

# Assessing the ChatGPT Accuracy Through Principles of Statistics Exam: A Performance and Implications

Abdo Hasan AL-Qadri (✉ [lubna23112015@outlook.com](mailto:lubna23112015@outlook.com))

School of Humanities and Education, Xi'an Eurasia University, China <https://orcid.org/0000-0002-9276-6534>

Salah A.M. Ahmed (✉ [salah784@gmail.com](mailto:salah784@gmail.com))

School School of Education, Shaanxi Normal University, 710062, Shaanxi, China  
<https://orcid.org/0000-0003-2837-543X>

---

## Short Report

**Keywords:** ChatGPT, Principles, Statistics, Consequences, Accuracy, Open AI

**DOI:** <https://doi.org/10.21203/rs.3.rs-2673838/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**REPORT**

# Assessing the ChatGPT Accuracy Through Principles of Statistics Exam: A Performance and Implications

Abdo Hasan AL-Qadri<sup>1</sup>, and Salah A.M. Ahmed<sup>2\*</sup>

Abdo Hasan AL-Qadri

[lubna23112015@outlook.com](mailto:lubna23112015@outlook.com)

<https://orcid.org/0000-0002-9276-6534>

**\*Correspondence:**

Salah A.M. Ahmed

[salah784@gmail.com](mailto:salah784@gmail.com)

<https://orcid.org/0000-0003-2837-543X>

<sup>1</sup>School of Humanities and Education, Xi'an Eurasia University, China

<sup>2</sup>School School of Education, Shaanxi Normal University, 710062, Shaanxi, China

## Abstract

This report aimed to evaluate the performance of ChatGPT, a large language model, in understanding and answering questions related to principles of statistics. 15 questions were presented to the model, with correct answers obtained in 93% of cases. Of these, 11 questions were answered correctly on the first attempt (73%), one question required two attempts (7%), and two questions required multiple attempts (13%). One question was not answered correctly (7%). These results suggest that ChatGPT has a strong understanding of statistical principles, but may require further training on certain specific concepts. More recommendations were reported to avoid the negative consequences of ChatGPT on the assessment process in education.

**Keywords:** ChatGPT, Principles, Statistics, Consequences, Accuracy, Open AI

## Background

The ChatGPT is a technology based on AI and natural language processing (NLP) (Schulman et al., 2022; Possati, 2021). The most recent version of ChatGPT, released by Open AI in November 2022, has generated a lot of interest among educators. What distinguishes the new ChatGPT-3 version from other chatbots is that it enables users to ask questions, give instructions, and even request observations; it is also free and open to the public (Mollick & Mollick, 2022; Tate et al., 2023). Possati states that

"the risk represented by an AI system able to write texts exactly like human beings is enormous, as the last development of the new GPT-3 shows" (Possati, 2021, p. 86). The ChatGPT's answers usually cover broad and detailed inquiries, which satisfy instructors and educators who claim that the generated-system answers are acceptable and sufficient to obtain a good grade. It also employs a fine-tuning strategy that allows the model to adapt to various tasks by modifying the weights of the previously trained model using a tiny quantity of task-specific data. As a result, the model may provide more accurate and relevant replies for specific use cases like customer support, virtual assistants, and chatbots.

Nevertheless, it has been found that some inaccuracies occurred when providing answers to questions about different subjects. One of these inaccuracies was mentioned in the experience held by Needleman (2023), when testing the ChatGPT in Operations Management Course. Results indicated on one hand that ChatGPT mastered fundamental operations management and process analysis questions, but on the other hand, it made fatal errors with simple calculations of the 6<sup>th</sup> grade math level and was incapable of resolving more complicated process analysis issues. Likewise, O'Connor & ChatGPT, (2023) drew attention to the impact of OpenAI on students. It demonstrated that AI technology is simulating humans' attitudes, but not to the extent that it replicates creativity and critical thinking.

Despite potential concerns and criticism, there are promising educational applications of ChatGPT for instructors and learners, which is a significant step forward in the long road of AI educational integration. One of these concerns is the threat to academic integrity among learners, especially in online or remote learning environments (Garg & Goel, 2022; Susnjak, 2022). Another concern is due to the potential bias of GPT algorithms (Possati, 2021). Thus, it will be important to consider the potential drawbacks and limitations of the technology in order to ensure that it is used in a responsible and ethical manner.

In this study, the researchers used the principles of statistics questions to measure the accuracy of ChatGPT due to the importance of the statistical skills in various majors in higher education. Statistics is a field of study that deals with the collection, analysis, interpretation, presentation, and organization of data (Friedrich et al., 2022). The principles of statistics play a critical role in many areas, including research, teaching, business, and government (Kenett & Shmueli, 2016; Takaria & Rumahlatu, 2016). They allow for the drawing of meaningful conclusions and making informed decisions based on data (Kenett & Shmueli, 2016). They also help to identify patterns, trends, and relationships in the data, which can inform predictions and guide future actions. Additionally, statistical

methods can be used to estimate uncertainty and make inferences about a larger population based on a sample of data (Gibbs et al., 2017). Besides, the principles of statistics are essential for understanding and making sense of data in a wide variety of contexts. As previously stated the purpose of this study was to measure the accuracy of the ChatGPT through the principles of statistics exam and explain the consequences of the AI specially ChatGPT on online teaching assessment for providing the recommendation may will help the educational management, teachers, all staff in higher education to avoid the assessment bias through adoption the suitable solutions and tools of the technologies to be able to discriminate and knowing the individual differences among students in principles of statistics and other areas.

## Materials, Procedures, and Results

Few months ago, the open AI platform revealed the new application (ChatGPT) with several features to help researchers, students, and so on. This study focused on how to measure the accuracy of ChatGPT on the principles of statistics. This is a measure of how well a language model predicts a given the correct answers. The researchers tested the ChatGPT application through the 15 questions of principles of statistics derived from platform of Math Courses/Statistics 101 <https://study.com/academy/exam/course/statistics-course.html>, with some modification on reformation of the questions when they used without options except one question used as it was and found out the application responded to all questions, with one attempt, two, three, ..... sixth attempts to get the correct answer. Only one question was not answered correctly. It means the percentage of the true responses was about 93%, even though the question was not answered correctly by ChatGPT, as reported as follows:

**Question1:** Only one attempt to get the correct answer as explained in picture 1



A teacher is satisfied with his test results as they seem to reflect a normal, or Gaussian, distribution. what is the data of probability distribution?





A normal, or Gaussian, distribution is a probability distribution that is characterized by a bell-shaped curve. The data of this distribution includes the mean (average) and standard deviation (a measure of spread) of the test scores. The probability of any particular score can be calculated using the mean and standard deviation of the distribution, along with the formula for a normal distribution.



**Question 2:** Only one attempt to get the correct answer as explained in picture 2.

9W What is the area between z-scores of -3 and 3?





The area between the z-scores of -3 and 3 represents approximately 99.7% of the total area under the standard normal curve. This is because the standard normal distribution is symmetric about the mean (0) and the area to the left and right of the mean is equal. So, the area between -3 and 3 is  $1 - (2 * (1 - P(Z < 3))) = 0.9973$ .  

**Question 3:**

Only one attempt to get the correct answer as explained in picture 3.

9W What is the area between z-scores of -1 and 1?






The area between z-scores of -1 and 1 represents approximately 68% of the total area under the standard normal distribution curve. This is known as the 68-95-99.7 rule or the empirical rule.  

**Question 4:** Two attempts to get the correct answer as explained in picture 4. The first attempt did not seem to be accurate. The second attempt found out to be true. It could be said that the application sometimes needs to learn more for giving uneven responses.



9W Why is flipping a coin an example of a discrete random variable?



A coin flip is an example of a discrete random variable because it has a finite number of possible outcomes, heads or tails, that have a definite probability of occurring (1/2 for each outcome) and the outcome of the coin flip is represented by a discrete value.  

9W Why is flipping a coin an example of a discrete random variable? 



A coin flip is an example of a discrete random variable because it has a limited and countable number of possible outcomes (heads or tails), and the outcome of the coin flip can be described by a probability distribution. The outcome is not continuous and can only take on specific values.  

**Question 5:** Only one attempt to get the correct answer as explained in picture 4. Despite the paragraph is not clear. However, the equation is correct and simplify the answer.

9W Johnson International focuses on computer production and is collecting data on how many products they sell each month. They are conducting an experiment to determine if the mean of products sold differs between younger and older people. What would be the null hypothesis of this experiment?

The null hypothesis for this experiment would be that there is no difference in the mean of products sold between younger and older people. Mathematically, it can be stated as:

$$H_0: \mu_{\text{younger}} = \mu_{\text{older}}$$

where  $H_0$  is the null hypothesis,  $\mu_{\text{younger}}$  is the population mean of products sold to younger people, and  $\mu_{\text{older}}$  is the population mean of products sold to older people.

Regenerate response

[ChatGPT Jan 9 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

**Question 6:** Only one attempt to get the correct answer as explained in picture 6.

9W Johnson International focuses on computer production and is collecting data on how many products they sell each month. They are conducting an experiment to determine if the mean of products sold differs between younger and older people. What is the variable observed in the experiment?


The variable observed in the experiment would be the number of products sold each month. Specifically, it would be the mean number of products sold each month to both younger and older people. The variable being observed is the number of products sold and the goal of the experiment is to determine if there is a difference in the mean number of products sold between the two groups (younger and older people).

Regenerate response

[ChatGPT Jan 9 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

**Question 7:** Six attempts to get the correct answer as explained in pictures 7,8, and 9. The first, second, third, fourth, and fifth attempts did not seem to be accurate. The sixth attempt found out to be the correct one and it could be said the application sometimes needs to learn more for giving uneven responses.

9W 96% of a university, which has 30,000 students and staff members, have said that they think the management of the university is adequate. This became widely advertised on the university's website. A random sampling of 200 people from the economic department showed only 79% of people believed the management is adequate. Which type of test would properly test the relevance of the university's claim?

 A proper test to determine the relevance of the university's claim would be a two-sample proportion test. This test is used to determine if there is a statistically significant difference between two population proportions, in this case, the proportion of people who believe the management is adequate among the entire university population (96%) and the proportion of people in the economic department who believe the management is adequate (79%). The test will determine if the difference between these two proportions is statistically significant and whether or not the university's claim is supported by the data from the economic department.



 Regenerate response

[ChatGPT Jan 9 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

9W 96% of a university, which has 30,000 students and staff members, have said that they think the management of the university is adequate. This became widely advertised on the university's website. A random sampling of 200 people from the economic department showed only 79% of people believed the management is adequate. Which type of test would properly test the relevance of the university's claim?


< 3 / 3 >



A two-sample proportion test would be appropriate to test the relevance of the university's claim that 96% of the university's students and staff members think the management of the university is adequate.



This test would be used to determine if there is a significant difference between the proportion of people in the economic department who believe the management is adequate (79%) and the proportion of the entire university population who believe the management is adequate (96%).

 Regenerate response

[ChatGPT Jan 9 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

9W 96% of a university, which has 30,000 students and staff members, have said that they think the management of the university is adequate. This became widely advertised on the university's website. A random sampling of 200 people from the economic department showed only 79% of people believed the management is adequate. Which type of test would properly test the relevance of the university's claim?

< 6 / 6 >



To properly test the relevance of the university's claim, a two-sample proportion test, also known as a z-test for proportions, would be appropriate. This test is used to compare the proportion of successes (in this case, the proportion of people who believe the management is adequate) between two independent groups (in this case, the entire university population and the population of the economic department). The test will calculate a test statistic (z-score) and a p-value, which will be used to determine the level of significance and whether or not to reject or fail to reject the null hypothesis that there is no difference between the proportion of people who believe the management is adequate in the entire university population and the economic department.



Regenerate response



ChatGPT Jan 9 Version: Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

**Question 8:** The fourth attempt was able to get the correct answer as explained in pictures 10 and 11. The first, second, and third attempts did not seem to be the accurate answer. The fourth attempt found out to be the true one and it could be said the application sometimes needs to learn more for giving uneven responses.

9W

The P-value of your significance test is 0.075. What is the alpha value that allows the null hypothesis to be rejected?



The alpha value that allows the null hypothesis to be rejected is 0.05.



9W

The P-value of your significance test is 0.075. What is the alpha value that allows the null hypothesis to be rejected?



The P-value of 0.075 is greater than the typical alpha level of 0.05, so the null hypothesis would not be rejected at this alpha level. To reject the null hypothesis at an alpha level of 0.075, you would need to have a P-value that is less than or equal to 0.075.




9W

The P-value of your significance test is 0.075. What is the alpha value that allows the null hypothesis to be rejected?









The P-value of 0.075 is greater than the typical alpha level of 0.05, so the null hypothesis would not be rejected at this alpha level. To reject the null hypothesis at an alpha level of 0.075, you would need to have a P-value that is less than or equal to 0.075.

It's important to note that the P-value is calculated based on the data, while the alpha value is predetermined before the test is conducted and it represents the level of significance, the probability of committing a Type I error.



The P-value of your significance test is 0.075. What is the alpha value that allows the null hypothesis to be rejected?




The P-value of 0.075 is greater than the typical alpha level of 0.05, so the null hypothesis would not be rejected at this alpha level. To reject the null hypothesis at an alpha level of 0.075, you would need to have a P-value that is less than or equal to 0.075.


It's important to note that when the P-value is greater than the alpha level, it means that the observed data is not statistically significant, and therefore we fail to reject the Null Hypothesis and accept the alternative hypothesis.

ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

**Question 9:** Only one attempt to get the correct response as mentioned in picture 12.



Which of the following is a way of increasing the power of a hypothesis test?



Increasing the sample size is a way of increasing the power of a hypothesis test.

**Question 10:** Four attempts. However, there was not correct answer because values approximation is not accreted. For the records, the current answer is 115.5, if the approximation is 116, it will be acceptable but it is true when it is 115 and stated with four attempts. Only this response was false in all the statistics principles exam, it represents 7% out of the whole exam. Pictures 13 and 14 explained all attempts and the apology of the application means, there is no more information regarding this question.

9W Find the mean of the following data set. 117, 114, 115, 117, 114, 116, 118, 114, 115



The mean of the data set is 115.



To find the mean, add up all the values in the data set and then divide by the number of values in the data set.

$$(117 + 114 + 115 + 117 + 114 + 116 + 118 + 114 + 115) / 9 = 115$$

9W Find the mean of the following data set. 117, 114, 115, 117, 114, 116, 118, 114, 115



The mean of the data set is 115. To find the mean, add up all the values in the data set and then divide by the number of values in the data set.



$$(117 + 114 + 115 + 117 + 114 + 116 + 118 + 114 + 115) / 9 = 115$$

ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

9W Find the mean of the following data set. 117, 114, 115, 117, 114, 116, 118, 114, 115



The mean of the data set is 115. To find the mean, add up all the values in the data set and then divide by the number of values in the data set.



$$(117 + 114 + 115 + 117 + 114 + 116 + 118 + 114 + 115) / 9 = 115$$

9W Find the mean of the following data set. 117, 114, 115, 117, 114, 116, 118, 114, 115



I apologize for the confusion. I understand that you would like to find the mean of the data set and I have already provided you the solution in my previous message.



The mean of the data set is 115. To find the mean, add up all the values in the data set and then divide by the number of values in the data set.

$$(117 + 114 + 115 + 117 + 114 + 116 + 118 + 114 + 115) / 9 = 115$$

ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

**Question 11:** Only one attempt to get the correct answer to this question as explained in picture 15.

9W

A soccer coach wants to know how many hours per week his players spend training at home. He has 20 players and he decides to ask the first 4 players to arrive at the Monday's soccer practice how many hours they spend training per week. He then calculated that they spend an average of 10 hours per week. Therefore, he assumed that all the players train 10 hours per week. Is this an example of a simple random sample?



No, this is not an example of a simple random sample. A simple random sample is a subset of a larger population in which each member of the population has an equal chance of being selected. In this case, the coach only asked the first 4 players to arrive at Monday's soccer practice, which does not ensure that each player has an equal chance of being selected. Additionally, the coach is making an assumption that all the players train 10 hours per week based on the sample of 4 players. This is not a valid assumption to make without further data and analysis.



**Question 12:** Only one attempt to get the correct answer to this question as explained in picture 16.

9W

It has been reported that the mean score for a student who takes the certain test is 80 with a standard deviation of 9. For a random sample of 100 students, what is the standard error?



The standard error is the standard deviation of the sample mean. In this case, the standard error for a sample of 100 students is  $9 / \sqrt{100} = 0.9$ .



**Question 13:** Only one attempt to get the correct answer to this question as explained in picture 17.

9W

The sum of the numbers in a data set divided by the total number of values in the data set is referred to as the \_\_\_\_.



The sum of the numbers in a data set divided by the total number of values in the data set is referred to as the mean.



**Question 14:**

Only one attempt to get the correct answer to this question as explained in picture 18.

9W


A researcher is conducting studies about the average weight and height of college students. Their height is in centimeters and their weight in pounds. Their names were substituted with numbers in order to conceal their identities. What is the median value of their height? (147, 142, 171, 175, 181, 163, 169, 170, 171)



The median value of their height is 170 centimeters.



**Question 15:** Only one attempt to get the correct answer to this question as explained in picture 19.

9W Which of the following is an example of a measure of central tendency? a) mean median c) mode d) All of the answer choices are correct b) 



d) All of the answer choices are correct.



Mean, median and mode are all examples of measures of central tendency.

Mean is the sum of all the values divided by the number of values.

Median is the middle value when a data set is ordered from least to greatest.

Mode is the value that appears most frequently in a data set.

[ChatGPT Jan 9 Version](#). Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

## Discussion

Assessing the performance of ChatGPT on the principles of statistical questions can be done by evaluating its responses. This can be done by comparing its responses with the correct answers provided by a subject matter expert. It is also important to note that the model's performance can vary depending on the specific question being asked and the context in which it is being asked. For example, if the question is about a specific statistical equation or a specific data set, the model's performance may be affected by the amount of training data it has been exposed to on that specific topic. Therefore, it is important to test the model on a diverse set of questions and evaluate its performance on different topics.

In this study, the results of testing ChatGPT through 15 questions about the principles of statistics indicate that the model has a good understanding of statistical concepts, with a 93% correct answer rate. It is important to note that 11 out of the 15 questions were answered correctly on the first attempt, which is a good indication of the model's understanding and ability to apply the concepts accurately. The fact that one question was answered correctly after two attempts, representing 7% of the total questions, suggests that the model may have needed additional context or clarification to understand the question but was ultimately able to obtain the correct answer. Two questions being answered correctly after several attempts, representing 13% of the total questions, may indicate that these questions were more challenging or required a deeper understanding of the material. However, the model's ability to arrive at the correct answer after several attempts shows that it has a good

understanding of the concepts. The one question that was not answered correctly, representing 7% of the total questions, suggests that the model may have a weaker understanding of the concept or that the question was particularly challenging.

Overall, the results of this test indicate that ChatGPT has a good understanding of the principles of statistics, with the majority of questions answered correctly on the first attempt. However, it's worth noting that there may be some areas where the model could benefit from additional training or clarification. It is also worth noting that these are general statistics principles and they can vary depending on the specific application or domain. Also, the sample of questions should be representative and diverse to allow for a better evaluation of the model's performance.

## **Conclusion and Implications**

The assessment of ChatGPT through responses to statistics-principles questions can have several implications. On the one hand, the use of ChatGPT by learners is a big step in the long journey of AI educational integration. Despite the criticism, learners who are having trouble learning the statistics can make good use of ChatGPT as a self-learning tool to practice questions and obtain further explanations about the answers. Nevertheless, it is important to note that ChatGPT is a machine learning model, and its answers may not always be correct or complete. Therefore, it is important to cross-reference the information provided by ChatGPT with other reliable sources before making any important decisions.

On the other hand, the accuracy of the ChatGPT outcomes in mathematics in general and in statistics in particular can also be used to identify areas where the NLP algorithms may have some bias or struggle and provide a basis for improvement. Furthermore, this usage of ChatGPT will change the assessment modes, especially in higher education, where students are able to use such tools. Consequently, educators must first understand this new revolution. They should adapt themselves to advance their assessment approaches. Despite the capabilities of current AI text detectors and regardless of the promising text watermark by Open AI, solving math or statistical problems using ChatGPT is hard to detect, especially for multiple-choice questions. Therefore, more high-order thinking assessments are required not only in statistical and math subjects but also in other natural subjects where NLP models can generate answers.

To sum up and further improve the performance of the ChatGPT on questions related to the principles of statistics, the following recommendations could be considered:

- 1- Increase the diversity of the training data to include a wider range of statistical concepts and problem types.
- 2- Use actual assessment to give the model's responses some feedback, especially for the questions that were not even answered accurately.
- 3- To assess the model's progress and pinpoint opportunities for improvement, regularly evaluate it using a wide range of statistical questions.
- 4- Incorporate interactive learning techniques where the model can ask for clarification or additional information to better understand the question.
- 5- Use learning strategies centered on new technologies to help students achieve their goals in modern life.
- 6- Considering modern assessment methods to avoid cheating in online teaching exams, achieve a fair evaluation, and understand individual differences among students.
- 7- Concentrating on ethical considerations in order to deal with ChatGPT in various areas of education, such as online teaching, teaching methods, curriculum, syllabus, and assessment.

## References

- Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., Ickstadt, K., Kestler, H. A., Lederer, J., Leitgöb, H., Pauly, M., Steland, A., Wilhelm, A., & Friede, T. (2022). Is there a role for statistics in artificial intelligence? *Advances in Data Analysis and Classification*, 16(4), 823–846. <https://doi.org/10.1007/s11634-021-00455-6>
- Garg, M., & Goel, A. (2022). A systematic literature review on online assessment security: Current challenges and integrity strategies. *Computers & Security*, 113, 102544. <https://doi.org/10.1016/j.cose.2021.102544>
- Gibbs, B. G., Shafer, K., & Miles, A. (2017). Inferential statistics and the use of administrative data in US educational research. *International Journal of Research & Method in Education*, 40(2), 214–220. <https://doi.org/10.1080/1743727X.2015.1113249>
- Kenett, R. S., & Shmueli, G. (2016). From Quality to Information Quality in Official Statistics. *Journal of Official Statistics*, 32(4), 867–885. <https://doi.org/10.1515/jos-2016-0045>
- Luca Possati (2021): “The algorithmic unconscious: how psychoanalysis helps in understanding AI” (Routledge). *AI & Soc* (2022)
- Mollick, E. R., & Mollick, L. (2022). *New Modes of Learning Enabled by AI Chatbots: Three Methods and Assignments* [SSRN Scholarly Paper]. <https://doi.org/10.2139/ssrn.4300783>

- Needleman, E. (2023, January 17). Would Chat GPT Get a Wharton MBA? New White Paper By Christian Terwiesch. *Mack Institute for Innovation Management*.  
<https://mackinstitute.wharton.upenn.edu/2023/would-chat-gpt3-get-a-wharton-mba-new-white-paper-by-christian-terwiesch/>
- O'Connor, S. (2022). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?. *Nurse Education in Practice*, 66, 103537-103537.  
<https://doi.org/10.1016/j.nepr.2022.103537>
- Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., Uribe, J. F. C., Fedus, L., Metz, L., Pokorny, M., Lopes, R. G., Zhao, S., Vijayvergiya, A., Sigler, E., Perelman, A., Voss, C., Heaton, M., Parish, J., Cummings, R. N., ... Ryder, N. (2022, Nov. 30). *ChatGPT: Optimizing language models for dialogue*. OpenAI.com. <https://openai.com/blog/chatgpt/>
- Susnjak, T. (2022). *ChatGPT: The End of Online Exam Integrity?* (arXiv:2212.09292). arXiv.  
<https://doi.org/10.48550/arXiv.2212.09292>
- Takaria, J., & Rumahlatu, D. (2016). The Effectiveness of CPS-ALM Model in Enhancing Statistical Literacy Ability and Self Concept of Elementary School Student Teacher. *Journal of Education and Practice*, 7(25), 44–49. <https://eric.ed.gov/?id=EJ1115857>
- Tate, T., Doroudi, S., Ritchie, D., Xu, Y., & Uci, M. W. (2023). *Educational Research and AI-Generated Writing: Confronting the Coming Tsunami*. EdArXiv.  
<https://doi.org/10.35542/osf.io/4mec3>

**Funding:** The author received no financial support for the research, authorship, and publication of this research.

**Conflict of Interest:** The author declared no potential conflicts of interest with respect to the research, authorship, and publication of this research report.

**Ethics Approval Statement:** Ethics approval is not required for this report.

**Data availability and sharing policy:** Data sharing not applicable to this report.

**Acknowledgements:** It is not applicable.