Question 99



6. *Simulation:* Requires accessing a website with an embedded simulation. Example: Chapter 2 Question 14- “Open the Build an Atom simulation (<http://openstax.org/l/16PhetAtomBld>) (a) Drag protons, neutrons, and electrons onto the atom template to make a neutral atom of Oxygen-16 and give the isotope symbol for this atom. (b) Now add two more electrons to make an ion and give the symbol for the ion you have created.”

GPT-5’s answers were graded as correct or incorrect, where answers that might otherwise be considered partially correct were labeled incorrect. Each section of a multi-part exercise was separately graded and tagged. Additionally, questions that GPT-5 gave incorrect answers for were resubmitted to the newly released GPT-5.1 Thinking model (subsequently referred to as GPT-5.1 for simplicity). Note that since we only submitted questions GPT-5 got incorrect to GPT-5.1, this is not an independent evaluation of the performance of GPT-5.1. We refer to this method of first using GPT-5, and then GPT-5.1 for any question GPT-5 got incorrect as +GPT-5.1 throughout the text.

RESULTS

The number of questions GPT-5 answered correctly is shown as a raw number and a percentage in Table 1. Percent correct scores are tabulated by cognitive skill tag as well. An additional percentage column appears on the right-hand side of the table (noted as +GPT-5.1) to indicate the revised score when allowing for the correction of answers by resubmitting to GPT-5.1, as described in the Research Method section. The performance for the entire set of questions is reported at the bottom of the table. The percentages reported in the table for GPT-5 performance and +GPT-5.1 revisions are visualized as blue and orange horizontal bars in Figure

2, respectively. Overall, GPT-5/+GPT-5.1 scored 93.4%/94.5%, demonstrating a mastery level of the material generally.

Figure 2
Performance of ChatGPT on Questions by Intellectual Tag with 95% Confidence Intervals

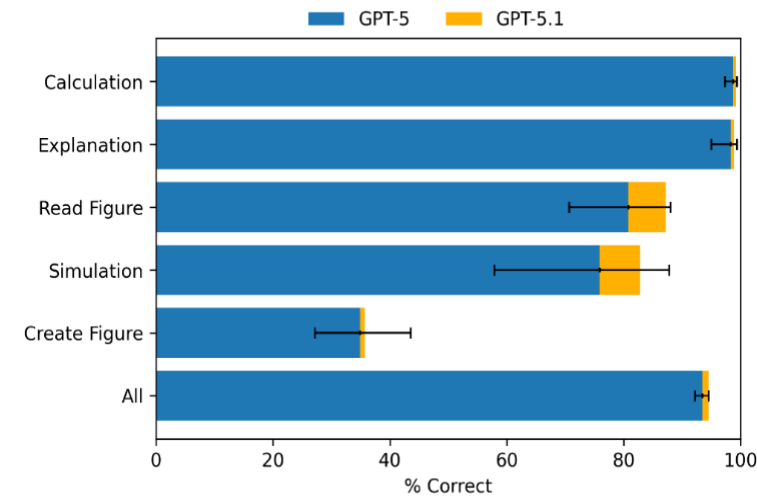


Table 1
GPT-5 and GPT-5.1 Performance by Cognitive Skill Tag

Cognitive Skill Tag	Number of Questions	GPT-5 Number Correct	GPT-5 % Correct	+GPT-5.1 % Correct
Calculation	534	527	98.7 (CI=97.3-99.4)	99.3
Explanation	171	168	98.2 (CI=95.0-99.5)	98.8
Read Figure	78	63	80.8 (CI=70.7-88.0)	87.2
Simulation	29	22	75.9 (CI=57.9-87.8)	82.8
Create Figure	126	44	34.9 (CI=27.2-43.6)	35.7
All Questions	1736	1622	93.4 (CI=92.2-94.5)	94.5

Note: CI represents the 95% confidence interval (Brown et al., 2001; GraphPad Software, 2025; Wilson, 1927)

As shown in the first two bars in Figure 2, GPT-5/+GPT-5.1 excels at providing correct solutions for questions that require calculations (98.7%/99.3%) and questions that require explanations/reasoning (98.2%/98.8%). The sample size for questions that are calculation-based (534, $\approx$ 31%) and explanation-based (171,

$\approx 10\%$) in this dataset are sufficiently large for broader conclusions of performance for AI-tools (see Discussion section).

However, GPT-5's performance on questions with the remaining cognitive skill tags decreases significantly. Its performance for questions tagged reading a figure (80.8%), interacting with a simulation (75.9%), and creating a figure/plot (34.9%) are shown in the 3rd through 5th bars in Figure 2. While resubmitting incorrect answers to GPT-5.1 shows improvement for questions requiring reading a figure (increased by 6.4%) and interacting with a simulation (increased by 6.9%), the performance remains below 90% for both categories. We note that in questions where GPT-5 is asked to generate a figure, the result is inaccurate the majority of the time, performing well below other cognitive categories. Furthermore, when resubmitting incorrect answers to GPT-5.1, the accuracy only increases by a meager 0.8%.

Our results presented above reveal a crucial benchmark for the most recent versions of ChatGPT accessible to students. In the following section we discuss specific cases where the AI-tool struggles and broader implications for designing homework problems in the modern age. It is important to note, however, that AI-tools have rapidly improved since their debut only 3 years ago, and there will certainly be further improvements to the current AI-tools, including future versions of ChatGPT. Inevitably, if our goal is to build curricula suitable for the contemporary educational landscape, we are presented with an ever-moving target.

DISCUSSION & CONCLUSIONS

In this study, we classified general chemistry questions into distinct cognitive skill categories and examined how GPT-5 (or GPT-5.1 for questions GPT-5 provided incorrect solutions to) performed across these different task types. The overall performance, as shown by the last bar in the chart of Figure 2 is exemplary, earning a score of 93.4% from the entire dataset of questions. These results indicate that a student who entirely uses solutions from GPT-5 would receive appreciable credit for any homework assignments built from a random sample from this set of chemistry questions. While creating homework assignments that cannot be trivially outsourced to common AI tools remains possible, doing so requires careful attention to the types of questions included. In this vein, we analyze ChatGPT's performance on general chemistry questions categorized by the distinct cognitive skills they require.

Calculation

There were only a few questions with the calculation tag that ChatGPT got incorrect. Among these, mistakes were not related to calculation or arithmetic error - they resulted from small conceptual mistakes or ChatGPT's inability to interact

with online simulations (vide infra). We note that this substantiates previous work that showed current LLMs have great accuracy for these types of questions (Alasadi & Baiz, 2024; Boiko et al., 2023; Bran et al., 2024; Kim et al., 2024; Schrier, 2024; Zhang et al., 2024).

Simulation

The textbook has problems that require the user to interact with PhET simulations, which ChatGPT cannot do. One example is Chapter 1 Question 42, which reads “Visit this density simulation (<http://openstax.org/l/16phetmasvolden>) and click the ‘reset’ button to ensure all simulator parameters are at their default values. (a) Use the water displacement approach to measure the mass and volume of the red block. (b) Use the measured mass and volume data from step (a) to calculate the density of the red block. (c) Use the vertical green slide control to adjust the fluid density to values well above, then well below, and finally nearly equal to the density of the red block, reporting your observations.” In part (a), GPT-5 identifies the red block as weighing 5kg and a volume of 1.25L, neither of which are correct. Those accessing the simulation could identify that the block weighs 2.5g and has a volume of 4.1mL. GPT-5 links to a URL that discusses a PhET simulation, where it describes a red block that has $m = 5\text{kg}$, $V = 1.25\text{L}$, indicating where GPT-5 is getting these numbers. In part (b), GPT-5 divides the mass by the volume from part (a) to get a density value of 4kg/L for the red block, which would be correct if those were the correct mass and volume values. In part (c), GPT-5 is able to correctly identify what would happen to the block if the fluid density was changed since that question is generalizable to situations not contained in the simulation. When this exercise was rerun through GPT-5.1, we saw very interesting improvements. In part (a), GPT-5.1 now correctly identifies the red block as weighing 2.5g, with a volume of 4.1 mL. GPT-5.1 now links to a different URL which has the exercise from the OpenStax textbook as well as the solution. Parts (b) and (c) are then done correctly (trivial tasks for ChatGPT). This presents an interesting situation: ChatGPT has no way to interact with the online simulation, however it is very much capable of searching the internet for any discussion as to its contents. There are some interesting implications here, since ChatGPT is pulling information directly from discussions of content it has not actually engaged with.

Reading a Figure

Reading figures is something ChatGPT struggles with somewhat - getting only 80.8% correct (87.2% when corrected by GPT-5.1). One question which was incorrect in GPT-5, but corrected in GPT-5.1 was Chapter 5 Question 18, which reads “Identify the hybridization of each carbon atom in the following molecule.

(The arrangement of atoms is given; you need to determine how many bonds connect each pair of atoms.).”

Figure 3

Image Associated with Chapter 5 Question 18



GPT-5 said that moving left to right along the chain, the formula implied is $\text{CH}_3\text{-CH}_2\text{-C}\equiv\text{C-CH}_3$, which is not the correct structure, particularly as it only has 5 carbons, rather than the 6 shown in the picture. It gives hybridizations for each carbon as sp^3 , sp^3 , sp , sp , and sp^3 , respectively, which would be correct for the structure it gives.

GPT-5.1 said the following:

We use the number of σ -bonding electron groups around each C (single, double, or triple bonds each count as one group) to get the hybridization. From the H pattern you can deduce the bonding pattern of the C chain as $\text{H}_3\text{C-CH}_2\text{-C}\equiv\text{C-CH=CH}_2$

Numbering carbons from left to right:

C_1 in $\text{CH}_3\text{-}$: 4 single bonds $\rightarrow$ 4 electron groups $\rightarrow \text{sp}^3$

C_2 in $\text{-CH}_2\text{-}$: 4 single bonds $\rightarrow \text{sp}^3$

C_3 in $\text{-C}\equiv\text{C-}$: one single bond + one triple bond $\rightarrow$ 2 electron groups $\rightarrow \text{sp}$

C_4 in $\text{-C}\equiv\text{C-}$: one triple bond + one single bond $\rightarrow \text{sp}$

C_5 in -CH= : one double bond + two single bonds $\rightarrow$ 3 electron groups $\rightarrow \text{sp}^2$

C_6 in =CH_2 : one double bond + two single bonds $\rightarrow \text{sp}^2$

So the hybridizations (left $\rightarrow$ right) are:

C_1 : sp^3 , C_2 : sp^3 , C_3 : sp , C_4 : sp , C_5 : sp^2 , C_6 : sp^2 .

GPT-5.1 now accurately solves both pieces of this exercise. ChatGPT does not, in either case, link to a reference, as it does in the simulation example, implying that between GPT-5 and GPT-5.1, the model did, in fact, become more readily able to identify pieces of an image.

Creating a Figure

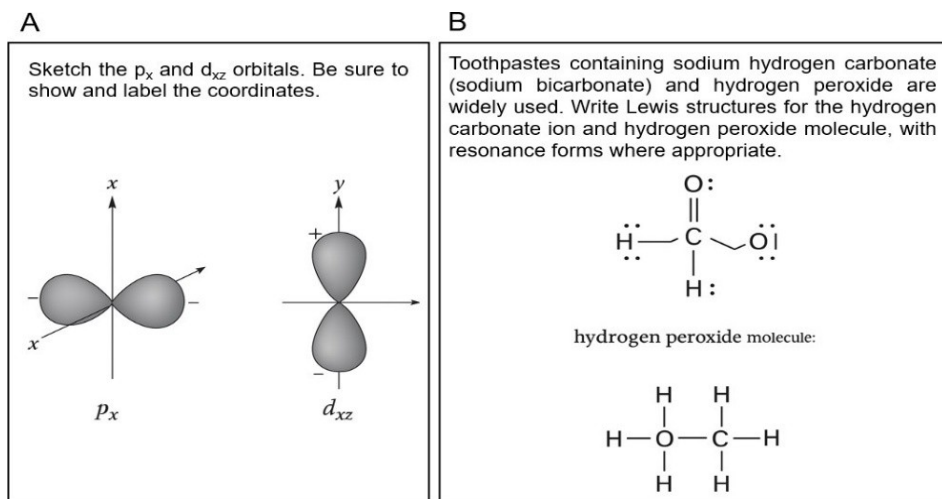
While AI has independently been able to perform complicated reasoning tasks and generate high-quality images, creating scientific figures - a task which draws heavily on both skills - is something ChatGPT struggles with. We saw both

GPT-5 and GPT-5.1 struggle with drawing accurate Lewis structures, orbital drawings, and graphs. To illustrate the types of fail points made when constructing figures, we have reproduced two incorrect solutions provided by GPT-5.1 in Figure 4, which are representative of the types of mistakes it makes. The prompt in Figure 4A, for instance, asks for a sketch of two different atomic orbitals, p_x and d_{xz} . The drawing GPT-5.1 produces for a p_x orbital gets the shape of the orbital correct but fails to properly label the axes. Egregiously, the incorrect solution labels two of the axes 'x' without providing an axis label for the third. Note also that there is no indication of a different phase between the two lobes of the orbit in the drawing. Meanwhile, the attempted drawing GPT-5.1 produces for a d_{xz} orbital fails to get even the shape of a d_{xz} orbital correct, provides only two axes, and only labels one of them. In contrast to the incorrect drawing of the p_x orbital, this attempted d_{xz} orbital drawing includes a relative phase of the lobes, though the overall solution is still incorrect, as it does not include the correct number of lobes and the reasons listed above. The question in Figure 4A directly tests crucial material that a student of general chemistry would be expected to master as part of their studies, and so the poor performance exhibited by GPT-5.0/GPT-5.1 is a clear indication that AI tools, at least in their current form, cannot reliably perform as well as a proficient general chemistry student. The fact that the responses from ChatGPT are not paired with any indication of the confidence interval that the solution is correct implies that unsuspecting students using ChatGPT may fall into a trap of blindly taking results like Figure 4A as true and reinforcing AI hallucinations as their working general chemistry knowledge.

When asked to draw Lewis structures of a hydrogen carbonate ion and a hydrogen peroxide molecule as shown in Figure 4B, GPT-5.1 produces a wholly incoherent molecule for the former and a different molecule entirely for the latter. These incorrect drawings also neglect to draw any resonance structures for either molecule. This Lewis structure drawing of a hydrogen carbonate ion has numerous problems: the bonds between carbon and hydrogen and between carbon and oxygen are bent, one hydrogen has two lone pairs of electrons, another hydrogen has a single lone pair of electrons (hydrogen should never get lone pair electrons), the oxygen that is double bonded only has one lone pair of electrons, carbon has five bonds, and the singly bonded oxygen has two lone pairs with what appears to be a bond drawn between the lone pairs. The overall atomic makeup of the molecule drawn is different than that of a hydrogen carbonate ion (HCO_3^-) as well. Meanwhile, when attempting to draw the Lewis structure of hydrogen peroxide, the response GPT-5.1 gives the Lewis structure of a non-physical molecule containing a quadruply bonded oxygen instead. We note in passing that accurate drawings of atomic orbitals and Lewis structures for simple molecules are widely available on the internet, obtaining correct answers to these specific types of question from non-AI sources is straightforward.

Figure 4

Exemplary Cases of Incorrect Scientific Figures GPT-5.1 Creates



The errors presented in Figure 4 and the poor overall performance of GPT-5/+GPT-5.1 (34.9%/35.7%) reveal a critical deficiency. Taken alongside the exemplary performance for tasks requiring an explanation (98.2%/98.8%) opens a window of opportunity for educators to consider when designing “AI-proof” assignments. We posit that questions formulated to explain course concepts using diagrams and graphs will encourage students to grapple with questions and formulate their own responses as those questions focus on cognitive skills AI-tools generally lack. In the section below we present another response from GPT-5.1 that substantiates this hypothesis.

Insubordination and Contradiction

In the previous section, we discussed GPT-5 and GPT-5.1’s poor performance when generating figures in response to a prompt. Here we present a case where GPT-5.1 receives a multi-part prompt asking for both an explanation and diagrams. Interestingly, even though the initial prompt for both GPT-5 and GPT-5.1 includes the phrase “Explain using an appropriate diagram” for two parts of the question, both versions of GPT included correct explanations of what the diagram would show in words, yet both failed to actually generate a diagram for either part. We provided a follow-up prompt asking for the diagrams and found both GPT-5.0/5.1 responses puzzling. In the case of GPT-5’s response, the response to the follow-up prompt included the two diagrams; however, one of the two diagrams was incorrect and, in fact, contradicted its correct explanation of what the figure should look like in response to the initial prompt. When we re-

prompted GPT-5.1 to produce diagrams when failing to do so to the initial prompt, the response only included a single diagram (see Figure 5) to a part of the question that did not request a diagram be drawn. The original part of the prompt that corresponds to Figure 5 asks: “In general, under what temperature conditions would you expect Z to have the largest deviations from the Z for an ideal gas?” GPT-5.1 correctly explains:

Deviations are largest when intermolecular interactions matter most:

Low temperature: molecules move slowly; attractions have more time and opportunity to pull them together → stronger deviations (big dip in Z , possible liquefaction).

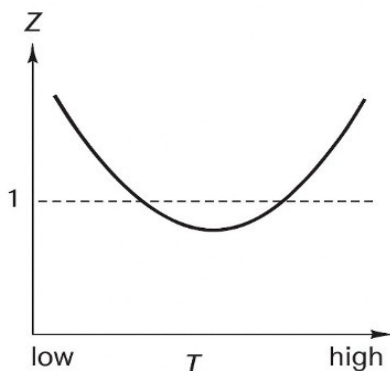
High temperature: molecules move fast; attractions are relatively less important → behavior is more ideal, $Z \rightarrow 1$.

So: Z deviates most from 1 at low temperatures (especially near the gas’ liquefaction/critical region) and becomes more ideal-like at high temperatures.

However, the diagram GPT-5.1 produced in Figure 5 appears to be completely detached from this explanation. The diagram shows a parabola with equal deviations at high and low temperatures where $Z > 1$ and the bottom of the parabola dipping below $Z = 1$. Awareness of this specific limitation of ChatGPT presents a useful case where students can work collaboratively with AI: the student ultimately needs to be the one creating the drawings, but they can receive guidance from ChatGPT in text form. We note in passing that we’ve encountered similar results elsewhere, e.g., for Lewis Structures.

Figure 5

GPT-5.1 Generated Plot of Compressibility Factor, Z vs. T



Our findings demonstrate a clear roadmap to an effective assignment design strategy requiring students to engage with course material that leverages the limitations discovered in this work. It is important to note that there are emerging

AI tools specifically designed for drawing scientific diagrams and chemical structures, which may perform better than general purpose language models like ChatGPT. Considering the rapid progress made in the AI/LLM chatbot space since the introduction of ChatGPT in 2022, one may reasonably expect models in the near future to potentially resolve the deficiencies we have discussed in this work. Nevertheless, it is crucial that educators are aware of the current capabilities of these tools as it has unequivocally changed the landscape of modern education.

Limitations and Future Directions

During the course of developing this manuscript using GPT-5, GPT-5.1 was released. This emphasizes how assessing AI's capabilities is a moving target. The release of another new model calls back into question whether there are still weaknesses in the abilities of AI as a whole. Therefore, our subsequent analysis with GPT-5.1 focused only on verifying that skill gaps were still present in the newer model. Therefore, only questions which GPT-5 answered incorrectly were tested using GPT-5.1. This does, however, have the limitation that we did not fully assess GPT-5.1's abilities and therefore cannot compare GPT-5.1's performance to GPT-5's. We also did not look at other LLMs to assess their performance on these questions. We suggest that testing their ability to correctly answer only the questions GPT-5 answered incorrectly would give insightful information about whether they have similar limitations as GPT-5. As new models continue to emerge, this presents a quick way to assess their abilities.

Conclusion and Implications

This work is motivated by an interest in creating assignments that cannot be trivially completed using widely available AI tools. Whether we as educators want our students to work independently or collaboratively with AI, our assessments should help students develop skill sets that current AI systems do not possess. However, the set of tasks that AI cannot perform reliably is shrinking rapidly. Each new model release often eliminates categories of difficulty that were previously out of reach. As such, evaluating the performance of new AI models across cognitively diverse types of tasks remains an urgent and ongoing research need.

In this work, we evaluate the recently released GPT-5 Thinking and GPT-5.1 Thinking models' ability to answer a large set of college-level general chemistry questions. Despite very good performance overall (94.5% correct), the models continue to show difficulty with tasks that require interpreting scientific figures (87.2% correct), interacting with online simulations (82.8% correct), and particularly creating scientific figures (35.7% correct). These limitations are expected given the underlying architecture: despite recent multimodal extensions,

ChatGPT is still a language- based system at its core; figure interpretation and generation are at the moment unreliable.

These results suggest that instructors should deliberately diversify the cognitive skills their assignments require. Emphasis should be placed on tasks that require students to manipulate visual information or generate scientific representations, which remain meaningful indicators of student learning. Incorporating such tasks into homework, laboratories, and examinations also aligns well with a long-standing push to have students use higher-order cognitive skills, as identified by Bloom et al., including evaluation, synthesis, and analysis (Bloom et al., 1956).

This will be important as introductory chemistry is important for understanding more advanced chemistry courses. We need to ensure students learn basic chemistry through problem sets before reaching more complex problems that ChatGPT may be unable to handle. This is especially true as organic chemistry, the next step in chemistry education, is less focused on cognitive skills ChatGPT excels at, including calculations and explanations, and more focused on skills it has difficulty with, including drawing figures.

ACKNOWLEDGEMENTS

The coauthor Jennifer R. DeRosa was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2236662. The views expressed are those of the authors and do not reflect the official guidance or position of the National Science Foundation.

REFERENCES

- Alampara, N., Schilling-Wilhelmi, M., Ríos-García, M., Mandal, I., Khetarpal, P., Grover, H. S., Krishnan, N. A., & Jablonka, K. M. (2025). Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 5, 952–961. <https://doi.org/10.1038/s43588-025-00836-3>
- Alasadi, E. A., & Baiz, C. R. (2024). Multimodal generative artificial intelligence tackles visual problems in chemistry. *Journal of Chemical Education*, 101(7), 2716–2729. <https://doi.org/10.1021/acs.jchemed.4c00138>
- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32, 347–364. <https://doi.org/10.1007/BF00138871>
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of*

- educational goals. *Handbook 1: Cognitive domain* (pp. 1103–1133). Longman.
- Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*, 624(7992), 570–578. <https://doi.org/10.1038/s41586-023-06792-0>
- Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2024). Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6, 525–535. <https://doi.org/10.1038/s42256-024-00832-8>
- Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2), 101–133. <https://doi.org/10.1214/ss/1009213286>
- Clark, T. M. (2023). Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *Journal of Chemical Education*, 100(5), 1905–1916. <https://doi.org/10.1021/acs.jchemed.3c00027>
- Clark, T. M., Anderson, E., Dickson-Karn, N. M., Soltanirad, C., & Tafini, N. (2023). Comparing the performance of college chemistry students with ChatGPT for calculations involving acids and bases. *Journal of Chemical Education*, 100(10), 3934–3944. <https://doi.org/10.1021/acs.jchemed.3c00500>
- Daher, W., Diab, H., & Rayan, A. (2023). Artificial intelligence generative tools and conceptual knowledge in problem solving in chemistry. *Information*, 14(7), 409. <https://doi.org/10.3390/info14070409>
- Emenike, M. E., & Emenike, B. U. (2023). Was this title generated by ChatGPT? considerations for artificial intelligence text-generation software programs for chemists and chemistry educators. *Journal of Chemical Education*, 100(4), 1413–1418. <https://doi.org/10.1021/acs.jchemed.3c00063>
- Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education*, 100(4), 1672–1675. <https://doi.org/10.1021/acs.jchemed.3c00087>
- Genovese, D. (2025, August 7). OpenAI GPT-5 is here: Company unveils 'PhD-level experts in your pocket'. *Fox Business*. <https://www.foxbusiness.com/technology/openai-gpt-5-here-company-unveils-phd-level-experts-your-pocket>
- GraphPad Software. (2025). *GraphPad Prism (Version 10.5.0)* [Computer software].
- Guo, T., Guo, K., Nan, B., Liang, Z., Guo, Z., Chawla, N. V., Wiest, O., & Zhang, X. (2023). What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. *NeurIPS '23: Proceedings of the*

- 37th International Conference on Neural Information Processing Systems, 59662 – 59688. <https://doi.org/10.52202/075280-2607>
- Guo, Y., & Lee, D. (2023). Leveraging ChatGPT for enhancing critical thinking skills. *Journal of Chemical Education*, 100(12), 4876–4883. <https://doi.org/10.1021/acs.jchemed.3c00505>
- Hallal, K., Hamdan, R., & Tlais, S. (2023). Exploring the potential of AI-chatbots in organic chemistry: An assessment of ChatGPT and Bard. *Computers and Education: Artificial Intelligence*, 5, 100170. <https://doi.org/10.1016/j.caeai.2023.100170>
- Hasrod, T., Nuapia, Y. B., & Tutu, H. (2024). ChatGPT helped me build a chemistry app, and here's how you can make one also. *Journal of Chemical Education*, 101(2), 653–660. <https://doi.org/10.1021/acs.jchemed.3c01170>
- Humphry, T., & Fuller, A. L. (2023). Potential ChatGPT use in undergraduate chemistry laboratories. *Journal of Chemical Education*, 100(4), 1434–1436. <https://doi.org/10.1021/acs.jchemed.3c00006>
- Jones, R. M., Bryan, E. M., & Preston, C. (2024). Exploring the concept of valence and the nature of science via generative artificial intelligence and general chemistry textbooks. *Journal of Chemical Education*, 101(8), 3276–3283. <https://doi.org/10.1021/acs.jchemed.4c00271>
- Kim, S., Jung, Y., & Schrier, J. (2024). Large language models for inorganic synthesis predictions. *Journal of the American Chemical Society*, 146(29), 19654–19659. <https://doi.org/10.1021/jacs.4c05840>
- Kılınç, S. (2023). Embracing the future of distance science education: Opportunities and challenges of ChatGPT integration. *Asian Journal of Distance Education*, 18(1), 205–237. <https://doi.org/10.5281/zenodo.7857396>
- Lear, B. J. (2024). Using ChatGPT-4 to teach the design of data visualizations. *Journal of Chemical Education*, 101(7), 2749–2756. <https://doi.org/10.1021/acs.jchemed.4c00228>
- Leon, A. J., & Vidhani, D. (2023). ChatGPT needs a chemistry tutor too. *Journal of Chemical Education*, 100(10), 3859–3865. <https://doi.org/10.1021/acs.jchemed.3c00288>
- Lopez-Gazpio, I. (2025). Integrating large language models into accessible and inclusive education: Access democratization and individualized learning enhancement supported by generative artificial intelligence. *Information*, 16(6), 473. <https://doi.org/10.3390/info16060473>
- Lopez-Perez, K., Benjamin, S. A., Veige, M. K. (2026). Outsmarting ChatGPT in Chemistry: Using Student-Formulated Questions for Critical AI Evaluation. *Journal of Chemical Education*. <https://doi.org/10.1021/acs.jchemed.5c01611>
- McLeod, G. M., Hanks, B., Ashby, H., Nolder, J., Sorenson, B., & Hanson, K. (2026). Generative AI vs Student Performance on Multiple-Choice, General

- Chemistry Exams: Insights into Question Design from Regression Analysis. *Journal of Chemical Education*, 103(2), 872–878.
<https://doi.org/10.1021/acs.jchemed.5c00979>
- Mirza, A., Alampara, N., Kunchapu, S., Ríos-García, M., Emoekabu, B., Krishnan, A., Gupta, T., Schilling-Wilhelmi, M., Okereke, M., Aneesh, A., Asgari, M., Eberhardt, J., Elahi, A. M., Elbeheiry, H. M., Gil, M. V., Glaubitz, C., Greiner, M., Holick, C. T., Hoffmann, T., Ibrahim, A., ... Jablonka, K. M. (2025). A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*. 17, 1027–1034.
<https://doi.org/10.1038/s41557-025-01815-x>
- Pereira, N. C. P., Bedouch, M. B., & Borges, E. M. (2026). Leveraging Microsoft Copilot (GPT-5) for Calculations and Interactive Data Visualization. *Digital*, 6(1), 16. <https://doi.org/10.3390/digital6010016>
- Ruff, E. F., Engen, M. A., Franz, J. L., Mauser, J. F., West, J. K., & Zemke, J. M. O. (2024). ChatGPT writing assistance and evaluation assignments across the chemistry curriculum. *Journal of Chemical Education*. 101(6), 2483–2492. <https://doi.org/10.1021/acs.jchemed.4c00248>
- Ruff, E. F., Franz, J. L., & West, J. K. (2024). Using ChatGPT for method development and green chemistry education in upper-level laboratory courses. *Journal of Chemical Education*, 101(8), 3224–3232.
<https://doi.org/10.1021/acs.jchemed.4c00193>
- Runcie N. T., Deane C. M., & Imrie F. (2026). Assessing the Chemical Intelligence of Large Language Models. *Journal of Chemical Information and Modeling*, 66(1), 216–227. <https://doi.org/10.1021/acs.jcim.5c02145>
- Schrier, J. (2024). Comment on “comparing the performance of college chemistry students with ChatGPT for calculations involving acids and bases”. *Journal of Chemical Education*, 101(5), 1782–1784.
<https://doi.org/10.1021/acs.jchemed.4c00058>
- Sharlin, S., Agbere, F., Ishimwe, K., Osifová, Z., Socha, O., Dračinský, M., & Josephson, T. R. (2026). NMR-challenge for LLMs: Evaluating chemical reasoning in humans and AI. *Machine Learning and Deep Learning*, 66(11), 6299–6312. <https://doi.org/10.1021/acs.jcim.6c00481>
- Tyson, J. (2023). Shortcomings of ChatGPT. *Journal of Chemical Education*, 100(8), 3098–3101. <https://doi.org/10.1021/acs.jchemed.3c00361>
- Uçar, S. Ş., Lopez-Gazpio, I., & Lopez-Gazpio, J. (2025). Evaluating and challenging the reasoning capabilities of generative artificial intelligence for technology-assisted chemistry education. *Education and Information Technologies*, 30, 11463–11482. <https://doi.org/10.1007/s10639-024-13295-6>
- Watts, F. M., Dood, A. J., Shultz, G. V., & Rodriguez, J.-M. G. (2023). Comparing student and generative artificial intelligence chatbot responses to

- organic chemistry writing-to-learn assignments. *Journal of Chemical Education*, 100(10), 3806–3817.
<https://doi.org/10.1021/acs.jchemed.3c00664>
- West, J. K., Franz, J. L., Hein, S. M., Leverentz-Culp, H. R., Mauser, J. F., Ruff, E. F., & Zemke, J. M. (2023). An analysis of AI-generated laboratory reports across the chemistry curriculum and student perceptions of ChatGPT. *Journal of Chemical Education*, 100(11), 4351–4359.
<https://doi.org/10.1021/acs.jchemed.3c00581>
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. <https://doi.org/10.2307/2276774>
- Yik, B. J., & Dood, A. J. (2024). ChatGPT convincingly explains organic chemistry reaction mechanisms slightly inaccurately with high levels of explanation sophistication. *Journal of Chemical Education*, 101(5), 1836–1846. <https://doi.org/10.1021/acs.jchemed.4c00235>
- Yuriev, E., Wink, D. J., & Holme, T. A. (2024). The dawn of generative artificial intelligence in chemistry education. *Journal of Chemical Education*, 101(8), 2957–2959. <https://doi.org/10.1021/acs.jchemed.4c00836>
- Zhang, D., Liu, W., Tan, Q., Chen, J., Yan, H., Yan, Y., Li, J., Huang, W., Yue, X., Ouyang, W., Zhou, D., Zhang, S., Mao, S., Zhong, H. S., & Li, Y. (2024). ChemLLM: A chemical large language model. *arXiv*.
<https://doi.org/10.48550/arXiv.2402.06852>
-

Bios

EMILIE V. ERNST, Master of Biomedical Sciences, Albert Einstein College of Medicine, New York City, United States. Email: emilie.v.ernst@gmail.com

JENNIFER R. DEROSA, Doctor of Chemistry, University of Pennsylvania, Philadelphia, United States. Email: jrderosa97@gmail.com

BRIAN G. ERNST, Doctor of Chemistry, Wentworth Institute of Technology, Boston, United States. Email: bge4043@gmail.com