

Article

Optimizing Human–AI Collaboration in Chemistry: A Case Study on Enhancing Generative AI Responses through Prompt Engineering

Dinesh V. Vidhani ^{1,*}  and Manoharan Mariappan ²
¹ Department of Math & Natural Science, Miami Dade College, 627 SW 27th Ave., Miami, FL 33135, USA

² Department of Natural Science, North Florida College, 325 Turner Davis Dr, Madison, FL 32340, USA; mariappanm@nfc.edu

* Correspondence: dvidhani@mdc.edu

Abstract: “Are we asking the right questions?” seems cliché, but for ChatGPT, it is a pivotal tool to ensure the accuracy of responses. While ChatGPT-3.5’s training on the vast database promises to revolutionize STEM education and research, this investigation shows the importance of precise communication and prompt engineering in guiding ChatGPT-3.5 toward reliable and accurate responses, particularly in chemistry. For instance, emphasizing context, clearly defining symbols, and focusing on field-specific instructions can dramatically improve its performance. Furthermore, avoiding open-ended prompts and strategically using repetition can further enhance its accuracy. The iterative prompt design, demonstrated through a series of adjustments, illustrates how seemingly minor refinements, such as substituting “least” for “lowest”, profoundly impact the output. This study highlights the essential role of human oversight, including the construction of well-crafted prompts, in guarding reliable information and nurturing a productive “Human–AI” (HAI) partnership.

Keywords: artificial intelligence; generative-AI; ChatGPT; chemical education; prompt engineering



Citation: Vidhani, D.V.; Mariappan, M. Optimizing Human–AI Collaboration in Chemistry: A Case Study on Enhancing Generative AI Responses through Prompt Engineering. *Chemistry* **2024**, *6*, 723–737. <https://doi.org/10.3390/chemistry6040043>

Academic Editors: Felix Plasser and Ryan C. Fortenberry

Received: 22 June 2024

Revised: 6 August 2024

Accepted: 8 August 2024

Published: 11 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The year 2022 is poised to mark the inception of the fourth industrial revolution with the widespread accessibility of sophisticated AI, ChatGPT-3.5, to the general public [1–3]. This powerful language model, capable of human-like text generation and engaging in meaningful conversation across diverse fields, has captured significant attention within STEM and healthcare disciplines. From medical applications, ranging from patient triage [4] to clinical research [5], to materials chemistry [6], ChatGPT-3.5’s potential is being actively explored across all key technology sectors. For example, the Yaghi group successfully employed strategies of ChemPrompt Engineering to fine-tune ChatGPT-3.5 and extract, classify, and summarize synthesis data related to metal–organic frameworks (MOFs) with a high degree of accuracy (90–99%) [7]. Furthermore, ChatGPT-3.5’s ability to analyze and interpret vast datasets, identify hidden patterns, and predict future trends in areas like epidemiology [8] and climate science [9] generates considerable interest within scientific research communities.

ChatGPT-3.5’s training on vast datasets and ability to understand patterns and generate contextually sound responses make it a prime candidate for its integration into higher education, particularly for exploring personalized learning [10], educator support [11], and online test preparation [12]. For example, a recent study by Humphrey and Fuller explored the potential of ChatGPT-3.5 to automatically generate the discussion section of chemistry lab reports while also investigating methods to detect AI-generated text in these reports [13]. In a separate investigation conducted by Emenike and Emenike, the authors accurately highlighted that while valid concerns surround the integration of ChatGPT, its vast potential across a spectrum of educational and professional domains, ranging from

instructional materials to research documents, cannot be ignored in the long term. The authors correctly mentioned that shortly, the discussion will not revolve around whether to integrate GAI into chemistry education but how to incorporate it methodically [14]. To this end, Leon et al. investigated the effectiveness of ChatGPT as a potential tutor for a college-level introductory chemistry course [15]. The authors employed prompts directly from question banks, anticipating that students would use these freely accessible platforms to obtain answers if tutors were unavailable or to cheat on online exams. Surprisingly, ChatGPT-3.5's performance was not only inconsistent but also unaffected by the structure of the prompt, whether a multiple-choice (MCQ) or free-response (FR) question. With a maximum score of only 37%, ChatGPT-3.5 failed a typical introductory chemistry final exam. It could correctly answer discipline-specific questions when presented with distractors from other subjects, indicating its ability to utilize context to guide its responses. However, it could not determine the correctness of its chosen answers using computational or analytical methods.

Amid its initial success, concerns regarding its potential shortcomings began shortly after its launch. For example, the phenomenon known as AI hallucinations, where ChatGPT-3.5 makes up facts seemingly from nowhere [16–18], or the model's propensity to repeatedly produce factually incorrect or misleading responses [19,20] or fabricate information [21,22], has raised legitimate concerns about its reliability in real-world applications. ChatGPT's response generation mechanism causes several of these limitations. Like other LLMs, the strength of ChatGPT lies in predicting the next character based on probabilities learned during its supervised training. For instance, it may provide accurate responses to math queries but does not solve them like a calculator, often leading to inconsistencies in complex mathematical prompts. Thus, the question arises whether the perceived shortcomings of ChatGPT-3.5 stem from its inherent flaws or the user's inability to operate an overly sophisticated technology [7]. Considering ChatGPT's ability to process concepts efficiently and reproduce them flawlessly, it is reasonable to conclude that using this tool is far more complex and nuanced than using it as an internet search engine. Poorly constructed prompts often lead to vague or broad responses [23]. Besides technical concerns, there has also been a growing debate around GAI's potential to disrupt the workplace and the concerns it raises regarding potential job displacement. In response to these concerns, the concept of human–AI intelligence, also known as HAI, where artificial intelligence (AI) complements human intelligence (HI) rather than replaces it, is gaining prominence [24–27]. This approach integrates human creativity and ethical considerations with AI's computing power, data analysis, and pattern recognition abilities. A key element on the human side of human–AI intelligence (HAI) involves the creation of meticulously crafted prompts. As a result, the scientific community is witnessing the emergence of prompt engineering as a critical tool to ensure the accuracy and reliability of ChatGPT-3.5 responses [7,28–32]. By using prompt engineering techniques [33] such as crafting clear and specific prompts, specifying the desired format and structure, providing context and background information, applying constraints and limitations, and employing iterative refining, the model can be guided towards generating accurate responses [34].

This investigation employs various prompt engineering techniques to demonstrate ChatGPT-3.5's capability to generate consistently accurate responses, particularly in higher education related to general chemistry. Notably, this study focuses on achieving correct responses without providing feedback in the form of a thumbs-up/thumbs-down in the chat window of ChatGPT-3.5, showing that the model did not adapt to user satisfaction during the investigation. The findings of this study will equip educators with a possible roadmap for incorporating ChatGPT into their classrooms.

2. The Experimental Design

Source of the Questions: The present investigation was conducted using ChatGPT-3.5 [35]. The case study involved 30 multiple-choice questions obtained on the key topics in the general chemistry I course, such as atomic and molecular structure, chemical reactions,

aqueous solutions, thermochemistry, electronic structure, periodic properties, chemical bonding, molecular geometry, gas behavior, and measurement techniques (See Table S1 in Supplementary Materials). These questions closely aligned with three cognitive levels in Bloom's taxonomy, viz. (1) the "*apply level*", assessing the application of chemical principles to solving problems; (2) the "*understand level*", evaluating the comprehension of chemical concepts; and (3) the "*remember level*", gauging the ability to recall this information [36–38].

Mode of Delivery: Initially, the exam questions were presented to ChatGPT-3.5 in two distinct ways:

1. **Comprehensive Presentation:** All 30 multiple-choice questions, each with five answer choices, were provided to ChatGPT-3.5 simultaneously, treating the entire exam as a single prompt.
2. **Sequential Presentation:** The questions were presented individually, with ChatGPT-3.5's response to each question recorded before proceeding to the next. Each multiple-choice question and its five answer choices were treated as a separate prompt.

The above presentation methods were repeated five times, resulting in a total of ten exams. From this dataset, the questions with an error rate exceeding 50%, meaning ChatGPT-3.5 answered incorrectly more than five times out of ten, were selected for prompt refining techniques. These techniques included providing contextual cues, clearly defining symbols, stating the inherent constraints, replacing open-ended prompts with direct questions, and iterative prompt refinement.

Erasing Chat History: The chat history was erased after each prompt revision to isolate the effects of each prompt revision on ChatGPT's output. This step was necessary because the model's self-attention mechanism considers the contexts from the prior interactions within the same context or session window before generating a response. It does that by assigning weights to previous tokens and incorporating information from earlier prompts that might not be relevant to the revised version. As a result, unrefined prompts and corresponding responses could unintentionally influence the outputs for the subsequently presented, more refined prompts.

Prompt Refinement and Testing: To ensure the dependability of our refined prompts, each was subjected to ten iterations of testing. We adopted a two-pronged approach to this evaluation: Firstly, we replicated the same prompt within the confines of a single context window ten times. Since no additional data were introduced during these repetitions, we anticipated that the model's responses would remain consistent, unaffected by any potential accumulation of information from previous outputs. Secondly, we ran each prompting strategy ten times in separate context windows. This method allowed us to ascertain whether ChatGPT exhibited any carryover of information from prior responses when a user repeated identical prompts within the same conversational thread. By comparing the results of these two testing approaches, we aimed to evaluate whether the buildup of responses in a given context influenced the model's output and whether ChatGPT's accurate answers were due to optimized prompts or simply carrying contextual information.

The Iterative Prompting Method: Iterative prompting involves continuously refining the AI prompts to achieve the desired outcomes. This process entails creating an initial prompt, evaluating its output, adjusting the initial prompt based on the analysis, running the modified prompt, examining the results, and refining the prompt until the desired results are achieved. For example, an unrefined prompt might initially work but eventually fail upon further repetitions in a separate context window, indicating the need for additional refinement.

Reliability and Replicability Checks: Our methodology adheres to a scientific method by incorporating reliability and replicability checks. We first confirm the reliability of our findings by repeating the experiment under the same conditions and demonstrating consistent results. Next, we establish their replicability by repeating the experiment with different conditions, verifying that the observed effects are held in broader contexts.

3. Results and Discussion

The present study is based on the premise that most ChatGPT users, mainly students, will use it in its default configuration. Thus, the primary goal of this investigation was to evaluate different prompt engineering strategies to incorporate ChatGPT as an instructional tool into higher education chemistry. First, ChatGPT-3.5's proficiency in answering general chemistry questions was analyzed using a sample of 30 multiple-choice questions which aligned with the "apply level" (21), the "understand level" (3), and the "remember level" (1) of Bloom's taxonomy (Table 1). Across ten trials, five involving all 30 questions given at once and the other five with questions given one at a time, ChatGPT-3.5 scored an average of 62% (Table S2 in Supplementary Materials). Notably, on three tests, ChatGPT-3.5 secured a passing score of 70%. This average, recorded in November and December 2023, surpassed ChatGPT-3.5's previously reported scores (18–22%) on comparable introductory chemistry assessments obtained between January and March 2023 [15], demonstrating a notable technological advancement within a few months.

Table 1. Classification of questions based on Bloom's taxonomy.

Bloom's Taxonomy	Type of Questions by Category	Number of Questions Answered Incorrectly out of Total Questions
The Apply Level	21	11
The Understand Level	9	3
The Remember Level	1	0

While the data followed a normal distribution, more than analyzing ten tests with a paired *t*-test may be required to say which delivery method is better. Still, ChatGPT-3.5 performed better, with an average of 65%, when questions were delivered individually rather than all 30 questions being given at once (60%) (See Table S2 in Supplementary Materials). Furthermore, in 10 trials, 14 questions showed an average score below 60%. Most of these questions were at the "apply level" (37%, Table 1), followed by the "understand level" (10%), within Bloom's classification. Also, category-wise, the occurrence of incorrect answers was significantly higher in the "apply level" (52%, Table 1) than within the "understand level" (38%, Table 1) or the "remember level". These findings were consistent with a brief examination conducted by Herrmann Werner and colleagues investigating ChatGPT-3.5's capabilities in answering medical science questions that were categorized using Bloom's taxonomy of cognitive skills [39]. This study demonstrated ChatGPT-3.5's inability to translate concepts into novel scenarios ("apply" level) or provide illogical explanations for a phenomenon ("understand" level). Recognizing the shortcomings of ChatGPT-3.5 in handling "apply"- and, to some extent, "understand"-level inquiries, the incorporation of human expertise (human intelligence—HI) through the application of prompt engineering techniques emerges as a critical factor in maximizing the efficacy of ChatGPT-3.5 [27].

3.1. Improving AI Accuracy: The Role of Contextual Cues and Symbol Clarity

One of the strategies in crafting prompts involves presenting context or relevant information before posing questions to ChatGPT-3.5, especially in specialized subjects. Since ChatGPT is trained on an extensive dataset, providing appropriate cues should guide it in utilizing the patterns and associations learned from its training phase to generate accurate responses. Moreover, clearly defining symbols employed within the prompt also prevents misinterpretations. A question with a success rate of 1 out of 10 trials, corresponding to the "apply" level of Bloom's taxonomy, was selected to illustrate this point (Figure 1). The original prompt, "Which isoelectronic series is correctly arranged in order of increasing radius?", lacked specific guidance and did not explicitly direct ChatGPT-3.5 to consider the concepts of ionic and atomic sizes before generating a response. Additionally, the prompt does not clarify the meanings of the ">" and "<" symbols, potentially leaving

ChatGPT-3.5 to rely on its existing understanding and increasing the likelihood of errors. As seen from the iterations (see Table S3 in Supplementary Materials), ChatGPT-3.5 gave a correct explanation but chose the incorrect order.

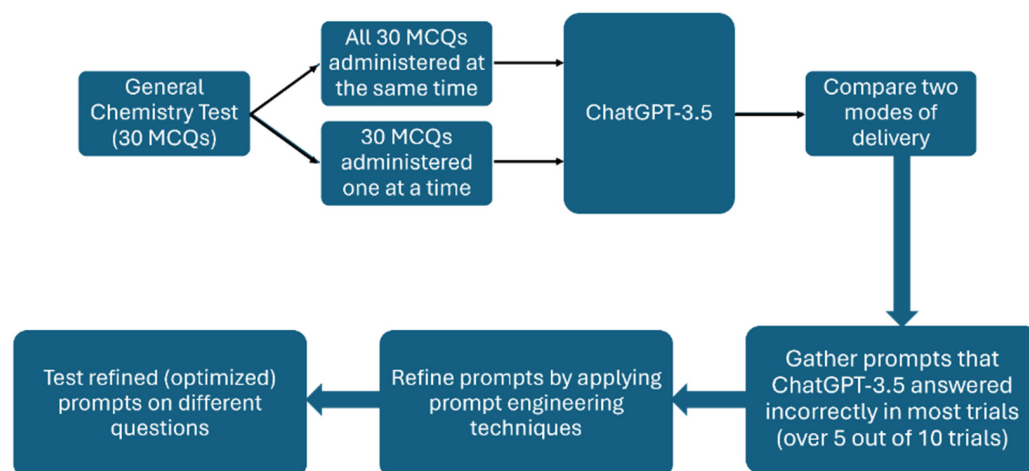


Figure 1. Flowchart illustrating the experimental design.

When the prompt was rephrased as “Review the concepts of ionic and atomic sizes. Know that “>” means greater than and “<” means less than in the following series. Which isoelectronic series is correctly arranged in order of increasing radius?”, ChatGPT-3.5 provided accurate responses in ten consecutive instances without errors (see Figure 2 and Table S4 in Supplementary Materials for iterations). Strikingly, ChatGPT-3.5 correctly placed argon (Ar) and the potassium ion (K^+) despite the potassium ion’s larger size. The modified prompt proved equally effective when applied to another isoelectronic series comprising O, F, Na, and Mg, excluding the noble gas neon (See Figure S1 in Supplementary Materials). Particularly noteworthy was the instance involving a 36-electron isoelectronic series. In this scenario, one of the options included Na^+ , which was not isoelectronic with the other ions. ChatGPT-3.5 provided a correct explanation for not choosing this option by highlighting that Na^+ did not share isoelectronic characteristics with the other ions (see Figure S1 and Table S4 in Supplementary Materials).

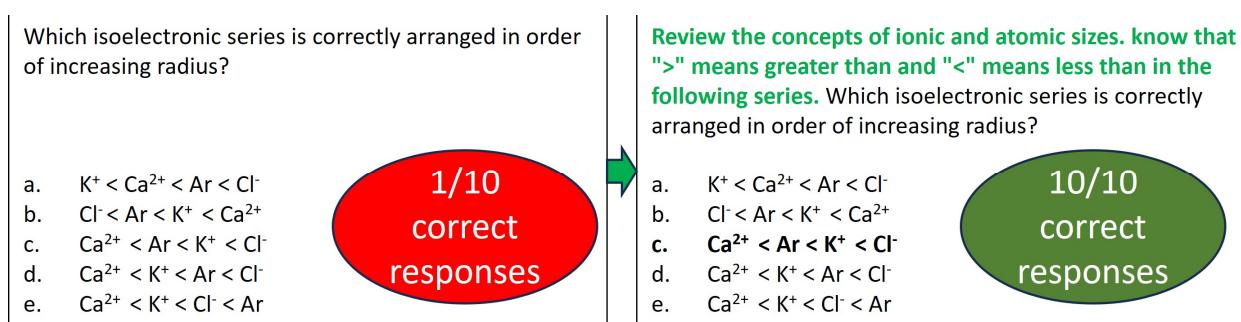


Figure 2. Effect of cues and symbol definitions on ChatGPT-3.5’s accuracy. **(Left)** One correct response in ten attempts on the unmodified question. **(Right)** Ten correct responses in ten attempts on the modified question.

3.2. Beyond Guesswork: Emphasizing Inherent Constraints for Accurate Responses

While ChatGPT-3.5 can handle conceptual prompts, hidden constraints can occasionally lead to inaccuracies in its responses. The concept of inherent constraints involves general expectations that users may have about ChatGPT’s understanding. As a result, users need to pay more attention to explicitly defining these constraints, which could lead ChatGPT to offer inaccurate responses. This became apparent when ChatGPT-3.5 was

directed to identify an invalid set of quantum numbers, a prompt that was categorized as the “apply” level of Bloom’s taxonomy. The unaltered prompt, presented as “Which of the following is not the valid set of four quantum numbers? (n, l, ml, ms)”, was directly copy-pasted into ChatGPT-3.5, mirroring a student’s approach to obtain an answer (Figure 3). Generally, a student would recollect the allowed values of n, l, and ml and apply them to the given choices. ChatGPT-3.5 gave five correct answers in ten trials without these constraints in the prompt. A closer analysis of its incorrect responses showed that ChatGPT-3.5 struggled with the relationship between principal and angular momentum quantum numbers (see Table S5 in Supplementary Materials for iterations). When ChatGPT-3.5 was provided with a cue to “consider the set of quantum numbers (n, l, ml, ms) for an electron in an atom”, initially, there was an improvement for a few iterations before the model began to falter. However, introducing an additional instruction, “Pay special attention to the allowed values of n and l”, altered the outcome. Thus, a simple fix—adding a clarifying phrase—not only transformed its performance from five correct answers in ten trials to all correct answers in ten consecutive trials but also worked for a different set of invalid quantum numbers (see Figure 3 and Figure S2 in Supplementary Materials).

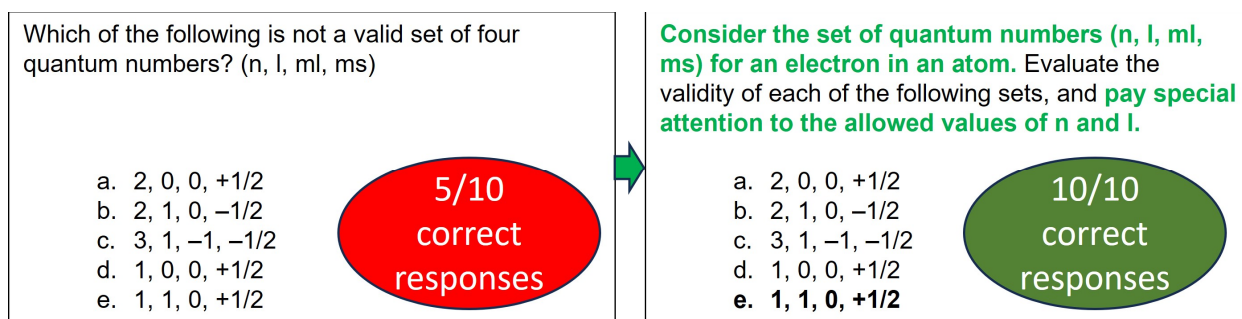


Figure 3. Effect of emphasizing inherent constraints on the response of ChatGPT-3.5. (Left) Unaltered prompt. (Right) Modified prompt.

Interestingly, a similar fix involving formatting the prompt worked for a question about significant figures in a multistep calculation. In ten successive attempts using the original unformatted prompt (Figure 4A), ChatGPT-3.5 consistently generated incorrect responses, revealing an interesting pattern of favoring the choice with three significant figures (see Table S6 in Supplementary Materials). This indicates that ChatGPT-3.5 recognized the concept of maintaining a minimal number of significant figures in a calculation but needed help to apply it. A closer examination of the prompt shows that the format of the mathematical expression lacks instructions regarding the order of operations [40]. Since solving algebraic expressions requires adhering to a specific sequence, such as PEMDAS, the AI could have been misled without such guidelines. Adding parentheses to the expression helped ChatGPT-3.5 to follow the rules of operation, boosting its accuracy to nine correct answers over ten trials (Figure 4B). Applying a similar technique worked successfully for a different set of numbers while preserving the same number of significant figures (Figure 4C). However, the accuracy dropped to 2 correct answers over 10 trials when ChatGPT-3.5 was provided with a similar problem but with different numbers and significant figures (Figure 4D). The analysis of the responses revealed the model’s inability to apply rounding rules consistently (see Table S7 in Supplementary Materials). Thus, providing specific instructions such as “Round the final answer based on significant figures rules” dramatically improved its performance, achieving correct responses over all ten trials (see Figure 4E and Table S8 in Supplementary Materials). These findings emphasize the importance of tailoring the prompts to the specific nuances of different disciplines. Even seemingly trivial details like parentheses can substantially impact the accuracy of AI responses. Interestingly, changing the mathematical operations while maintaining the numbers caused ChatGPT-3.5 to revert to its erratic behavior (Figure 4F). As a result, the model may require further refinement for its general use in calculating significant figures. This

also agrees with previous studies demonstrating ChatGPT-3.5's limitations in responding to mathematical queries [41,42].

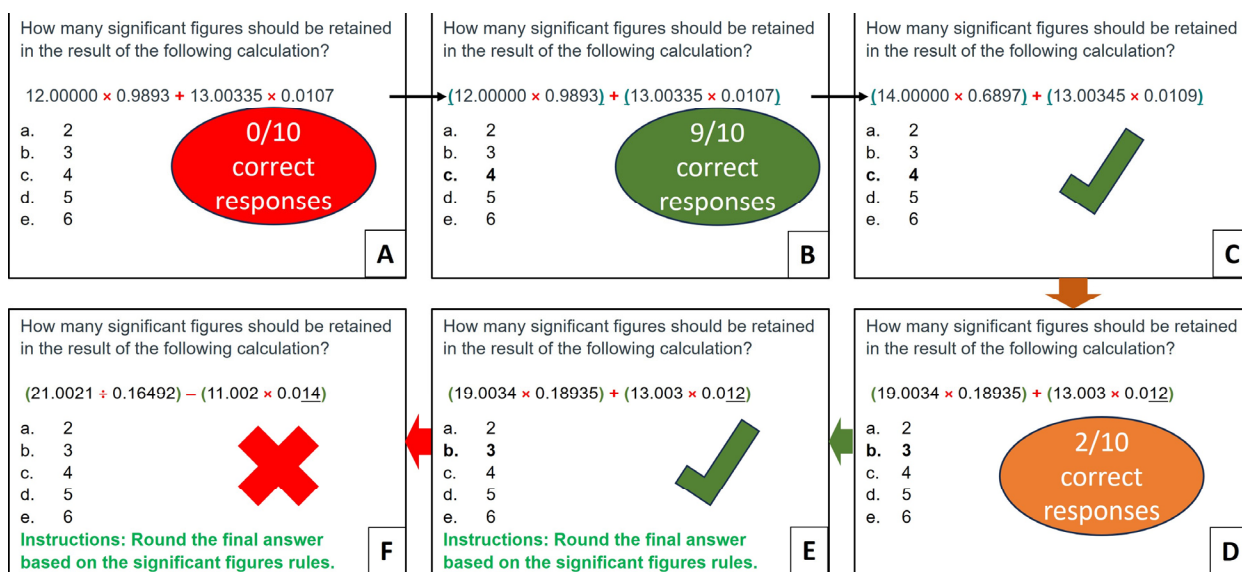


Figure 4. Impact of explicitly stating the inherent constraints of PEMDAS and adjusting formatting on ChatGPT-3.5's accuracy. (A) Original prompt without changes. (B) Prompt with added parentheses for improved clarity. (C) Formatted prompt with different sets of numbers while preserving the same number of significant figures. (D) Formatted prompt with different sets of numbers and significant figures. (E) Revised prompt with special instructions on how to follow the rules for rounding the significant figures. (F) Revised prompt with different mathematical operations.

3.3. Balancing Variation and Accuracy: Avoiding Open-Ended Prompts

While fill-in-the-blank prompts are common in higher education assessment, administering such prompts to large language models (LLMs) like ChatGPT-3.5, particularly in STEM fields, can generate erroneous responses. Unlike students, who generally stay within the boundaries of topic and context, ChatGPT-3.5 deviates from the standard boundaries set by human norms and mores when filling in missing information. While ChatGPT-3.5 generates responses demonstrating large variations, users may find them confusing, potentially leading to a misunderstanding of essential concepts and hindering the learning process. To prevent ChatGPT-3.5 from providing false or misleading information, replacing open-ended formats like fill-in-the-blank prompts with more direct alternatives is desirable. This was evident in a fill-in-the-blank question where ChatGPT-3.5 had to identify a gas from multiple choices based on mass, volume, pressure, and temperature (see Figure 5 and Table S9 in Supplementary Materials). Specifically, ChatGPT-3.5 was required to calculate moles using the ideal gas equation, determine the molar mass of the unknown gas, compute the molar masses of all the gases provided in the choices, and match the unknown gas to the provided options for identification. This question, aligned with the “apply” level in Bloom’s taxonomy, exhibited an accuracy of three correct responses over ten trials when presented in its unmodified form: “The volume of a sample of gas (2.49 g) was 752 mL at 1.98 atm and 62 °C. The gas is _____.” Interestingly, the responses suggested that the incorrect answers were likely the result of guesswork rather than robust reasoning processes. Additionally, the multiple choices only included the formulas instead of listing both formulas and their names (Figure 5).

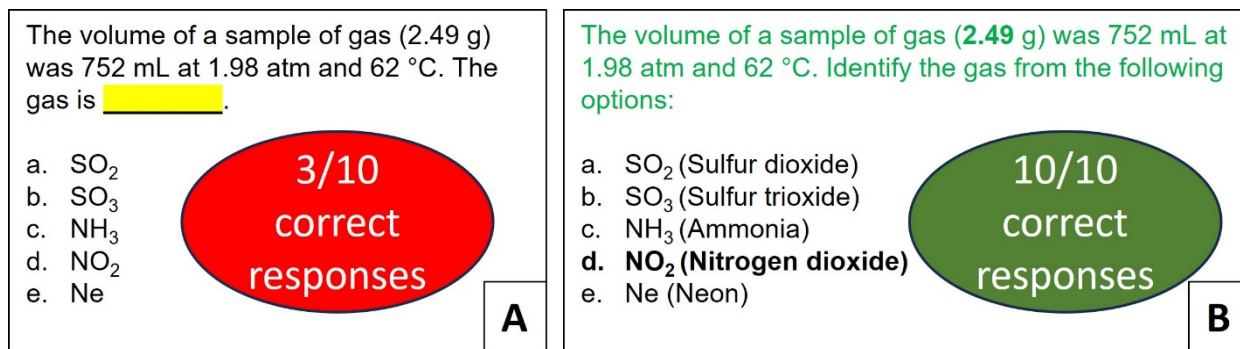


Figure 5. Exploring the effects of shifting from fill-in-the-blank to direct question format: (A) examining accuracy in the fill-in-the-blank format; (B) evaluating accuracy in the direct prompt format.

Switching from a fill-in-the-blank format to a prompt structured as "The volume of a sample of gas (2.49 g) was 752 mL at 1.98 atm and 62 °C. Identify the gas from the following options:" and listing the formulas and corresponding names of the gases, enhanced ChatGPT-3.5's accuracy to correct responses across all ten iterations (see Figure S3 and Table S10 in Supplementary Materials). In addition, ChatGPT-3.5 successfully identified various gases from the same choices when parameters such as temperatures, weights, and pressures were modified (see Figure S3 and Table S10 in Supplementary Materials). To further examine this approach, we selected another "apply"-level prompt from Bloom's taxonomy, which contained two fill-in-the-blanks and was structured as "The hybridizations of iodine in IF₃ and IF₅ are _____ and _____, respectively" (see Figure 6 and Table S11 in Supplementary Materials). The unmodified prompt yielded four correct answers over ten trials, whereas with the format of a direct question, the accuracy increased to 100% over ten trials (see Figure S4 and Table S12 in Supplementary Materials). Furthermore, ChatGPT-3.5 continued to provide correct responses with the modified prompt even when choices were altered (see Figure S4 in Supplementary Materials). Notably, the revised prompt also proved effective for identifying the hybridization of other classes of compounds, such as hydrocarbons (see Figure S4 in Supplementary Materials) and both high- and low-spin iron (III) and Co(III) complexes (see Table S12 in Supplementary Materials).

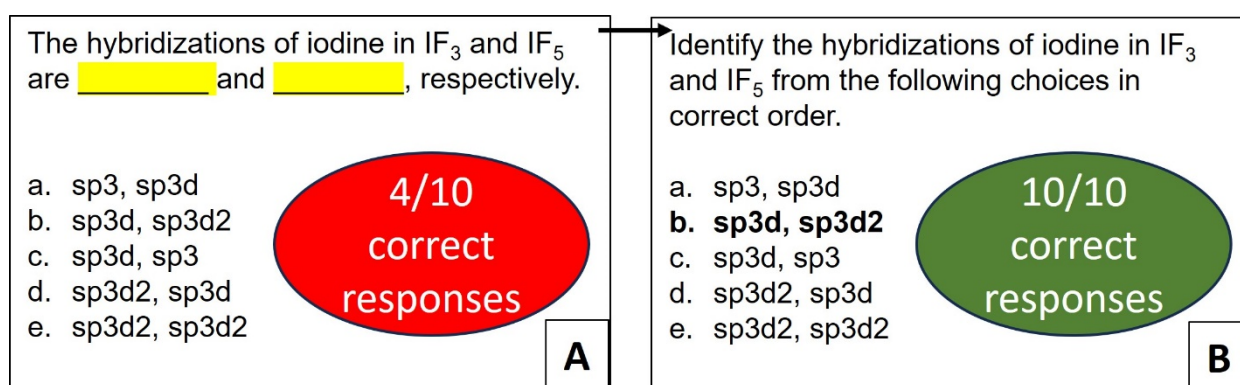


Figure 6. Effects of substituting a double fill-in-the-blank question with a direct question format. (A) illustrates four correct responses over ten trials for the original prompt. (B) Modified prompts gives correct responses across all ten trials.

3.4. Striking a Balance: Designing Prompts with Optimal Information for Accurate Responses

While background information is crucial for accurate responses from ChatGPT-3.5 in complex prompts, overindulging or misplacing it can lead to inaccurate responses. Finding the balance between sufficient context and concise framing is critical to unlocking

the full potential of ChatGPT-3.5 [43–45]. For instance, in a general chemistry exam, simply removing the introductory context from the propane combustion prompt, which is categorized as an “apply”-level question in Bloom’s taxonomy, and focusing directly on the specific data led to a notable improvement in ChatGPT-3.5’s accuracy, increasing it from 50% to 100% (see Figure 7 and Tables S13 and S14 in Supplementary Materials). The revised prompt exhibited remarkable generalizability, accurately predicting the target percent yields across a spectrum of scenarios, from different weights of CO₂ (see Figure S5 and Table S14 in Supplementary Materials), with different combinations of weights of CO₂ and propane (see Figure S5 in Supplementary Materials), and even in the combustion of octane with varying numerical weights for both CO₂ and octane (see Figure S5 in Supplementary Materials). While both prompts involve a chemical reaction between propane and oxygen, their approach differs. Prompt 1 immediately delves into a specific experiment, detailing precise quantities of reactants (38.0 g of carbon dioxide, 22.05 g of propane, and excess oxygen), clearly emphasizing calculating the percentage yield for this unique scenario.

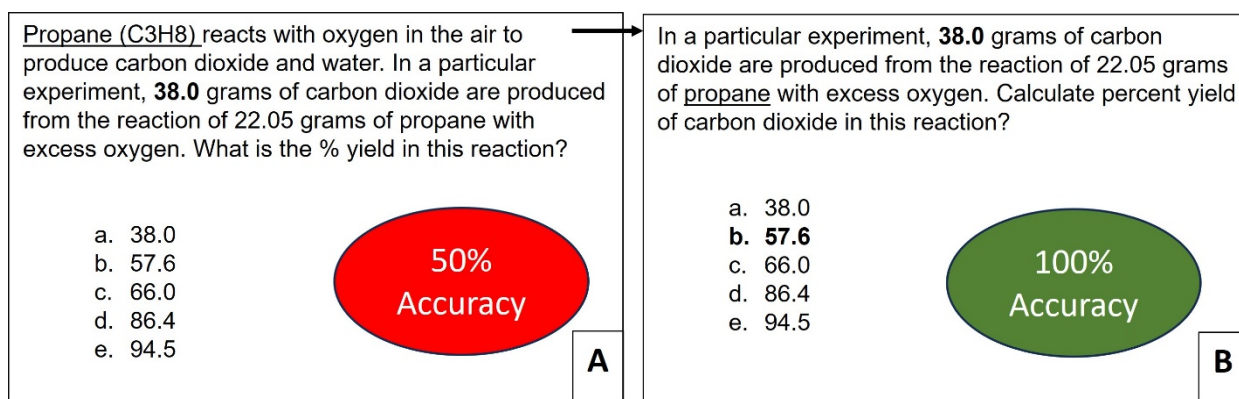


Figure 7. (A) Five correct responses over ten attempts on an unmodified question. (B) Ten correct responses in ten attempts on a modified question.

On the other hand, Prompt 2 begins with a general statement about the propane and oxygen reaction, providing context for the chemical process, and subsequently introduces specific quantities for a particular experiment. The emphasis in Prompt 2 is on the broader chemical reaction, followed by a focus on the specific experimental data and a request to calculate the percent yield in that specific context. Thus, the prompt engineering might be more intricate than initially anticipated. While context can be beneficial, this study showcases how excessive background information can lead to unexpected issues. It can be confusing initially to learn that providing context should improve the performance but then encounter situations where excessive information can hinder the model, particularly with unpredictable outcomes.

3.5. Field-Specific Accuracy: Directing AI with Specialized Chemistry Prompts

In some instances, ChatGPT-3.5 may require specific instructions related to a particular field provided within the prompt. Take, for example, the concept of ionization energy within a general chemistry course, denoting the energy needed to remove electrons from the valence shell of an element or ion (Figure 8A). Typically, an undergraduate student taking this course would be familiar with expecting electrons on the product side of the ionization equation in MCQ. Nevertheless, asking ChatGPT-3.5 a seemingly straightforward question like “Which of the following equations correctly represents the third ionization of aluminum?” without specifying the electron placement resulted in surprisingly inaccurate responses (see Figure 8A and Table S15 in Supplementary Materials). However, there was a notable improvement in its accuracy when ChatGPT-3.5 was explicitly instructed to pay special attention to the placement of the electrons before generating the responses to such

queries (Figure 8B). The validity of this method was confirmed for both the first and second ionization energies of aluminum (Figure 8C,D), as well as for the second ionization energy of calcium (see Table S16 in Supplementary Materials).

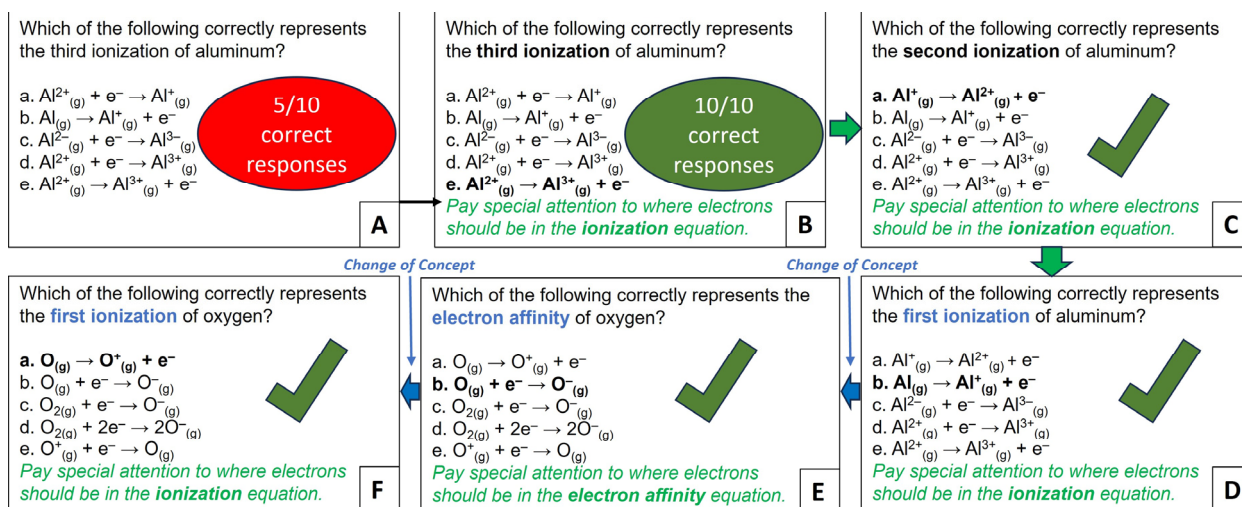


Figure 8. Influence of specialized cues on ChatGPT-3.5 output. (A) Five correct responses over ten attempts on unmodified questions. (B) Ten correct responses in ten attempts on modified questions. (C,D) depict the effectiveness of the revised prompt on aluminum's second and first ionization energies, respectively. (E) represents a conceptual shift in investigating the electron affinity of oxygen using the revised prompt. (F) highlights a conceptual shift, transitioning from electron affinity to the first ionization energy of oxygen, with the choices maintained as in (E).

Remarkably, ChatGPT-3.5 gave accurate responses when a similar approach was applied to an electron affinity, a concept opposite to ionization energy, involving bromine (see Table S16 in Supplementary Materials). The most impressive feat came when it accurately answered a multiple-choice question about oxygen's electron affinity (Figure 8E), followed by a similar question about its first ionization energy (Figure 8F), even though all the answer choices were identical to the previous question. This test underscores the importance of well-crafted prompts, enabling accurate responses and seamless navigation across diverse chemical concepts, from aluminum and calcium ionization to bromine and oxygen electron affinity, culminating in distinguishing between electron affinity and ionization for oxygen (Figure 8).

3.6. Iterative Prompt Design: Enhancing ChatGPT-3.5's Accuracy via Direct Questions, Calculations, and Language Nuances

Iterative prompt design is a loop-based approach that involves continuously refining AI prompts to achieve the desired outcomes. This process involves creating an initial prompt, evaluating its output, adjusting the initial prompt based on the analysis, running the modified prompt, examining the results, and refining the prompt until the desired results are achieved [46]. To illustrate this strategy, a low-accuracy question (three correct responses over ten trials) requiring ChatGPT-3.5 to identify the lowest-energy photon emitted in hydrogen atom transitions was selected (see Figure 9A and Table S17 in Supplementary Materials). The initial prompt can be classified as a borderline case between the "understand" and "apply" levels of Bloom's taxonomy, and it was structured as a multiple-choice, fill-in-the-blank question. Such borderline queries may require prior knowledge and a grasp of concepts but may not explicitly require application to novel scenarios. In these scenarios, students typically show a preference for leveraging their pre-existing knowledge base ("understand") rather than treating the situation as entirely new ("apply"). They may utilize their understanding of Bohr's equation and algebra to deduce that wider energy gaps predict higher-energy photons, guiding their attention toward smaller differences.

Similarly, familiarity with fractions could lead them to infer that $n = 6$ -to- $n = 3$ transitions emit low-energy photons, even though transitions like $n = 5$ to $n = 2$ or $n = 4$ to $n = 1$ appear similar, ultimately leading them to the correct answer through these intuitive leaps. Alternatively, students could meticulously calculate the energy for each transition using Bohr's equation, applying their knowledge more directly but with increased time investment.

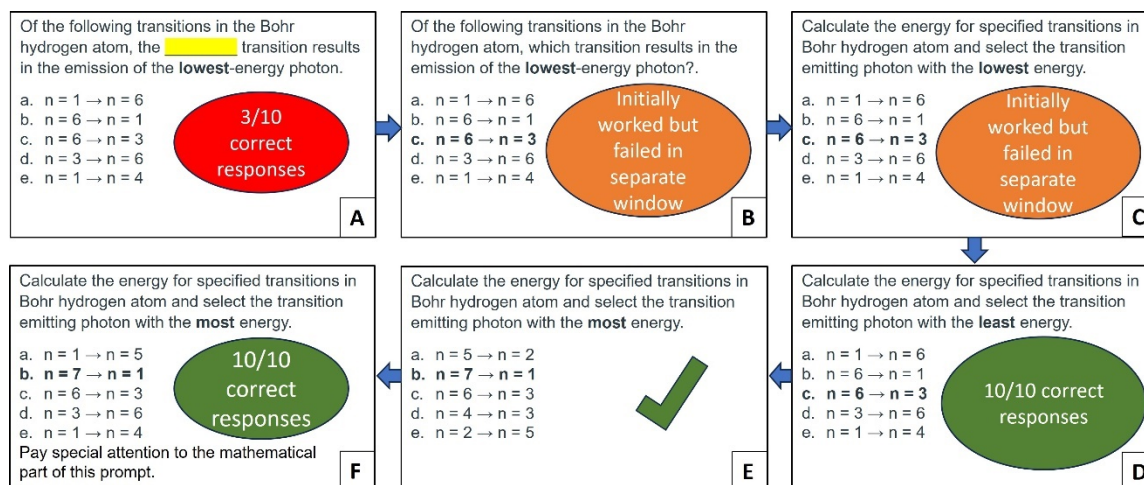


Figure 9. (A) Three correct responses over ten attempts on unmodified question. (B) involves modifying the prompt into a direct question-based query. (C) ChatGPT-3.5 was instructed to perform calculations before providing a response. (D) entails changing the term “lowest” to “least”. (E) demonstrates a case where the antonym of “least”, which is “most”, proved effective. (F) includes the addition of special instructions, directing ChatGPT-3.5 to pay particular attention to the mathematical components of the prompt.

Interestingly, ChatGPT-3.5 used a similar approach and demonstrated an understanding of the concept involved by correctly identifying the relationship between the energy gap ΔE and the energy of the photon. However, this understanding did not translate into choosing the correct transition, with ChatGPT-3.5 consistently opting for the highest-to-lowest ($n = 6 \rightarrow n = 1$) transition instead of selecting the transition with the smallest energy difference (see Figure 9A and Table S17 in Supplementary Materials). So, the prompt was modified from a fill-in-the-blank format to a direct question (Figure 9B). Initially, ChatGPT-3.5 gave correct responses with the modified prompt, but after a couple of iterations, it reverted to selecting transitions from the highest to lowest energy. Subsequently, ChatGPT-3.5 was instructed to process the prompt at the “apply” level, requiring the calculation of photon energies for each transition presented in the multiple choices (Figure 9C). Once again, the refined prompt yielded the correct responses for a brief period before reverting to selecting transitions from the highest to lowest energy. The comprehensive analysis of all the iterations indicated the tendency of ChatGPT-3.5 to focus on the most negative value in terms of both sign and magnitude. However, this approach was incorrect, as the photon with the least negative value should have been the accurate answer.

Interestingly, substituting “lowest” with “least” resulted in ChatGPT-3.5 achieving 100% accuracy across ten trials (see Figure 9D and Table S18 in Supplementary Materials). Furthermore, this approach worked with other electronic transitions (Figure 9E,F). On closer examination, it is evident that the adjective “least” conveys the minimal quantity of a substance, in this context referring to energy. In contrast, “lowest” functions as a comparative adjective, generally referring to the minimum point on a scale. As a result, the discrepancy arose due to the subtlety of language, as ChatGPT-3.5 interpreted the phrase “lowest energy” as a comparative adjective, leading it to select electron transitions from the highest ($n = 6$) to the lowest energy ($n = 1$) levels.

3.7. Repetition within Same and Different Context Windows

Given that we repeated each prompt ten times to test the reliability of the optimized prompts, it was essential to explore the effect of repetitions within the same context window versus separate context windows. In the first method, we repeated the optimized prompts ten times consecutively within the same context window. In the second method, we executed the optimized prompts ten times in ten context windows to compare the results. The goal was to determine whether repeating the same prompt in the same context window yields correct answers because ChatGPT maintains the same context throughout the conversation or because it is inherently the correct answer. Our findings showed that ChatGPT's responses were consistent regardless of the technique used, indicating that the model's performance depends more on the clarity and formulation of the prompt than on the accumulated context. These results align with the existing literature, suggesting that while ChatGPT can use context for chain-of-thought reasoning [47], its core problem-solving abilities remain strong even when such reasoning is limited [48].

3.8. Premium ChatGPT-4.0 vs. Free ChatGPT-3.5 Versions

To assess the impact of the premium version ChatGPT-4's impact, we reevaluated using the same 30 general chemistry questions. ChatGPT-4 achieved 93%, outperforming ChatGPT-3.5's 65%. This translates into 28 correct answers for ChatGPT-4 and only 2 incorrect responses (see Table S19 in Supplementary Materials for the iterations). Both ChatGPT-3.5 and premium ChatGPT-4.0 struggled with identifying the lowest-energy photon in Bohr's hydrogen atom and the order of isoelectronic ions/atoms. Nevertheless, using optimized prompts that successfully provided accurate responses from ChatGPT-3.5 also enabled ChatGPT-4 to provide correct responses. In this context, it is noteworthy that ChatGPT-3.5 continues to be widely utilized, with 27 million users in the USA, of which only 3.9 million (14%) are paid subscribers. As a result, our investigation primarily aimed to establish a benchmark for prompt effectiveness using the freely accessible ChatGPT-3.5, ensuring that our findings were relevant and applicable to the broadest possible audience.

4. Concluding Remarks

The present investigation highlights the role of effective communication in steering ChatGPT-3.5 toward reliable and accurate responses, particularly in chemistry. On the one hand, emphasizing the context and clearly defining mathematical symbols within the prompt dramatically improve the accuracy of ChatGPT-3.5's responses, as observed in the questions related to the isoelectronic series. On the other hand, excessive introductory elaboration could confuse ChatGPT-3.5 and lead to inaccurate responses. Cases like the combustion of propane and octane demonstrated how a direct focus on the subject matter yields far better results than one with an introductory context. This study goes beyond general instructions to emphasize the critical role of field-specific prompts in prompt engineering. For example, directing ChatGPT-3.5 to pay special attention to where the electrons should be in ionization or electron affinity equations proved remarkably effective, boosting the results while transitioning between opposite concepts like electron affinity and ionization equations. To avoid misleading responses, replacing open-ended prompts with direct questions can dramatically improve ChatGPT-3.5's reliability. For example, when identifying unknown gases using the ideal gas equation or predicting the hybridization of compounds, direct questions led to significantly more accurate results. Furthermore, the iterative prompt design demonstrated the refinement process through a trial, analysis, revision, and trial cycle to achieve the desired results. It underscored how seemingly trivial changes, like switching from the word "lowest" to "least", could impact ChatGPT-3.5's output. Finally, the repetition of optimized prompts in the same and different context windows showed that ChatGPT's accurate responses to optimized prompts were due to the prompts themselves and not influenced by repeated contextual information.

The present investigation emphasizes the critical role of human oversight (human intelligence—HI) and highlights the potential of integrating ChatGPT-3.5 into higher ed-

ucation. In terms of implementation, this study proposes utilizing ChatGPT-based large language models (LLMs) to enhance critical thinking skills within chemistry education. For instance, the free version of ChatGPT-3.5 is susceptible to generating inaccurate responses with unoptimized prompts. This very characteristic can be harnessed to stimulate critical analysis when strategically employed in classroom discussions. Instructors can cultivate a learning environment that fosters essential critical thinking abilities by prompting students to assess the veracity of ChatGPT's outputs and justify their reasoning. While this study sheds light on specific "known unknowns" within the risk profile quadrant of ChatGPT-3.5, a broader investigation encompassing uncharted dimensions is currently underway. This more extensive study includes AI hallucinations and potential outcomes of conducting similar studies in various global locations across diverse IP addresses, disciplines, and user accounts.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/chemistry6040043/s1>.

Author Contributions: Conceptualization, D.V.V.; methodology, D.V.V. and M.M.; software, D.V.V.; validation, D.V.V. and M.M.; formal analysis, D.V.V. and M. M.; investigation, D.V.V.; resources, D.V.V.; data curation, D.V.V.; writing—original draft preparation, D.V.V.; writing—review and editing, D.V.V. and M.M.; visualization, D.V.V.; supervision, D.V.V.; project administration, D.V.V.; funding acquisition, D.V.V. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Mr. Antonio Delgado, Ms. Marimar Molinary and Dr. Elodie Billioniere at the Miami Dade College.

Data Availability Statement: The original contributions presented in the study are included in the article/Supplementary Materials, further inquiries can be directed to the corresponding author.

Acknowledgments: D.V.V. wishes to express deepest gratitude to Jyrko Correa, Chair of the Department at MDC, whose unwavering support and encouragement were instrumental in the success of this project. D.V.V. is also grateful to Abhijit Pandya (FAU), Yanet Cusido (MDC), and Sudha for their insightful discussion and Diana Vidhani for her invaluable contributions to this work's computer-related and digital aspects. D.V.V. is also grateful to Elodie Billioniere (MDC), Antonio Delgado (MDC), and Marimar Molinary (MDC) for their crucial financial support, which enabled the publication of this investigation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hwang, Y.H.; Um, J.; Pradhan, B.; Choudhury, T.; Schlüter, S. How does ChatGPT evaluate the value of spatial information in the fourth industrial revolution? *Spat. Inf. Res.* **2024**, *32*, 187–194. [\[CrossRef\]](#)
2. Javaid, M.; Haleem, A.; Singh, R.P. A study on ChatGPT for Industry 4.0: Background, potentials, challenges, and eventualities. *J. Econ. Technol.* **2023**, *1*, 127–143. [\[CrossRef\]](#)
3. Agbaji, D.; Lund, B.; Mannuru, N.R. Perceptions of the Fourth Industrial Revolution and artificial intelligence impact on society. *arXiv* **2023**, arXiv:2308.02030. [\[CrossRef\]](#)
4. Ali, S.R.; Dobbs, T.D.; Hutchings, H.; Whitaker, I.S. Using ChatGPT to write patient clinic letters. *Lancet Digit. Health* **2023**, *5*, e179–e181. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Khan, R.A.; Jawaid, M.; Khan, A.R.; Sajjad, M. ChatGPT-Reshaping medical education and clinical management. *Pak. J. Med. Sci.* **2023**, *39*, 605–607. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Park, Y.J.; Jerng, S.E.; Park, J.S.; Kwon, C.; Hsu, C.W.; Ren, Z.; Yoon, S.; Li, J. 1.5 million materials narratives generated by chatbots. *arXiv* **2023**, arXiv:2308.13687. [\[CrossRef\]](#)
7. Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J.T.; Yaghi, O.M. ChatGPT Chemistry Assistant for Text Mining and Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **2023**, *145*, 18048–18062. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Biswas, S.S. Role of ChatGPT in public health. *Ann. Biomed. Eng.* **2023**, *51*, 868–869. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Biswas, S.S. Potential use of chat gpt in global warming. *Ann. Biomed. Eng.* **2023**, *51*, 1126–1127. [\[CrossRef\]](#)
10. Baidoo-Anu, D.; Owusu Ansah, L. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *J. AI* **2023**, *7*, 52–62. [\[CrossRef\]](#)
11. Francl, M. ChatGPT saves the day. *Nat. Chem.* **2023**, *15*, 890–891. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Tlili, A.; Shehata, B.; Adarkwah, M.A.; Bozkurt, A.; Hickey, D.T.; Huang, R.; Agyemang, B. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ.* **2023**, *10*, 15. [\[CrossRef\]](#)

13. Humphry, T.; Fuller, A.L. Potential ChatGPT Use in Undergraduate Chemistry Laboratories. *J. Chem. Educ.* **2023**, *100*, 1434–1436. [CrossRef]
14. Emenike, M.E.; Emenike, B.U. Was this title generated by ChatGPT? Considerations for Artificial Intelligence text-generation software programs for chemists and chemistry educators. *J. Chem. Educ.* **2023**, *100*, 1413–1418. [CrossRef]
15. Leon, A.; Vidhani, D.V. ChatGPT needs a chemistry tutor too. *J. Chem. Educ.* **2023**, *100*, 3859–3865. [CrossRef]
16. Alkaissi, H.; McFarlane, S.I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* **2023**, *15*, e35179. [CrossRef] [PubMed]
17. Salvagno, M.; Taccone, F.S.; Gerli, A.G. Artificial intelligence hallucinations. *Crit Care* **2023**, *27*, 180. [CrossRef] [PubMed]
18. Goddard, J. Hallucinations in ChatGPT: A cautionary tale for biomedical researchers. *Am. J. Med.* **2023**, *136*, 1059–1060. [CrossRef]
19. Veyseh, A.P.B.; Lai, V.; Dernoncourt, F.; Nguyen, T.H. Unleash GPT-2 power for event detection. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Stroudsburg, PA, USA, 1–6 August 2021; Volume 1.
20. Walters, W.H.; Wilde, E.I. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci. Rep.* **2023**, *13*, 14045. [CrossRef]
21. Emsley, R. ChatGPT: These are not hallucinations—they’re fabrications and falsifications. *Schizophrenia* **2023**, *9*, 52. [CrossRef]
22. Bhattacharyya, M.; Miller, V.M.; Bhattacharyya, D.; Miller, L.E. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus* **2023**, *15*, e39238. [CrossRef]
23. Zamfirescu-Pereira, J.D.; Wong, R.Y.; Hartmann, B.; Yang, Q. Why johnny can’t prompt: How non-ai experts try (and fail) to design llm prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI’23), Hamburg, Germany, 23–28 April 2023; pp. 1–22.
24. Yang, J.H.; Ogata, H.; Matsui, T.; Chen, N.S. Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100008. [CrossRef]
25. Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum. Comput. Interact.* **2020**, *36*, 495–504.
26. Santos, R.P.D. Enhancing Physics Learning with ChatGPT, Bing Chat, and Bard as Agents-to-Think-With: A Comparative Case Study. *arXiv* **2023**, arXiv:2306.00724. [CrossRef]
27. Santos, R.P.D. Enhancing Chemistry Learning with ChatGPT and Bing Chat as Agents to Think with: A Comparative Case Study. *arXiv* **2023**, arXiv:2305.11890.
28. Zhu, J.-J.; Jiang, J.; Yang, M.; Ren, Z.J. ChatGPT and Environmental Research. *Environ. Sci. Technol.* **2023**, *57*, 17667–17670. [CrossRef] [PubMed]
29. Zheng, Z.; Rong, Z.; Rampal, N.; Borgs, C.; Chayes, J.T.; Yaghi, O.M. A GPT-4 Reticular Chemist for Guiding MOF Discovery. *Angew. Chem. Int. Ed.* **2023**, *62*, e202311983. [CrossRef] [PubMed]
30. White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D.C. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv* **2023**, arXiv:2302.11382.
31. Giray, L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Ann. Biomed. Eng.* **2023**, *51*, 2629–2633. [CrossRef]
32. Meskó, B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *J. Med. Internet Res.* **2023**, *25*, e50638. [CrossRef]
33. Polverini, G.; Gregorcic, B. How understanding large language models can inform the use of ChatGPT in physics education. *Eur. J. Phys.* **2024**, *45*, 025701.
34. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* **2023**, *3*, 121–154. [CrossRef]
35. OpenAI ChatGPT Login. Available online: <https://chatgpt.com/> (accessed on 31 December 2023).
36. Bloom, B.S. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain*; Longmans, Green: New York, NY, USA, 1956.
37. Bloom, B.S.; Englehard, M.D.; Furst, E.J.; Hill, W.H.; Krathwohl, D. *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*; Longmans: New York, NY, USA, 1956.
38. Anderson, L.W.; Krathwohl, D.R. *A Taxonomy for Learning, Teaching and Assisting: A Revision of Bloom’s Taxonomy of Education Objectives*; Longmans: New York, NY, USA, 2001.
39. Herrmann Werner, A.; Festl-Wietek, T.; Holderried, F.; Herschbach, L.; Griewatz, J.; Masters, K.; Zipfel, S.; Mahling, M. Assessing ChatGPT’s Mastery of Bloom’s Taxonomy using psychosomatic medicine exam questions. *J. Med. Internet Res.* **2024**, *26*, e52113. [CrossRef] [PubMed]
40. Lee, J.K.; Licwinko, S.; Taylor-Buckner, N. Exploring Mathematical Reasoning of the Order of Operations: Rearranging the Procedural Component PEMDAS. *J. Math. Educ. Teach. Coll.* **2023**, *4*, 73–78.
41. Frieder, S.; Pinchetti, L.; Griffiths, R.R.; Salvatori, T.; Lukasiewicz, T.; Petersen, P.C.; Chevalier, A.; Berner, J. Mathematical capabilities of chatgpt. *arXiv* **2023**, arXiv:2301.13867.
42. Kortemeyer, G. Could an Artificial-Intelligence agent pass an introductory physics course? *Phys. Rev. Phys. Educ. Res.* **2023**, *19*, 010132. [CrossRef]
43. Liu, Y.; Ju, S.; Wang, J. Exploring the potential of ChatGPT in medical dialogue summarization: A study on consistency with human preferences. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 75. [CrossRef]

44. Shue, E.; Liu, L.; Li, B.; Feng, Z.; Li, X.; Hu, G. Empowering beginners in bioinformatics with ChatGPT. *Quant. Biol.* **2023**, *11*, 105–108. [[CrossRef](#)] [[PubMed](#)]
45. Zhang, X.; Talukdar, N.; Vemulapalli, S.; Ahn, S.; Wang, J.; Meng, H.; Mehtab, S.; Murtaza, B.; Leshchiner, D.; Dave, A.A.; et al. Comparison of Prompt Engineering and Fine-Tuning Strategies in Large Language Models in the Classification of Clinical Notes. *AMIA Summits Transl. Sci. Proc.* **2024**, *2024*, 478–487. [[CrossRef](#)]
46. Oppenlaender, J. A taxonomy of prompt modifiers for text-to-image generation. *Behav. Inf. Technol.* **2023**, 1–14. [[CrossRef](#)]
47. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2023**, arXiv:2201.11903.
48. Hagendorff, T.; Fabi, S.; Kosinski, M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat Comput Sci.* **2023**, *3*, 833–838. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.