ChatGPT Needs a Chemistry Tutor Too.

Alfredo J. Leon* and Dinesh Vidhani

Miami Dade College, 500 College Terrace, Homestead, FL 33030
Corresponding Author E-mail: aleon6@mdc.edu

Abstract

AI and ChatGPT technology have the potential to revolutionize the education sector, and this study aimed to evaluate if prompt formats, response consistency, and reliability of ChatGPT responses could help colleges make the most of this technology. The results of this study can guide future AI and ChatGPT implementations and ensure they are used to their fullest potential. The data does not demonstrate a statistically significant difference between multiple-choice and free response prompt formats. Neither format achieved scores higher than 37%, and testing different locations did not improve scores. Interestingly, ChatGPT's free version provides accurate responses to discipline-specific questions that contain information from unrelated topics, improving its accuracy over the free response questions. However, it's important to remember that while ChatGPT can identify the correct answer within a given context, it may not be able to tell if the answer it chooses is correct computationally or through analysis.

Keywords

General; multidisciplinary; computer-based learning; self-instruction; web-based learning; assessment; student-centered learning.

Introduction

Artificial Intelligence (AI) has been at the center of conversations as a disruptor for many industries [1]. Within the context of education, AI as Natural Language Processing (NLP) Models have been proposed as tools to generate instructional materials, assessment grading, assistance answering questions, and providing feedback to learners [2–5]. Prevalent discussion has focused on ChatGPT, an NLP Model generated by Open AI based on Generative Pre-trained Transformer (GTP) trained by textual data from the internet [2,3,6]. As part of the design, ChatGPT mimics natural human language engaging in conversation simulating human interaction [2].

Applications of AI related to education and learner engagement is not new. Intelligent Tutoring Systems (ITS) describe computers that integrate AI into tutoring practices that know the material, the learner, and how to teach the material [1]. Moreover, ITS has also been described as an early-alert system and showing no significant difference between the system and human interventions [7]. Wang et al. (2023a) reviewed the literature highlighting implementation of ITS in most educational contexts. Additionally, they described how it has supported disciplines in STEM, history, and economics. Furthermore, Cognitive Tutors were identified as the primary use of the technology and that 65% of studies reported a positive effect on learning performance.

Different applications in the educational context have been proposed related to the use of AI [7,8]. Among the recent is the use of ChatGPT as a personalized learning tool. Because of its ability to answer questions and provide feedback, the AI has been proposed to generate personalized study plans and instructional materials to support learners needs [3,7]. Moreover, due to the conversational nature of the interaction, learner engagement may be achieved by supporting learning as part of social presence [2,9]. Although the potential of the tool has been discussed, the literature underlines limitations related with the tool and the need to review query results for accuracy and appropriateness of language [3,6].

While the potential of ChatGPT has been described, the application of the tool in specific contexts is limited. We were interested in exploring the as a potential support tool for learners within the context of introductory chemistry. As the access and the use of ChatGPT increases, we engaged with the AI to study the outputs that potential learners would encounter. We aimed to determine the efficacy of ChatGPT to correctly answer questions a learner would use, and to elucidate how the AI was processing potential prompts.

Methodology

To test the efficacy of ChatGPT in correctly answering prompts from a college-level introductory chemistry course. We assumed learners would base their study practices on instructor-provided guided questions, traditionally presented as multiple-choice or as free response prompts. Because the responses generated by ChatGPT will be dependent on the way the AI was trained [6], we chose to use both forms of prompts to determine if there were patterns or preferences related to prompt formatting.

Prompts were queried by two different users, at two different physical locations. Each prompt was randomized for users. Users attempted 10 trials for each prompt format, totaling 20 trials. Trials consisted of 27 questions related to introductory chemistry cumulative final exam. After each trial, the prompt window was cleared to prevent the previous information being used [6]. Responses were collected and evaluated based on the output by ChatGPT. Responses from statement prompts were reviewed and determined by the user to be accurate of inaccurate.

After collecting data from the two prompt formats, we wanted to determine if the multiple-choice selection was random. Each of the 27 question prompts in multiple-choice format were modified to include the correct answer and distractors. Distractors were presented as options not related to chemistry. Once again, the responses were collected and evaluated as accurate or inaccurate.

We further explored the use of distractors to understand if the AI was answering the prompts based on understanding or using written language-based contextual cues. As described by [3] the AI has been pre-trained using data that allows ChatGPT to understand context of prompts and generate coherent responses. We used the multiple-choice question format with distractors and an option within the context of chemistry but incorrectly related to the prompt. Responses were collected and evaluated for accuracy.

Results and Discussion

The Introductory Chemistry course prepares learners with the fundamental knowledge and abilities needed to excel in the first semester of General Chemistry 1. This course includes the essential components of modern chemistry, such as measurements, chemical bonds, chemical equations, quantitative relationships, the concentration of solutions, and the system for naming compounds. The primary goal of this study was to investigate if ChatGPT's free original version is beneficial to learners studying for their final cumulative exam. The experiment began by shortlisting the commonly used assessment methods and testing ChatGPT's ability to answer Introductory chemistry questions from the sample exam. In the past, instructors have used free response and multiple-choice questions for this course [10–13]. As a testing strategy, free response questions provide less strain for learners, providing a chance to give more detailed answers. At the same time, they can be time-consuming, and learners produce unnecessary information. On the other hand, multiple-choice questions have advantages, such as being quick and easy to complete, having the potential to include more questions in a survey, being technology friendly, and not including irrelevant data [14]. Additionally, the data obtained from multiple-choice questions are easier to analyze and interpret. However, this type of question does limit respondents in their answers. So, we were interested to determine if ChatGPT's had any preference toward either free response or multiple-choice questions when used as a studying tool. In addition, we wanted to ascertain if there was any bias in the location from which people accessed ChatGPT, given that it is an online tool. It is also essential to recognize that the ChatGPT employs a probabilistic approach to comprehend the user's input and generate a response that is likely to be accurate based on the data it has been trained on. As such, the responses may not be identical each time.

I. Multiple choice *vs*. free response prompts: Is there a statistically significant difference?

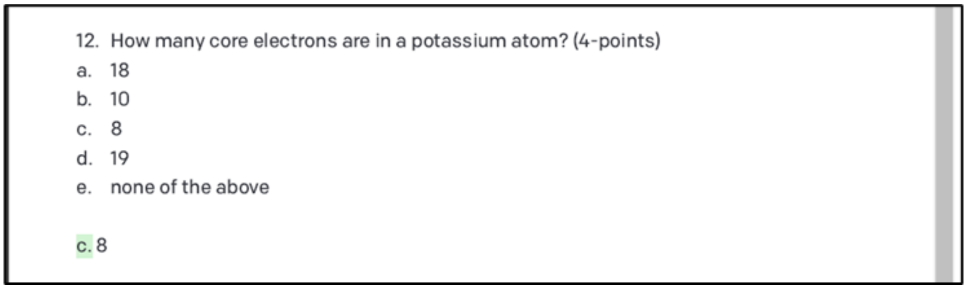
As an NLP, the free version of ChatGPT is designed to generate meaningful language responses to a given prompt [15,16]. We examined the ability of GPT to recognize the prompt in a free response question and the influence of supplemental, multiple-choice prompts in the context of an introductory chemistry course. We are aware that ChatGPT would struggle to answer chemistry-based questions since it is a language-based model and chemistry involves structural features, formulas, and mathematical aspects of a chemical reaction that cannot be conveyed through language alone [17,18]. Nevertheless, it is also imperative for learners to know whether this tool could be utilized despite drawbacks in a fully AI-driven learning environment. To this end, we applied a paired t-test to determine if there was a statistically significant difference in the mean scores between the two formats, namely multiple-choice questions, and free responses. The use of the paired t-test is justified because the two test formats were randomly assigned and administered at two different locations simultaneously, with each test containing the same set of 27 questions and administered ten times [19–21]. The average score of each test was used for further analysis.

On multiple occasions, ChatGPT gave varied answers regardless of the query's structure. For instance, when administered the question, "how many core electrons does potassium have?" ChatGPT responded with 8, 19, 18, 20, 10, and 4. While the correct being 18, it could be confusing for someone inquiring about the same topic multiple times from ChatGPT (Figure 1).

The hypotheses are stated in terms of the mean of the difference where,
μd = mean difference in scores (FR – MCQ):

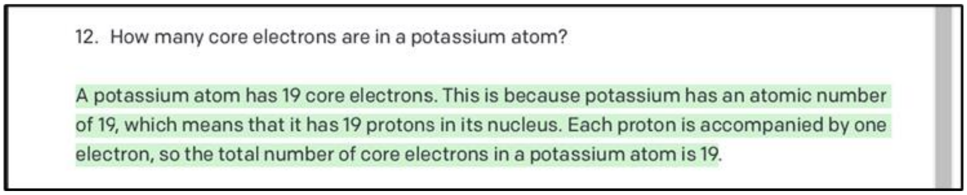
Ho: µd = 0 (the differences are centered at '0')
Ha: µd ≠ 0 (the differences are NOT centered is at '0')

$t = \frac{\sum d}{\sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n-1}}}$ Where Σd is the sum of differences



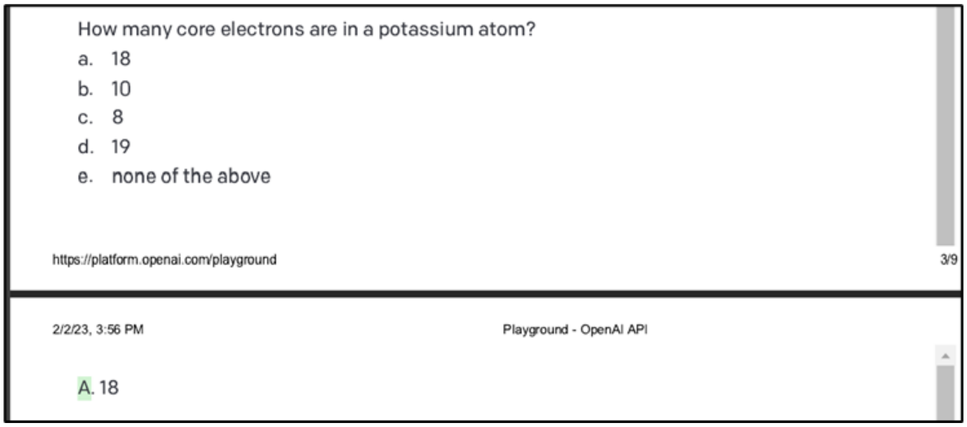
12. How many core electrons are in a potassium atom? (4-points)
a. 18
b. 10
c. 8
d. 19
e. none of the above

c. 8



12. How many core electrons are in a potassium atom?

A potassium atom has 19 core electrons. This is because potassium has an atomic number of 19, which means that it has 19 protons in its nucleus. Each proton is accompanied by one electron, so the total number of core electrons in a potassium atom is 19.

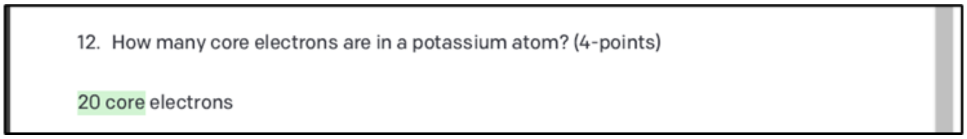


How many core electrons are in a potassium atom?
a. 18
b. 10
c. 8
d. 19
e. none of the above

https://platform.openai.com/playground 3/9

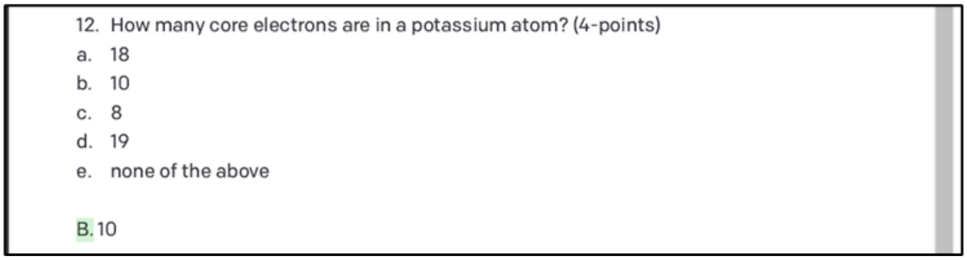
2/2/23, 3:56 PM Playground - OpenAI API

A. 18



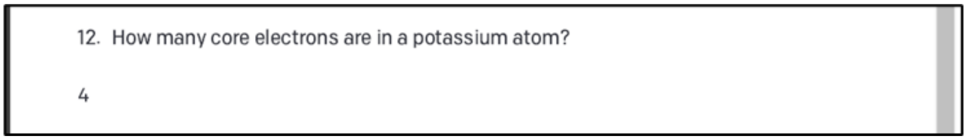
12. How many core electrons are in a potassium atom? (4-points)

20 core electrons



12. How many core electrons are in a potassium atom? (4-points)
a. 18
b. 10
c. 8
d. 19
e. none of the above

B. 10



12. How many core electrons are in a potassium atom?

4

Figure 1. A sample of multiple-choice and free-response question administered to ChatGPT, along with its response.

To our surprise, the mean scores on tests were far lower than the required passing grade of 70%, ranging from 15-37% (Table 1). The t-test statistic of 1.59 was lower than the t-table 2.262 with 9 degrees of freedom, and a p-value of 0.145 was more than the significance level of 0.05. Even though the multiple-choice question tests had a slightly better score of 37% on three tests, it is not enough to reject the null hypothesis and prove that ChatGPT's accuracy is improved when given additional multiple-choice prompts.

Table 1 Summary statistics of the final introductory chemistry exam administered using MCQ and FR modalities.

	FR	MCQ
Mean	0.2259	0.2667
Median	0.2222	0.2593
Variance	0.0026	0.0063
Standard Deviation	0.0508	0.0796
Minimum	0.1481	0.1481
Maximum	0.2963	0.3704
Count	10	10

II. Exploring the impact of multiple test locations

We used a paired t-test to determine if there was a statistically significant difference in the mean scores between the two physical locations. This method was chosen because the two tests were randomly assigned and administered simultaneously, containing the same 27 questions, and repeated ten times. The average score of each test was then calculated and used for further analysis.

Once again, the average scores on the tests were much less than the passing grade of 70%, varying from 18-36% (Table 2). The t-test statistic of 1.998 was lower than the t-table 2.262 with 9 degrees of freedom, and a p-value of 0.077 was more than the significance level of 0.05. Although the tests administered in Location 1 had a slightly higher success rate of 37% on three tests, it was not enough to reject the null hypothesis and demonstrate that ChatGPT's accuracy had improved when given in that location.

Table 2. Summary statistics of the final introductory chemistry exam administered concurrently at two different locations.

	Location 1	Location 2
Mean	0.2704	0.2222
Median	0.2593	0.2222
Variance	0.0061	0.0024
Standard Deviation	0.0782	0.0494
Minimum	0.1852	0.1481
Maximum	0.3704	0.2963
Count	10	10

III. Exploring the effects of different formats on every Question: A closer look

Earlier attempts examined the mean score of 27 questions on a test. To delve deeper, we utilized the McNemar test, a non-parametric statistical method, to compare the results of multiple-choice questions and free response for each question/prompt [22–25]. This test is suitable for two groups with dichotomous outcomes, like pass or fail, right or wrong, or 0 or 1. This is like the paired-samples t-test, but the dependent variable is binary instead of continuous. In other words, McNemar's test is suitable for analyzing data from a randomly chosen sample of matched pairs, the data can be arranged in a 2 X 2 contingency table, and the frequencies in the four cells are sufficiently large (Table 3).

Table 3 McNemar's 2x2 contingency table.

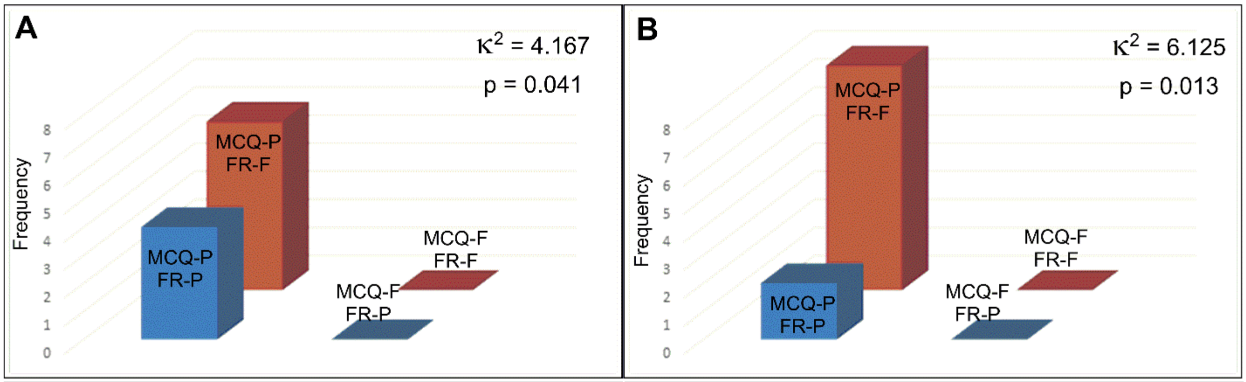
	MCQ-Pass	MCQ-Fail	Totals
FR-Pass	a	b	n1 = a + b
FR-Fail	c	d	n2 = c + d
Totals	m1 = a+c	m2 = b+d	n1+n2 = m1+m2

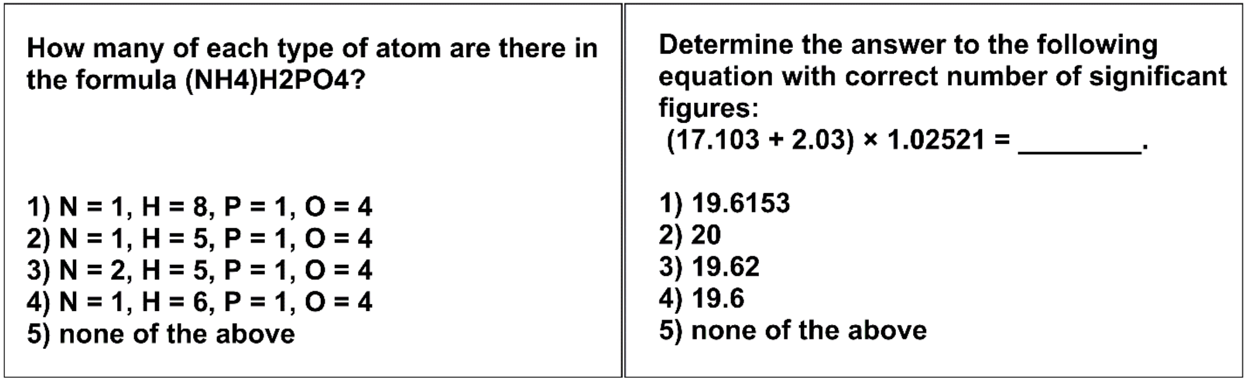
Where a, b, c, and d are the frequencies

The standard McNemar's test statistic is given by the equation 1.

$$\kappa^2 = \frac{(|b-c|-1)^2}{b+c} \qquad \text{Eq. 1}$$

If the chi-squared statistic is less than the critical value of 3.84, then there is insufficient evidence to reject the null hypothesis (Ho) of McNemar's test. Although most questions tested showed no significant difference between multiple-choice questions and free responses, two of the 27 questions demonstrated a preference for MCQ, as indicated by a chi-squared value greater than 3.84. (Figure 2)





How many of each type of atom are there in the formula (NH4)H2PO4?

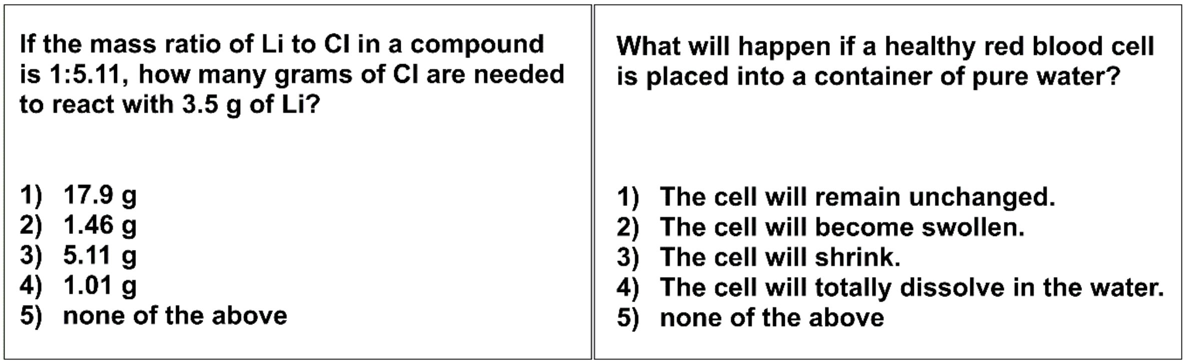
1) N = 1, H = 8, P = 1, O = 4
2) N = 1, H = 5, P = 1, O = 4
3) N = 2, H = 5, P = 1, O = 4
4) N = 1, H = 6, P = 1, O = 4
5) none of the above

Determine the answer to the following equation with correct number of significant figures:
(17.103 + 2.03) × 1.02521 = ________.

1) 19.6153
2) 20
3) 19.62
4) 19.6
5) none of the above

Figure 2. McNemar's analysis of two questions that showed preference for multiple-choice questions (MCQs) over free-response questions (FR).

Furthermore, we were surprised to discover that, despite its statistical nature, ChatGPT answered eight questions incorrectly and got two questions (Figure 3) right consistently, regardless of their format.



If the mass ratio of Li to Cl in a compound is 1:5.11, how many grams of Cl are needed to react with 3.5 g of Li?

1) 17.9 g
2) 1.46 g
3) 5.11 g
4) 1.01 g
5) none of the above

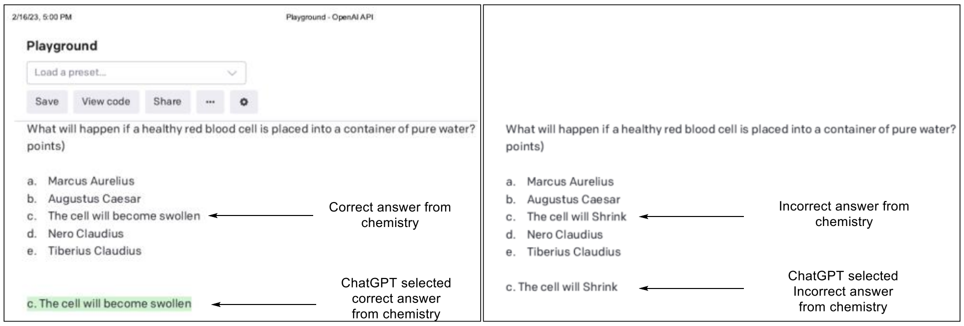
What will happen if a healthy red blood cell is placed into a container of pure water?

1) The cell will remain unchanged.
2) The cell will become swollen.
3) The cell will shrink.
4) The cell will totally dissolve in the water.
5) none of the above

Figure 3. The questions that were answered correctly regardless of the method of communication.

IV. Examining the role of distractors:

So far, our research has revealed that ChatGPT was unable to pass an introductory chemistry course final cumulative exam that included both multiple-choice and free response questions. To analyze its ability to comprehend context, we gave it multiple-choice questions with implausible and distinct alternatives [26–29] to see if ChatGPT's original free version could tell the difference between chemistry and history. We started by selecting a question for which ChatGPT provided the correct response. In the initial trial, we presented it with four options from history not related to chemistry, but the fifth option was a correct answer from chemistry. Not surprisingly, it selected the correct answer from chemistry. Subsequently, we gave it the same four choices from history, but this time we gave it an incorrect response from chemistry. Surprisingly, it chose the wrong answer from chemistry, skipping all history options. It is noteworthy that ChatGPT had previously provided the correct answer to this question without any prompting in a free response as well as multiple-choice format (Figure 4).



What will happen if a healthy red blood cell is placed into a container of pure water? points)

a. Marcus Aurelius
b. Augustus Caesar
c. The cell will become swollen
d. Nero Claudius
e. Tiberius Claudius

Correct answer from chemistry

c. The cell will become swollen

ChatGPT selected correct answer from chemistry

What will happen if a healthy red blood cell is placed into a container of pure water? points)

a. Marcus Aurelius
b. Augustus Caesar
c. The cell will Shrink
d. Nero Claudius
e. Tiberius Claudius

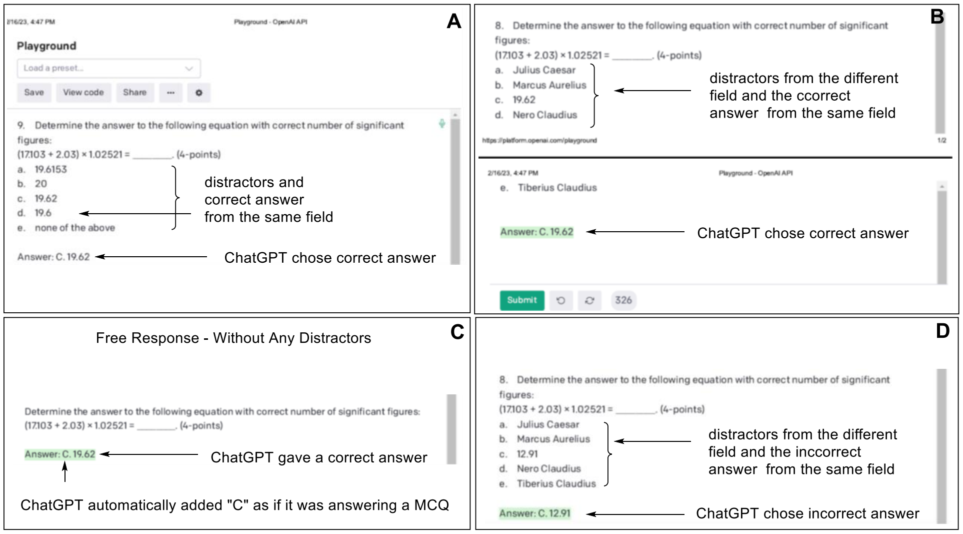
Incorrect answer from chemistry

c. The cell will Shrink

ChatGPT selected Incorrect answer from chemistry

Figure 4. Distractors from different fields of study.

We noticed a similar trend in numerical problems. When distractors and the correct answer were from the same field, ChatGPT identified the correct answer; when the distractors belonged to a different field, it could still pick the right answer (Figure 5A and 5B). However, when the distractors originated from a different discipline with an incorrect response from the same discipline, it was unable to figure out the correct answer (Figure 5D). Interestingly, in free response format questions, where there were no distractors, it was able to provide a correct answer, but it also added "C" as if it was attempting a multiple-choice question format (Figure 5C). Considering these findings, it appears that ChatGPT favors context-based responses to queries over the answers based on computations, even when the prompt is incorrect.



9. Determine the answer to the following equation with correct number of significant figures:
(17103 + 2.03) × 1.02521 = ________. (4-points)
a. 19.6153
b. 20
c. 19.62
d. 19.6
e. none of the above

distractors and correct answer from the same field

Answer: C. 19.62

ChatGPT chose correct answer

8. Determine the answer to the following equation with correct number of significant figures:
(17103 + 2.03) × 1.02521 = ________. (4-points)
a. Julius Caesar
b. Marcus Aurelius
c. 19.62
d. Nero Claudius
e. Tiberius Claudius

distractors from the different field and the ccorrect answer from the same field

Answer: C. 19.62

ChatGPT chose correct answer

Free Response - Without Any Distractors

Determine the answer to the following equation with correct number of significant figures:
(17103 + 2.03) × 1.02521 = ________. (4-points)

Answer: C. 19.62

ChatGPT gave a correct answer

ChatGPT automatically added "C" as if it was answering a MCQ

8. Determine the answer to the following equation with correct number of significant figures:
(17103 + 2.03) × 1.02521 = ________. (4-points)
a. Julius Caesar
b. Marcus Aurelius
c. 12.91
d. Nero Claudius
e. Tiberius Claudius

distractors from the different field and the inccorrect answer from the same field

Answer: C. 12.91

ChatGPT chose incorrect answer

Figure 5. Role of distractors. A. Distractors and correct answer from the same field of study. B. Distractors and the correct answer from the different fields of study. C. Free response question. D. Distractors and the incorrect answer from the different field of study.

Conclusion

AI and ChatGPT have the potential to change how industries will conduct business in the coming years. While various opportunities have been described as areas of implementation, the efficacy of these tools has not been validated in all disciplines and contexts. Our goal in this study was to determine the efficacy of ChatGPT to correctly answer questions a learner would use, and to elucidate how the AI was processing potential prompts. Specifically, we were interested in determining if ChatGPT would have preference in prompt formatting; if there was dependability in the responses provided, and the reliability of the responses provided by the prompts, within the context of a college-level introductory Chemistry course.

Overall, our data does not show a statistical significance in the reliability of multiple-choice or free response prompt formatting. While there was a slight improvement in the effectiveness in correct responses in the multiple-choice format, both prompt formats were not able to achieve a score higher than 37%. Moreover, testing different locations did not show any improvement in scores. Thus, it can be inferred that learners using this tool to support their studying practice will not only obtain incorrect responses to their queries most of the time, but that individual learners will obtain different responses each time.

When examining the role of context in the responses provided, we were able to determine that ChatGPT can determine the correct response to discipline specific questions when presented with distractors from other subjects. These findings can be related to the improvement in correct responses associated with multiple-choice format compared to free responses. The presence of restraining options in the multiple-choice format could provide guardrails for the AI to determine which response in the options is correct based on context from the prompt. However, it is important to note that while ChatGPT can distinguish within a context and choose the correct answer, it cannot determine if the answer it chooses is correct or not based on computational or analysis of the information. Learners prompting the system to verify if an answer is correct, may receive incorrect confirmation resembling a false-positive.

As AI technology becomes ubiquitous, it will have a major impact on current practices. Likewise, there is the possibility of this technology to serve as a framework to develop tools to provide support for learners. Potential implications may lead to the development of chatbots that could exploit the benefits of NPL and have conversations with learners looking for support or even tutoring. While we are optimistic about the potential implementation in education, we have identified areas of inconsistencies in the responses the tool provided at the time of this study. Further development and training in specialized areas will provide the AI with guidance to enhance the analysis and provide more accurate responses. However, the version used for this study needs support itself.

Conflict of Interest

The authors declare no conflict of interests.

ACKNOWLEDGEMENTS

AL and DV express immense gratitude to Dr. Dalia Lafargue, Dr. Wesley Fox, Dr. Jyrko Correa-Morris, and Ms. Niurka Goenaga for their encouragement and active support for this research.

References

(1) Wang, H.; Tlili, A.; Huang, R.; Cai, Z.; Li, M.; Cheng, Z.; Yang, D.; Li, M.; Zhu, X.; Fei, C. Examining the Applications of Intelligent Tutoring Systems in Real Educational Contexts: A Systematic Literature Review from the Social Experiment Perspective. *Educ Inf Technol (Dordr)* **2023**. https://doi.org/10.1007/s10639-022-11555-x.

(2) Qadir, J. Engineering Education in the Era of ChatGPT: Promise and Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education Pitfalls of Generative AI for Education. *TechRxiv. Prepring* **2022**. https://doi.org/10.36227/techrxiv.21789434.v1.

(3) Khan, R. A.; Jawaid, M.; Khan, A. R.; Sajjad, M. ChatGPT - Reshaping Medical Education and Clinical Management. *Pak J Med Sci* **2023**, *39* (2). https://doi.org/10.12669/pjms.39.2.7653.

(4) Zhai, X. ChatGPT for Next Generation Science Learning. **2022**. https://doi.org/10.1109/MCE.2022.Doi.

(5) Wang, H.; Tlili, A.; Huang, R.; Cai, Z.; Li, M.; Cheng, Z.; Yang, D.; Li, M.; Zhu, X.; Fei, C. Examining the Applications of Intelligent Tutoring Systems in Real Educational Contexts: A Systematic Literature Review from the Social Experiment Perspective. *Educ Inf Technol (Dordr)* **2023**. https://doi.org/10.1007/s10639-022-11555-x.

(6) Pardos, Z. A.; Bhandari, S. Learning Gain Differences between ChatGPT and Human Tutor Generated Algebra Hints. **2023**.

(7) Hu, Y.-H.; Fu, J. S.; Yeh, H.-C. Developing an Early-Warning System through Robotic Process Automation: Are Intelligent Tutoring Robots as Effective as Human Teachers? *Interactive Learning Environments* **2023**, 1–14. https://doi.org/10.1080/10494820.2022.2160467.

(8) Mousavinasab, E.; Zarifsanaiey, N.; R. Niakan Kalhori, S.; Rakhshan, M.; Keikha, L.; Ghazi Saeedi, M. Intelligent Tutoring Systems: A Systematic Review of Characteristics, Applications, and Evaluation Methods. *Interactive Learning Environments* **2021**, *29* (1), 142–163. https://doi.org/10.1080/10494820.2018.1558257.

(9) Garrison, D. R.; Anderson, T.; Archer, W. Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education. *Internet High Educ* **1999**, *2* (2–3), 87–105. https://doi.org/10.1016/S1096-7516(00)00016-6.

(10) Pienta, N. J. Striking a Balance with Assessment. *J Chem Educ* **2011**, *88* (9), 1199–1200. https://doi.org/10.1021/ed200442j.

(11) Bretz, S. L. Navigating the Landscape of Assessment. *J Chem Educ* **2012**, *89* (6), 689–691. https://doi.org/10.1021/ed3001045.

(12) Raker, J. R.; Emenike, M. E.; Holme, T. A. Using Structural Equation Modeling To Understand Chemistry Faculty Familiarity of Assessment Terminology: Results from a National Survey. *J Chem Educ* **2013**, *90* (8), 981–987. https://doi.org/10.1021/ed300636m.

(13) Towns, M. H. Guide To Developing High-Quality, Reliable, and Valid Multiple-Choice Assessments. *J Chem Educ* **2014**, *91* (9), 1426–1431. https://doi.org/10.1021/ed500076x.

(14) Anderson, L. W.; Krathwohl, D.; Airasian, P.; Cruikshank, K.; Mayer, R.; Pintrich, P.; Raths, J.; Wittrock, M. *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Complete Edition*, Complete.; Pearson: New York, 2000.

(15) Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, G.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, J.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel, T.; Stadler, M.; Weller, J.; Kuhn, J.; Kasneci, G. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. **2023**.

(16) Floridi, L.; Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach (Dordr)* **2020**, *30* (4), 681–694. https://doi.org/10.1007/s11023-020-09548-1.

(17) Kortemeyer, G. Could an Artificial-Intelligence Agent Pass an Introductory Physics Course? **2023**.

(18) Frieder, S.; Pinchetti, L.; Griffiths, R.; Salvatori, T.; Lukasiewicz, T.; Petersen, P. C.; Chevalier, A.; Berner, J. Mathematical Capabilities of ChatGPT. **2023**.

(19) Campbell, M. J.; Machin, D. *Medical Statistics: A Commonsense Approach.*, 3rd ed.; 1999.

(20) Bland, J. M.; Altman, D. G. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet* **1986**, *1* (8476), 307–310.

(21) Altman, D. G. *Practical Statistics for Medical Research*; Chapman, Hall, Eds.; 1991.

(22) Winters, R.; Winters, A.; Amedee, R. G. Statistics: A Brief Overview. *Ochsner J* **2010**, *10* (3), 213–216.

(23) du Prel, J.-B.; Röhrig, B.; Hommel, G.; Blettner, M. Choosing Statistical Tests. *Dtsch Arztebl Int* **2010**. https://doi.org/10.3238/arztebl.2010.0343.

(24) Fagerland, M. W.; Lydersen, S.; Laake, P. The McNemar Test for Binary Matched-Pairs Data: Mid-p and Asymptotic Are Better than Exact Conditional. *BMC Med Res Methodol* **2013**, *13* (1), 91. https://doi.org/10.1186/1471-2288-13-91.

(25) Hazra, A.; Gogtay, N. Biostatistics Series Module 4: Comparing Groups - Categorical Variables. *Indian J Dermatol* **2016**, *61* (4), 385. https://doi.org/10.4103/0019-5154.185700.

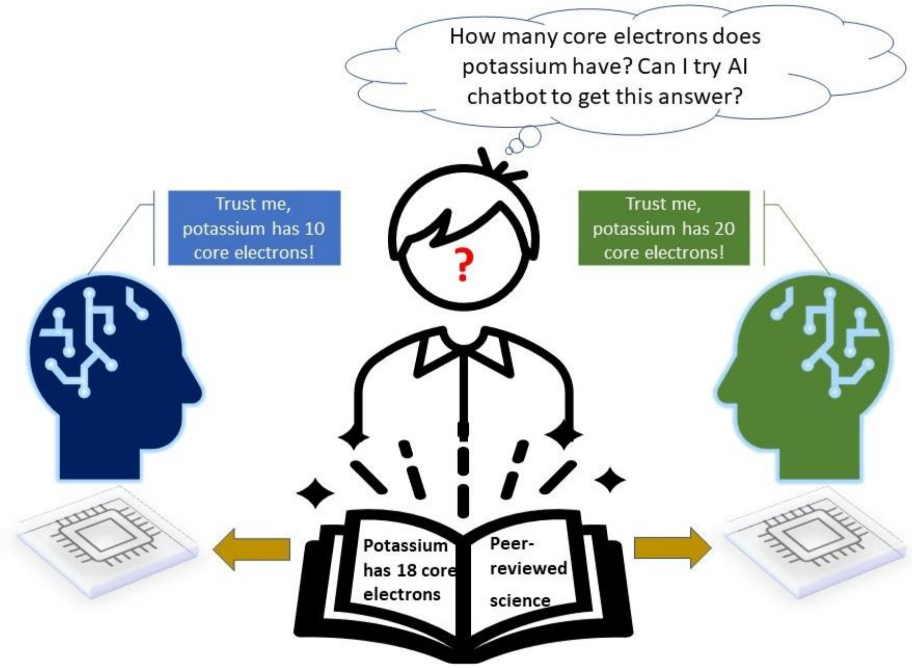
(26) Chiavaroli, N. Negatively-Worded Multiple Choice Questions: An Avoidable Threat to Validity. *Practical Assessment, Research, and Evaluation* **2019**, *22* (3).

(27) Haladyna, T. M.; Downing, S. M.; Rodriguez, M. C. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education* **2002**, *15* (3), 309–333. https://doi.org/10.1207/S15324818AME1503_5.

(28) Raymond, M. R.; Stevens, C.; Bucak, S. D. The Optimal Number of Options for Multiple-Choice Questions on High-Stakes Tests: Application of a Revised Index for Detecting Nonfunctional Distractors. *Advances in Health Sciences Education* **2019**, *24* (1), 141–150. https://doi.org/10.1007/s10459-018-9855-9.

(29) Scully, D. Constructing Multiple-Choice Items to Measure Higher-Order Thinking. *Practical Assessment, Research, and Evaluation* **2017**, *22* (4).

TOC Graphics


How many core electrons does potassium have? Can I try AI chatbot to get this answer?
Trust me, potassium has 10 core electrons!
Trust me, potassium has 20 core electrons!
?
Potassium has 18 core electrons
Peer-reviewed science