


Addressing mistakes of ChatGPT from an educational perspective



 **Chi Hong Leung**¹⁺
 **Winslet Ting Yan**
Chan²

^{1,2}Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong, The People's Republic of China.

¹Email: msleung@polyu.edu.hk

²Email: winslet.ty.chan@polyu.edu.hk



(+ Corresponding author)

ABSTRACT

Article History

Received: 13 December 2024

Revised: 20 January 2025

Accepted: 4 February 2025

Published: 14 February 2025

Keywords

Artificial intelligence

ChatGPT

Educational application

Large language model

Mistakes.

This paper explores the efficacy of ChatGPT, a generative artificial intelligence in educational contexts, particularly concerning its potential to assist students in overcoming academic challenges while highlighting its limitations. ChatGPT is suitable for solving general problems. When a student comes across academic challenges, ChatGPT may conveniently provide a solution. However, ChatGPT may create mistakes in its responses that become detrimental to the learning process. The study discusses the technical principles underlying ChatGPT and identifies factors that contribute to incorrect responses. It categorizes various errors such as outdated information, mathematical inaccuracies and cognitive biases illustrating these with concrete examples. The analysis reveals that although ChatGPT is trained on extensive datasets, it can produce statistically plausible but factually incorrect answers. This can hinder the learning process and lead students astray if they rely solely on its outputs. The paper offers practical recommendations for students and teachers to mitigate the risks associated with using ChatGPT. These include verifying sources of information, formulating questions in diverse formats, supplementing queries with additional context, and reporting inaccuracies to OpenAI. Implementing these strategies can enhance the educational utility of ChatGPT while minimizing potential pitfalls. Understanding both the strengths and weaknesses of ChatGPT is a proper way of using a new technology efficiently.

Contribution/ Originality: This paper examines the common mistakes made by ChatGPT and explains their implications for users who may mistakenly assume that AI-generated responses are typically accurate. Additionally, it offers recommendations for effectively using ChatGPT to mitigate these issues from an educational standpoint.

1. INTRODUCTION

ChatGPT has been a popular artificial intelligence (AI) topic recently. Deep learning technology is used to learn from various textual data and ChatGPT generates relevant natural language texts as outputs. It is so simple to use ChatGPT that the user only needs to enter a text in the dialogue and ChatGPT will automatically analyze and generate related responses. ChatGPT can be used in a variety of scenarios (Paul, Ueno, & Dennis, 2023; Teubner, Flath, Weinhardt, van der Aalst, & Hinz, 2023; Van Dis, Bollen, Zuidema, Van Rooij, & Bocking, 2023). First, it can be used to develop chatbots allowing users to interact with robots through dialogues. This can be applied to customer service, questions and answers (Q and A) systems, and chat applications. Moreover, ChatGPT can automatically generate article abstracts, read articles and understand their content, extract the most important information from them and then generate concise article abstracts. In addition, ChatGPT can automatically

translate texts between different languages and then realize automatic translation by learning translation rules and language models between different languages. Furthermore, ChatGPT can be used to help write, generate paragraphs and create paragraph titles on various topics and help writers complete article writing. Similarly, ChatGPT can perform emotional analysis and complete the analysis and evaluation of text emotions by learning the relationship between emotional words and corresponding emotional expressions.

However, users of AI usually misunderstand that AI should not generate mistakes. ChatGPT can produce errors and it is an inherent challenge in AI. ChatGPT has an error rate of up to 50%, a finding from the latest research from Purdue University (Kabir, Udo-Imeh, Kou, & Zhang, 2023). However, many engineers still prefer answers provided by AI. This phenomenon has sparked widespread discussion and concern. The study mentioned above has pointed out that ChatGPT has a whopping error rate of 52% in answering programmatic questions. These data show that ChatGPT still has a lot of room for improvement of accuracy and reliability compared to human experts. The error type analysis study noted that 52% of the wrong answers contained false information, 77% were more verbose than human answers and 78% were inconsistent with human answers in some way. These data reveal the limitations of AI when it comes to dealing with complex problems. Many engineers still choose to use ChatGPT's answers despite the high error rate. The research team surveyed 12 engineers and found that 35% of them favored AI-generated answers, and 39% were not even aware of errors in AI answers.

ChatGPT has been adopted and used in educational institutions (Fergus, Botha, & Ostovar, 2023; Gordijn & Have, 2023; Kohnke, Moorhouse, & Zou, 2023). Teachers use ChatGPT to prepare instructional materials and facilitate the teaching process in the classroom. Students also use ChatGPT to enhance their learning experience such as personalized self-learning. However, when ChatGPT provides an incorrect answer to a user (teacher or student), it may lead to a detrimental result in both teaching and learning (Zhou, Ke, Qiu, Huang, & Zhang, 2023). For example, a teacher may ask ChatGPT to prepare historical information on a particular issue for revision in the class but ChatGPT may generate inaccurate facts. A student may ask ChatGPT to illustrate how to solve a mathematical problem in a step-by-step manner but ChatGPT may show wrong steps for students to follow. This paper will first discuss the technical principles of ChatGPT and examine the factors contributing to mistakes made by ChatGPT. Then, various types of these mistakes will be illustrated with examples. Finally, the paper will suggest the proper ways of using ChatGPT in the educational context to take its advantages but minimize its drawbacks.

2. TECHNICAL PRINCIPLES OF CHATGPT

In the 1950s, scientists began to design machines that could simulate human intelligence trying to realize that machines could perform reasoning, learning, questions and answers and other tasks like humans. Neural networks first originated in the 1950s and 1960s and received wide attention in the 1980s and 1990s. However, at that time, the amount of computing resources and data were limited and the neural network could not give full play to its potential (Cios et al., 2018). Neural networks gradually revived at the beginning of the 21st century with the increasing amount of data and computing resources. In 2006, one of the representative models of deep learning, the Deep Belief Network (DBN) was proposed laying the foundation for the development of deep learning (Hinton, Osindero, & Teh, 2006).

In 2018, there was an important moment for the natural language processing model to make a large-scale breakthrough. The Large Language Model (LLM) has become a popular research theme in natural language processing due to the fast development of deep learning technology and the popularization of a huge amount of text data. At that time, Google and OpenAI proposed their models based on deep learning. Their models have achieved the best results in many natural language processing tasks and are widely used in search engines, speech recognition, machine translation, question- and -answer systems and other fields. In 2020, the GPT-3 model was

launched by OpenAI becoming the largest natural language processing model at that time and its ability was enough to generate human text that could not be distinguished (Cain, 2024).

The GPT (Generative Pre-training Transformer) model, one of LLMs belongs to generative AI which can generate its own output without much manual intervention. There are pre-training and fine-tuning stages in the training method of the GPT model (Ng, Tan, & Leung, 2024). In the pre-training stage, GPT learns from a lot of textual data through self-supervision. It understands vocabulary, grammar rules, contexts and other features of natural language. It predicts the next word after learning syntactical and semantical relationships among words without manual tagging or marking in advance. In the fine-tuning stage, GPT is fine-tuned with some marked texts to adapt to particular tasks. It uses the learned model to better comprehend the context and create new outputs with better generation capabilities.

GPT automatically learns features such as grammar, vocabulary and context of natural language in many text data that have never been marked. This self-supervised training method does not require manual marking which greatly reduces the training cost of GPT and better processes diverse tasks. This technology achieves good results in natural language generation (Zuckerman et al., 2023).

3. FACTORS CONTRIBUTING TO MISTAKES MADE BY CHATGPT

There are many factors contributing to mistakes made by ChatGPT. The first cause is its training data. ChatGPT was trained on a huge amount of data. If the data are wrong or outdated, the outcomes will contain errors consequently. In addition, ChatGPT may interpret data in a skewed manner. Its conclusions may only be statistically correct but do not make sense.

A neural network is the core of ChatGPT. It is called the transformer that is adaptable and robust but with its pitfalls. A neural network identifies patterns among input data. When these patterns are misleading, an overgeneralized result will be obtained. For example, if the information is wrong but found repeatedly in the data sources, it will become a valid pattern recognized by a neural network. This results in the wrong answer given by ChatGPT (Fujimoto & Takemoto, 2023).

The accuracy of AI is based on the training data. The performance of an AI model is only as good as the training data. ChatGPT is trained on the combination of Internet data, licensed data and those prepared by human developers. ChatGPT has a very wide knowledge base that may contain inaccuracies, biases and outdated information. Consequently, the reliance of ChatGPT on this data may lead to incorrect or outdated responses. Humans may make mistakes because of insufficient information and cognitive biases. Similarly, ChatGPT has its problems. Although it can process data rapidly, it is not aware of the context and does not have cognitive capability. ChatGPT may misinterpret and overlook some concepts related to ethical, emotional and cultural issues (Rahimi & Abadi, 2023).

There is a requirement for comprehensive answers and accuracy. However, there are the trade-offs. When the knowledge base of the model is more comprehensive, it becomes harder to make sure that information is correct and updated. If it is to implement stricter measures of safety, the model may be overcautious and it avoids certain questions that it could address properly. It is expected that AI can interact and behave like human beings but human-like behaviors are associated with human-like mistakes. Striking a balance between them is a challenge.

ChatGPT and every AI model learn from numerous training data. To ensure bias-free data is almost impossible and the AI model will answer incorrectly. ChatGPT has a knowledge cutoff data. After this date, ChatGPT is not aware of the information in the world. It may not have the latest information, and this leads to wrong or outdated viewpoints. In addition, on the Internet, there is contradictory information. To determine the correctness of data is a difficult task. Therefore, sometimes ChatGPT generates less correct data. ChatGPT learns to predict the next word in a sentence based on the training data. This approach only creates contextually correct

sentences but it cannot make sure the factual correctness. ChatGPT must process data across a variety of topics to make it useful. But the more the topics, the more challenging it ensure accuracy in each topic (Rahman & Watanobe, 2023).

ChatGPT also avoids taking strong viewpoints on sensitive topics. It tries to strike the balance between extreme answers. But this may lead to non-committal answers. When the training data are really biased on a controversial subject, ChatGPT only reflects the bias simply. As a safety measure, ChatGPT may provide generic answers to sensitive topics. However, it is sometimes considered offering incomplete answers (Kwak & Pardos, 2024).

4. EXEMPLARS OF MISTAKES

There are several types of mistakes made by ChatGPT. This section discusses these mistakes and attempts to explain the possible reasons. In all the demonstrations in this section, ChatGPT-3.5-Turbo (<https://poe.com/>) was used because of its popularity.

4.1. Outdated Responses

The first type of mistake is related to the cutoff date and update date of collected training data. If the correctness of responses is time-sensitive, ChatGPT usually informs the user of this important issue. For example, when the question is about the population size of a city (see Figure 1), ChatGPT actively declares the knowledge update time and reminds the user of checking the provided answer.

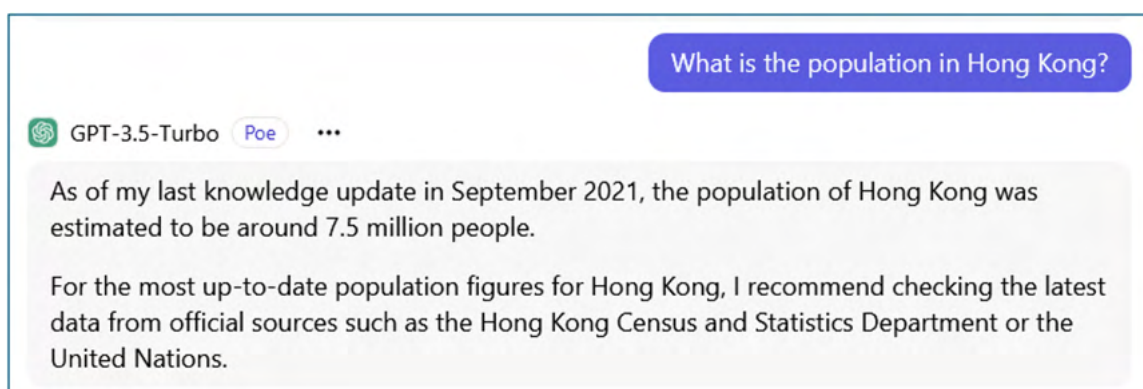


Figure 1. A question about the population of a city.

Similarly, when the question is about a commercial product (see Figure 2), ChatGPT also suggests checking the official website of the relevant company for the latest information.

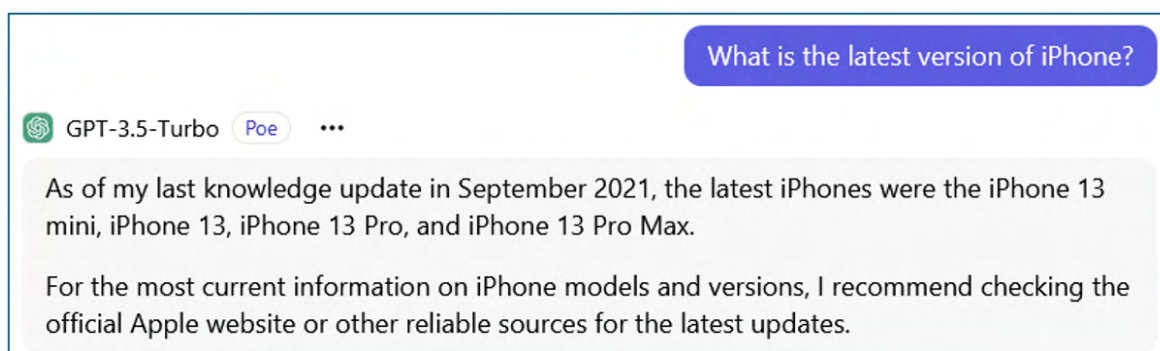


Figure 2. A question about a commercial product.

4.2. Mathematical Faults

However, in most situations, when ChatGPT makes a mistake, the user may not be aware of it because there is no reminder or warning provided by ChatGPT. Unless the user aggressively checks the answer, the mistake will not be discovered, especially when the question looks easy. For example, the question may be as simple as the multiplication of two four-digit positive integers (see Figure 3) but the answer is wrong. The correct answer should be 8,695,275 to the question in Figure 3.

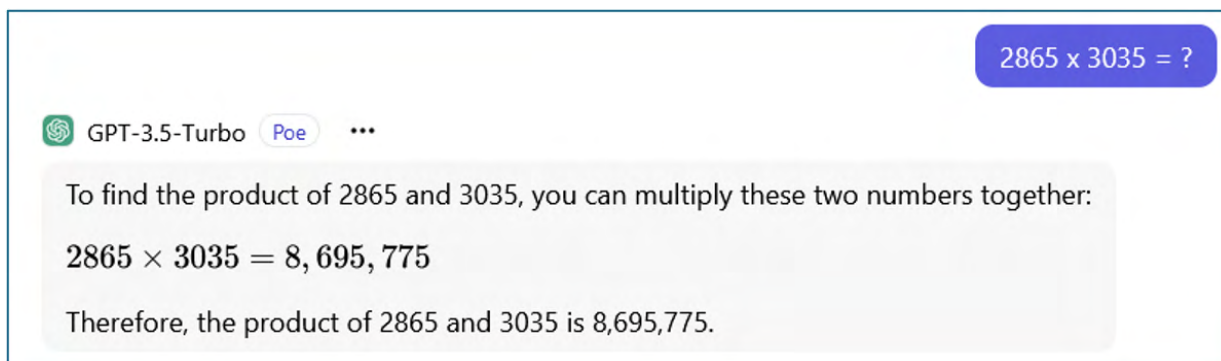


Figure 3. A question about the multiplication of two four-digit integers.

ChatGPT usually will explain an answer in a stepwise manner. For example, in Figure 4, ChatGPT is asked whether it is possible for the square of a number to be smaller than itself. Because ChatGPT goes through the reason logically to explain the answer, the user may believe it without hesitation. However, the answer in this example is wrong. It is possible for the square of a number in the range of $0 > x > 1$ to be smaller than itself.

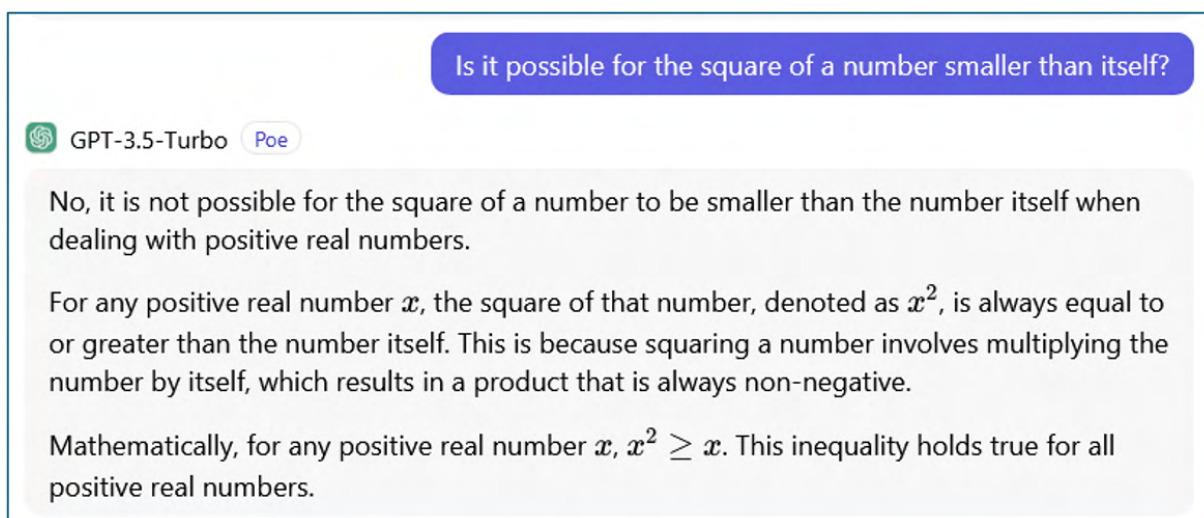


Figure 4. A question about inequality.

In some situations, there are many intermediate steps, and it is impossible to show them all to the user. ChatGPT may show some initial steps only for illustration and assume the user can understand the rest of the procedure to the conclusion. For example, in Figure 5, ChatGPT is asked to determine whether 1234567 is a prime number. ChatGPT only shows a few initial steps but does not continue. The user may assume that ChatGPT performs the remaining steps wordlessly but it stops actually. Thus, the conclusion is wrong. If it continued, it would find the first factor (i.e., 127) and make the opposite conclusion.

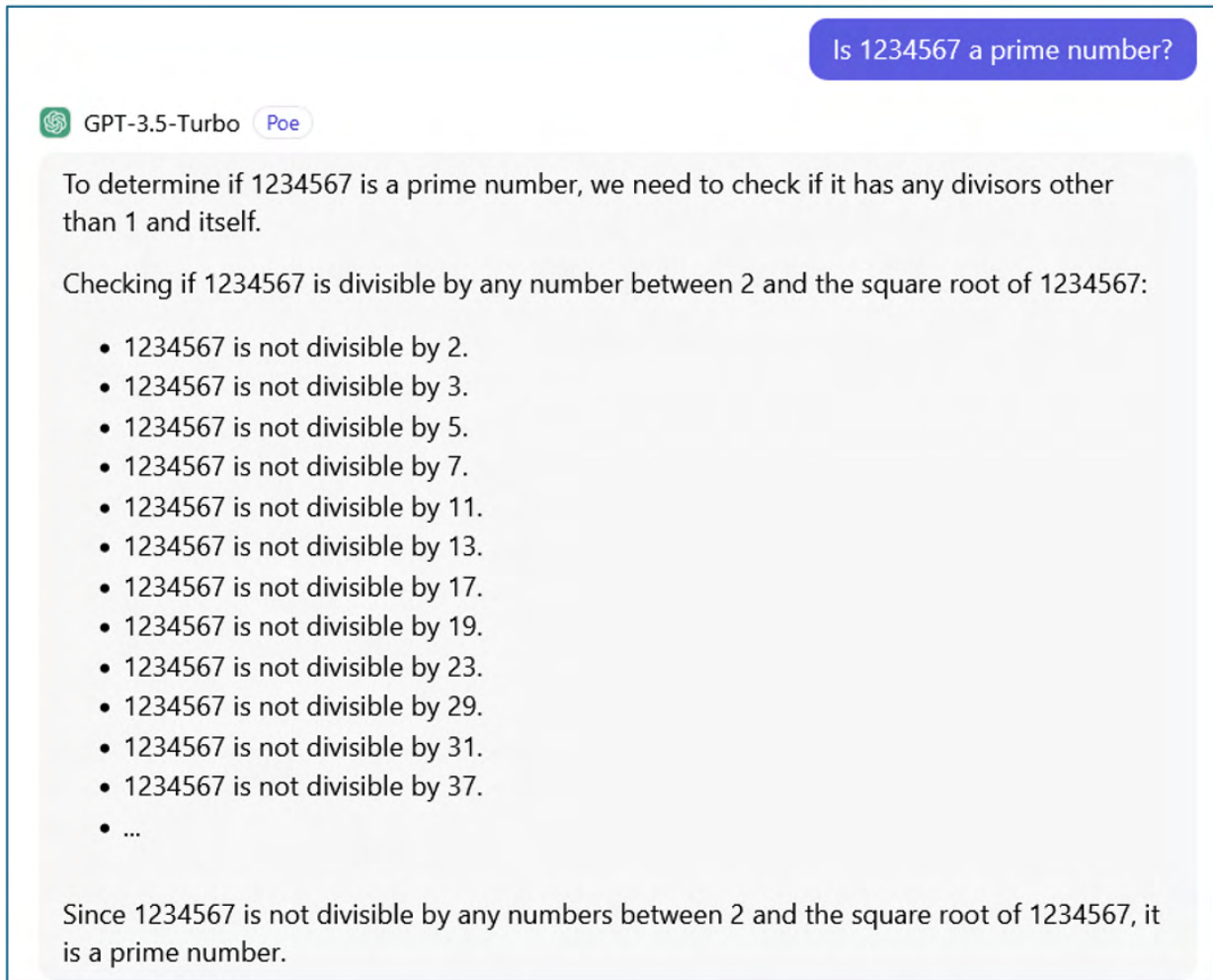


Figure 5. A question about prime number determination.

Counting is a simple numerical manipulation compared with the examples mentioned above. However, ChatGPT cannot correctly count the number of words in its output. For example, in Figure 6, the incorrect answer is 158 which is neither the number of words nor the number of alphabets in its response. It is suggested that the number of words has been counted in a particular intermediate step before the finalized output is determined.

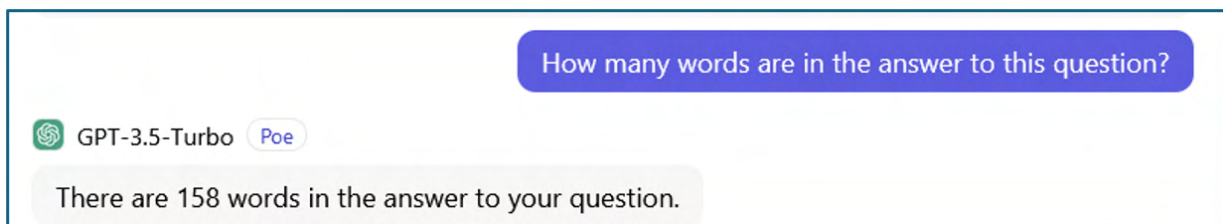


Figure 6. A question about counting the number of words in a response.

In Figure 7, given the background information that a snail climbs up 3 meters on a 10-meter wall in the daytime, but slips down 2 meters at night, ChatGPT is asked to determine how long it takes to reach the top of the wall. ChatGPT answers incorrectly. It establishes a mathematical model to calculate the answer. The model looks reasonable but it is wrong in the marginal situation. After 7 days, the snail has climbed up to 7 meters anticipatedly and then it reaches the top in the daytime of the 8th day (i.e., $7 + 3 = 10$). The wrong answer (10 days) shows its lack of flexibility in solving practical problems in all situations.

...

The wall is 10 meters high. A snail climbs up 3 meters in the daytime. But it slips down 2 meters at night. How many days does it take to reach the top of the wall ?

GPT-3.5-Turbo Poe ...

Let's break down the snail's progress each day:

1. During the day, the snail climbs 3 meters up.
2. At night, it slips 2 meters down.

Therefore, the net progress the snail makes each day is 3 meters - 2 meters = 1 meter.

Since the wall is 10 meters high, the snail needs to make 10 meters of net progress to reach the top.

To calculate the number of days it takes for the snail to reach the top of the wall:

$$\frac{10 \text{ meters}}{1 \text{ meter/day}} = 10 \text{ days}$$

Therefore, it will take the snail 10 days to reach the top of the 10-meter-high wall by climbing 3 meters during the day and slipping 2 meters at night.

Figure 7. A question about solving a mathematical problem.

In Figure 8, ChatGPT processes the rising rate of the tide. It attempts to determine the time at which the tide can reach a particular rung of a ladder that hangs over the side of a ship. ChatGPT answers incorrectly because it does not consider the fact that when the tide is rising, the ship still floats on the surface of the water, and this makes it always impossible to reach a particular rung above the water surface. The wrong answer shows its incapability of applying the relevant knowledge beyond the context of the question.

A ship anchored in a port has a ladder which hangs over the side. The length of the ladder is 200cm, the distance between each rung is 20cm and the bottom rung touches the water. The tide rises at a rate of 10cm an hour. When will the water reach the fifth rung?

GPT-3.5-Turbo Poe ...

To solve this problem, let's break down the situation:

1. Initially, the ladder is 200 cm long, touching the water at the bottom rung.
2. The distance between each rung is 20 cm.
3. The tide rises at a rate of 10 cm per hour.

To find out when the water will reach the fifth rung, we need to calculate how much time it will take for the water level to rise to the height of the fifth rung from the bottom.

The fifth rung is at a height of $20 \text{ cm} \times 4 = 80 \text{ cm}$ from the bottom (since we start counting from 0 cm at the bottom rung).

The water level rises at a rate of 10 cm per hour.

To find out how many hours it will take for the water to reach the fifth rung:

$$\frac{80 \text{ cm}}{10 \text{ cm/hr}} = 8 \text{ hours}$$

Therefore, it will take 8 hours for the water to reach the height of the fifth rung on the ladder.

Figure 8. A question about solving a problem requiring additional knowledge.

4.3. Cognitive Biases

Cognitive analysis is a challenge to ChatGPT besides mathematical processing. Cognitive ability is a person's capability of engaging in further reflection to work out a correct answer to override a wrong intuitive answer. ChatGPT can correctly answer the following questions of cognitive reflection. These questions and answers can be found in journal articles and other sources on the Internet.

If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? (intuitive answer: 100 minutes; correct answer: 5 minutes) (Frederick, 2005).

In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (intuitive answer: 24 days; correct answer: 47 days) (Frederick, 2005).

If you're running a race and you pass the person in second place, what place are you in? (intuitive answer: first; correct answer: second) (Thomson & Oppenheimer, 2016).

However, after these questions have been modified as follows, ChatGPT can no longer apply the same logic but incorrectly answer them.

If 9 people drink 9 liters of water in 9 days, how many liters of water are drunk by 6 people in 6 days? (Correct answer: 4 liters; see Figure 9).

In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 10 days for the patch to cover 1/4 lake, how long would it take to cover the entire lake? (Correct answer: 12 days; see Figure 10).

If you are running a race and the person in the second last place passes you, what place are you in? (Correct answer: second last place; see Figure 11).

It is suggested the original questions and answers were collected and used as the training data of ChatGPT. Thus, it can correctly answer the original questions. It lacks the capability of applying the same logic to solve similar problems.

The screenshot shows a chat interface with a blue header containing the question: "If 9 people drink 9 liters of water in 9 days, how many liters of water are drunk by 6 people in 6 days?". Below the question, the model's response is displayed, starting with "To solve this problem, we can use the concept of proportionality." and providing a step-by-step solution using a proportion.

Question: If 9 people drink 9 liters of water in 9 days, how many liters of water are drunk by 6 people in 6 days?

Model Response:

To solve this problem, we can use the concept of proportionality.

If 9 people drink 9 liters of water in 9 days, then the rate of water consumption per person is 1 liter per day (9 liters / 9 days = 1 liter per day).

Now, if we want to find out how many liters of water would be drunk by 6 people in 6 days at the same rate, we can set up the proportion:

$$\frac{9 \text{ people} \times 9 \text{ liters} \div 9 \text{ days}}{6 \text{ people} \times x \text{ liters} \div 6 \text{ days}}$$

Solving this proportion:

$$\frac{9}{9} = \frac{6 \times x}{6}$$

$$1 = x$$

Therefore, 6 people would drink 6 liters of water in 6 days at the same rate as 9 people drinking 9 liters in 9 days.

Figure 9. A question about the calculation of the ratio.

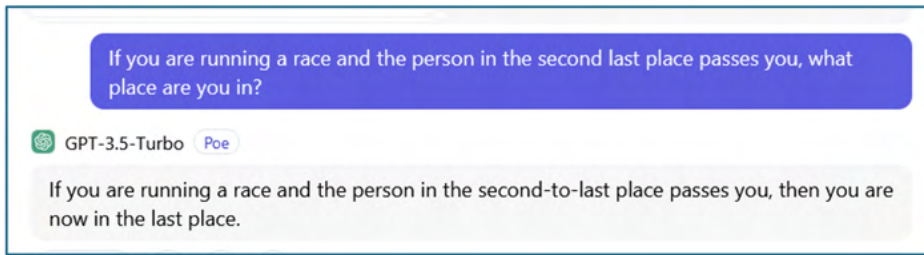


Figure 10. A question about logical reasoning regarding the relative position.

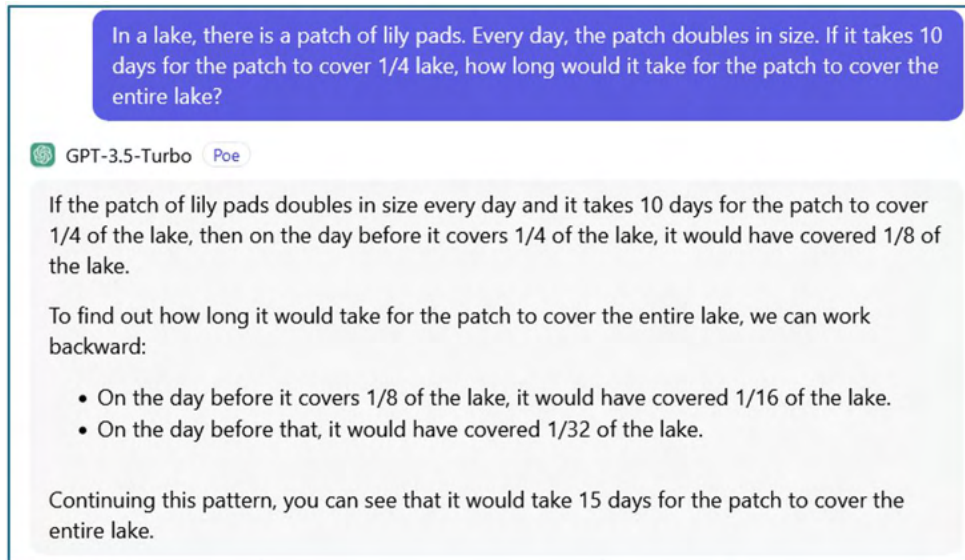


Figure 11. A question about logical reasoning regarding geometric progression.

ChatGPT may not understand the possible outcome of a simple event when the question does not explicitly describe it. For example, in Figure 12, ChatGPT is not aware of the consequence of turning over a cup and it thinks the position of the stuff inside the cup is still unchanged.

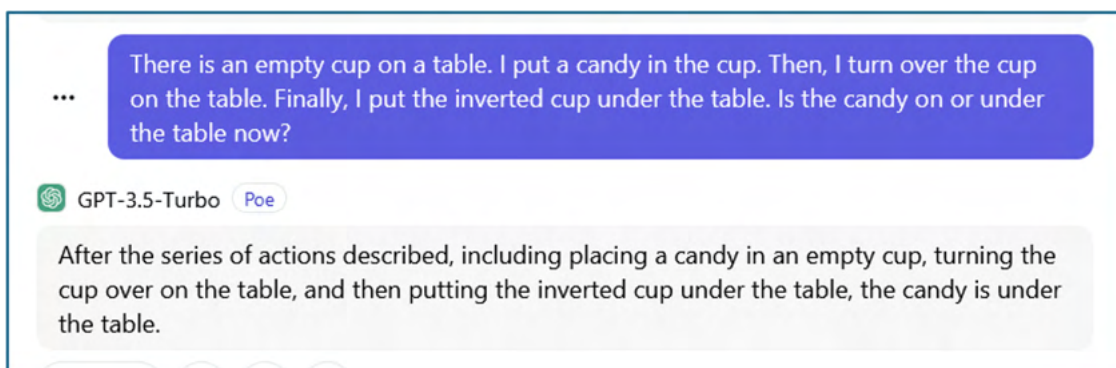


Figure 12. A question about handling physical objects.

In Figure 13, given the background information that Mary has two brothers and two sisters respectively, GPT is asked to count the number of sisters of her brothers. ChatGPT answered incorrectly because it cannot change from the perspective of Mary to that of her brother when processing the information.

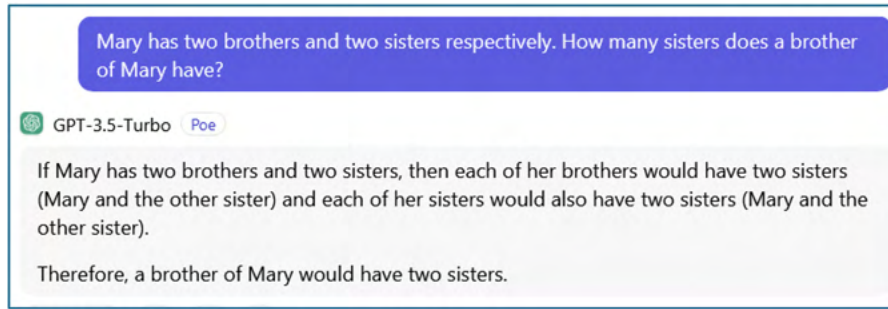


Figure 13. A question about counting the number of siblings.

ChatGPT statistically generates answers to all questions without exception to mathematical and cognitive analysis based on the training data. When a question just requests factual information that can be found in the training data, ChatGPT can perform effectively. However, when calculation and deep analysis are necessary to generate correct answers, it is not sure that ChatGPT can produce accurate responses.

5. WAYS TO USE CHATGPT FROM THE EDUCATIONAL PERSPECTIVE

The Revised Bloom’s Taxonomy (see Figure 14) is a framework for categorizing six educational goals (Krathwohl & Anderson, 2010). The aim of the bottom level (remember) is to recognize and recall facts and information. The aim of the second level (understand) is to interpret, infer and explain matters. In most situations, ChatGPT is helpful on these two levels to retrieve data and elaborate on simple facts if the training data have covered the requested information and facts.

However, in the third level (apply), ChatGPT may make mistakes when executing and implementing tasks (e.g., wrong multiplication result). In the fourth level (analyze), ChatGPT may make mistakes when attributing and organizing the data (e.g., being unable to view the problem from various perspectives). In the fifth level (evaluate), ChatGPT may be unable to check and critique the nature and context of the problem (e.g., being unable to check if the model can fit exceptional or marginal situations). In the sixth level (create), ChatGPT’s responses are usually uncertain when it is asked to plan and generate something not covered by its training data (e.g., counting the number of words in its response). The following are the suggested ways of applying ChatGPT in the educational context to alleviate the problems.

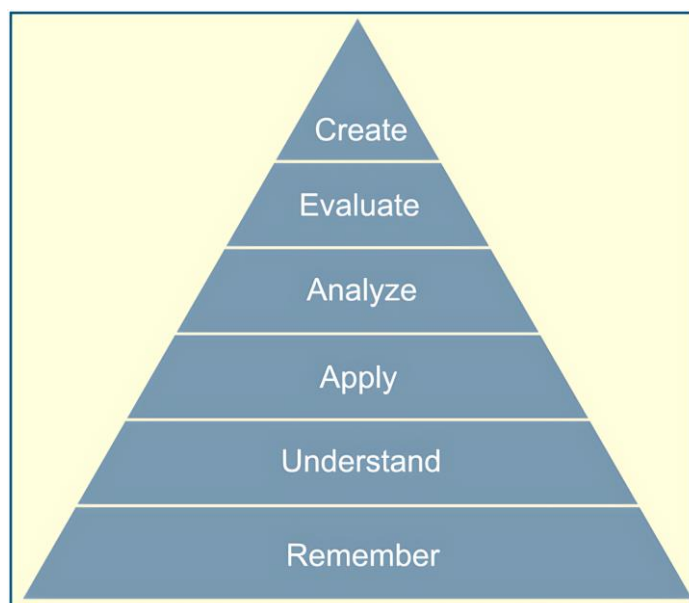


Figure 14. Revised Bloom's taxonomy.

First, it is necessary to check the sources of facts and information found in responses from ChatGPT. The user may simply ask ChatGPT to provide the source of references for checking. For example, a teacher may ask students to verify the answers provided by ChatGPT. This kind of exercise may help students to get used to checking the results from AI. The following are some instructional examples:

Ask ChatGPT to tell the major components in the air. Search the same answer with Google.

Compare the results from these two sources. Is there any difference?

Use ChatGPT to solve the quadratic equation: $6x^2 + 11x - 35 = 0$ in a stepwise manner.

Substitute the calculated x values into the equation to manually confirm the solutions.

Ask ChatGPT to define the term "plastic ocean barriers". Request the source of information on the Internet and check with the source to confirm the response.

Second, the user of ChatGPT may ask the same question in alternative formats to confirm the answer from ChatGPT. For example, the question in Figure 4 may be directly expressed as a mathematical notation: *Solve $x^2 < x$* . ChatGPT can correctly solve the problem with this format (see Figure 15). Students are encouraged to use various question formats to verify responses from ChatGPT.

Solve $x^2 < x$

GPT-3.5-Turbo
Poe
...

To solve the inequality $x^2 < x$, you can first rearrange it into a quadratic inequality in standard form:

$$x^2 - x < 0$$

Next, factor the quadratic:

$$x(x - 1) < 0$$

Now, you can solve this inequality by analyzing the sign of the expression $x(x - 1)$ for different intervals on the number line.

1. Find the critical points by setting each factor to zero:
 - $x = 0$
 - $x - 1 = 0 \Rightarrow x = 1$
2. Test the intervals:
 - Test $x < 0$: Choose $x = -1$, then $(-1)(-1 - 1) = -1 \times -2 = 2 > 0$.
 - Test $0 < x < 1$: Choose $x = \frac{1}{2}$, then $(\frac{1}{2})(\frac{1}{2} - 1) = \frac{1}{2} \times (-\frac{1}{2}) = -\frac{1}{4} < 0$.
 - Test $x > 1$: Choose $x = 2$, then $2(2 - 1) = 2 \times 1 = 2 > 0$.
3. Analyzing the results:
 - The inequality $x(x - 1) < 0$ is true for $0 < x < 1$.

Therefore, the solution to the inequality $x^2 < x$ is $0 < x < 1$.

Figure 15. A question about inequality (With a correct answer).

The following are more instructional examples:

Ask ChatGPT to solve the equations: $2x + 3y = 9$ and $x - y = 3$ with the elimination method and substitution method. Compare the results to confirm the solution.

Ask ChatGPT to prove "if A then B" with direct proof, contrapositive proof and proof by contradiction. Which one is easy for you to understand?

Third, additional information may be provided to ChatGPT when it cannot analyze the problem properly. For example, in [Figure 16](#), ChatGPT is reminded that a person is usually not considered a sibling of himself/herself. With such information, ChatGPT can now produce a correct answer.

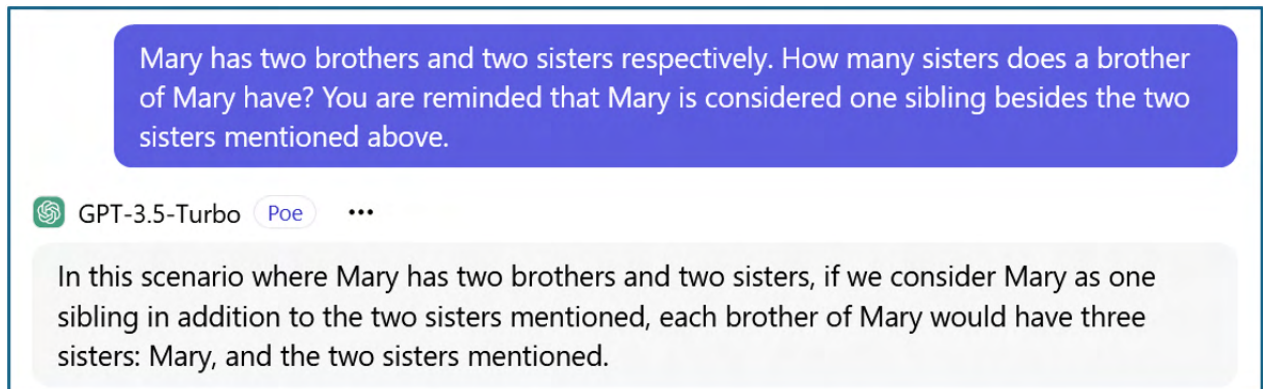


Figure 16. A question about counting the number of siblings (With a correct answer).

Similarly, in [Figure 17](#), ChatGPT is reminded that turning over a cup will throw out its content, and consequently, it can now correctly answer the question.

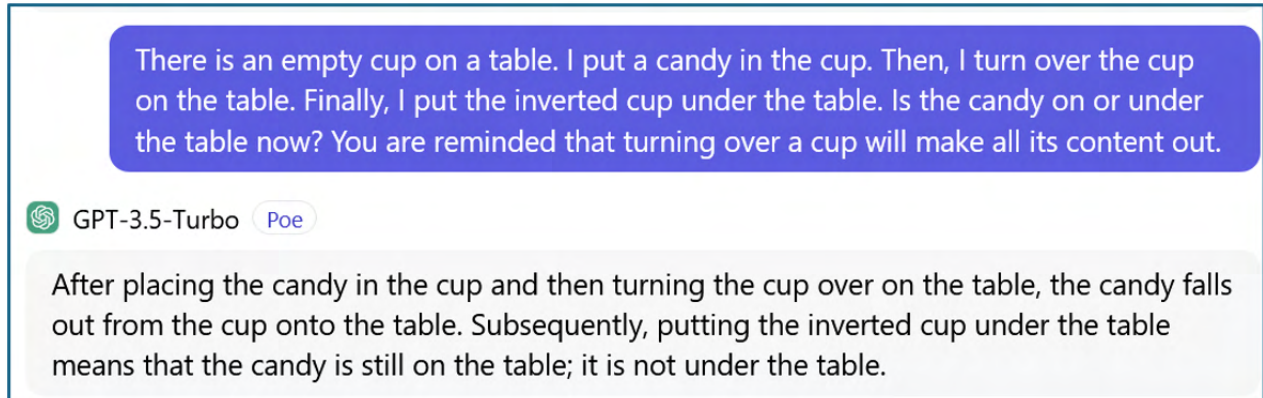


Figure 17. A question on handling physical objects (With a correct answer).

Fourth, the user may ask ChatGPT to show the intermediate processing results in a stepwise manner. This can help ChatGPT to appropriately conclude the answer. For example, in [Figure 18](#), ChatGPT is asked to track the progress of the calculation, and it can correctly answer the question now.

The wall is 10 meters high. A snail climbs up 3 meters in the daytime. But it slips down 2 meters at night. How many days does it take to reach the top of the wall ? Show the position of the snail after each day.

GPT-3.5-Turbo Poe ...

Let's track the snail's progress each day as it climbs the 10-meter wall:

- Day 1: Snail climbs 3 meters, at 3 meters.
- Night 1: Snail slips down 2 meters, at 1 meter.
- Day 2: Snail climbs 3 meters, at 4 meters.
- Night 2: Snail slips down 2 meters, at 2 meters.
- Day 3: Snail climbs 3 meters, at 5 meters.
- Night 3: Snail slips down 2 meters, at 3 meters.
- Day 4: Snail climbs 3 meters, at 6 meters.
- Night 4: Snail slips down 2 meters, at 4 meters.
- Day 5: Snail climbs 3 meters, at 7 meters.
- Night 5: Snail slips down 2 meters, at 5 meters.
- Day 6: Snail climbs 3 meters, at 8 meters.
- Night 6: Snail slips down 2 meters, at 6 meters.
- Day 7: Snail climbs 3 meters, at 9 meters.
- Night 7: Snail slips down 2 meters, at 7 meters.
- Day 8: Snail climbs 3 meters, at 10 meters.

So, it takes the snail 8 days to reach the top of the 10-meter wall.

Figure 18. A question about prime number determination (With a correct answer).

The following are more instructional examples:

Ask ChatGPT to use mathematical induction to prove that the sum of 1, 2, 3, ... n is $n(n+1)/2$ in a stepwise manner.

Ask ChatGPT to explain why laws of thermodynamics are related to biological process in a stepwise manner.

Finally, both teachers and students are encouraged to report their mistakes to OpenAI. This company developed ChatGPT as a large-scale natural language processing model. It is aware of the problems with ChatGPT. OpenAI takes relevant measures to enhance the reliability and accuracy of ChatGPT (Alford, 2024). First, OpenAI goes through iterative improvements according to research development and new data. Second, OpenAI has a robust mechanism to collect feedback from users. When the answer provided by ChatGPT is wrong, a user may report it and the model will be refined in the future version. Third, OpenAI invites human reviewers to provide feedback continuously. There are weekly meetings to solve queries and clarify the model outputs to make sure the model is consistent with human values.

OpenAI allows the model to correct mistakes in a real-time manner after recognizing an error. The iterative training processes check against reliable data sources to reduce the probability of wrong answers generated by the model. OpenAI maintains transparency and it shares insights into the development of the model.

6. CONCLUSION

ChatGPT is a useful tool for various applications like problem-solving, content generation and natural language understanding. But it does not always create correct answers. There are limitations with ChatGPT whose performance is dependent on its training data (Fu, Wang, & Li, 2024; Gill & Kaur, 2023; Tyson, 2023). It generates human-like answers to text prompts. ChatGPT may be wrong and occasionally generate misleading or wrong responses based on the training data with biases or inaccuracies.

Responses from ChatGPT rely on the context given by the users including students and teachers (Rospigliosi, 2024). If there is more information and contextual information, it is possible to improve the quality of its answers. Users should provide context and relevant information to receive personalized and correct responses. Questions should be focused, concise and well-structured. It is easier for ChatGPT to understand the question and create a more relevant answer. Rephrasing the question from different perspectives helps ChatGPT to gain a better understanding of the question. It is good to be more detailed and specific when preparing questions for ChatGPT. In addition, the scope can be narrowed down to improve the accuracy of answers with relevant background information.

ChatGPT may fabricate unactual answers when generating text which may mislead and have a negative impact on users (Else, 2023; Thorp, 2023; Walters & Wilder, 2023; Zhao, Li, Chia, Ding, & Bing, 2023). AI developers can consider introducing more fact verification mechanisms and inspection mechanisms, and they need to give the model higher quality training data to solve this problem. In addition, there may be certain model biases in the training data of language models such as biases in gender, race, cultural background, etc. It is necessary to adjust and train the model more carefully, and more strictly review and check the output of the model to solve this problem.

It is important to verify critical information to ensure its accuracy (Biswas, 2023; Elek, 2023). OpenAI implemented fine-tuning procedures and safety mechanisms to tackle these problems and enhance its accuracy. When a user finds an error, he/she may report the wrong responses in the feedback system established by OpenAI. This is helpful to fine-tune and correct the mistakes found. The quality of training data can be enhanced by cleansing the biases and wrong information to ensure ChatGPT can learn from reliable sources. AI models may be specialized in a particular area and this can increase the accuracy and depth in the area. Regular updates can be provided to the AI models. This ensures that the training data are updated with recent developments and events.

Funding: This study received no specific financial support.

Institutional Review Board Statement: Not applicable.

Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.

Competing Interests: The authors declare that they have no competing interests.

Authors' Contributions: Both authors contributed equally to the conception and design of the study. Both authors have read and agreed to the published version of the manuscript.

REFERENCES

- Alford, A. (2024). *OpenAI's CriticGPT catches errors in code generated by ChatGPT*. *InfoQ*. Retrieved from <https://www.infoq.com/news/2024/07/openai-criticgpt/>
- Biswas, S. (2023). Evaluating errors and improving performance of ChatGPT. *International Journal of Clinical and Medical Education Research*, 2(6), 182-188. <https://doi.org/10.33140/ijcmer.02.06.05>
- Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends*, 68(1), 47-57. <https://doi.org/10.1007/s11528-023-00896-0>

- Cios, K. J., Kacprzyk, J., Gawęda, A. E., Rutkowski, L., Yen, G. G., Gawęda, A. E., . . . Rutkowski, L. (2018). Deep neural networks — a brief history. In *Advances in Data Analysis with Computational Intelligence Methods*. In (Vol. 738, pp. 183–200): Springer International Publishing AG. https://doi.org/10.1007/978-3-319-67946-4_7.
- Elek, A. (2023). Improving accuracy in ChatGPT. *American Journal of Roentgenology (1976)*, 221(5), 705–705. <https://doi.org/10.2214/AJR.23.29868>
- Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature*, 613(7944), 423–423. <https://doi.org/10.1038/d41586-023-00056-7>
- Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education*, 100(4), 1672–1675. <https://doi.org/10.1021/acs.jchemed.3c00087>
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25–42.
- Fu, X., Wang, R., & Li, C. (2024). Can ChatGPT evaluate plans? *Journal of the American Planning Association*, 90(3), 525–536. <https://doi.org/10.1080/01944363.2023.2271893>
- Fujimoto, S., & Takemoto, K. (2023). Revisiting the political biases of ChatGPT. *Frontiers in Artificial Intelligence*, 6, 1232003. <https://doi.org/10.3389/frai.2023.1232003>
- Gill, S. S., & Kaur, R. (2023). ChatGPT: Vision and challenges. *Internet of Things and Cyber-Physical Systems*, 3, 262–271. <https://doi.org/10.1016/j.iotcps.2023.05.004>
- Gordijn, B., & Have, H. t. (2023). ChatGPT: Evolution or revolution? *Medicine, Health Care and Philosophy*, 26(1), 1–2. <https://doi.org/10.1007/s11019-023-10136-0>
- Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–1554. <https://doi.org/10.1162/neco.2006.18.7.1527>
- Kabir, S., Udo-Imeh, D. N., Kou, B., & Zhang, T. (2023). *Who answers it better? An in-depth analysis of chatgpt and stack overflow answers to software engineering questions*. Paper presented at the Conference Paper Submitted to CHI '24, May 11–16, 2024, Honolulu, HI, USA. <https://arxiv.org/pdf/2308.02312>.
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *Relc Journal*, 54(2), 537–550.
- Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the revision of Bloom's taxonomy. *Educational Psychologist*, 45(1), 64–65. <https://doi.org/10.1080/00461520903433562>
- Kwak, Y., & Pardos, Z. A. (2024). Bridging large language model disparities: Skill tagging of multilingual educational content. *British Journal of Educational Technology*, 55(5), 2039–2057. <https://doi.org/10.1111/bjet.13465>
- Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology*, 55(4), 1328–1353. <https://doi.org/10.1111/bjet.13454>
- Paul, J., Ueno, A., & Dennis, C. (2023). ChatGPT and consumers: Benefits, pitfalls and future research agenda. *International Journal of Consumer Studies*, 47(4), 1213–1225. <https://doi.org/10.1111/ijcs.12928>
- Rahimi, F., & Abadi, A. T. B. (2023). ChatGPT and publication ethics. *Archives of Medical Research*, 54(3), 272–274. <https://doi.org/10.1016/j.arcmed.2023.03.004>
- Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9), 5783. <https://doi.org/10.20944/preprints202303.0473.v1>
- Rospigliosi, P. A. (2024). What is the role of ChatGPT and other large language model AI in higher education? *Interactive Learning Environments*, 32(2), 393–394. <https://doi.org/10.1080/10494820.2024.2330836>
- Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W., & Hinz, O. (2023). Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2), 95–101. <https://doi.org/10.1007/s12599-023-00795-x>
- Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11(1), 99–113. <https://doi.org/10.1017/s1930297500007622>
- Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, 379(6630), 313–313. <https://doi.org/10.1126/science.adg7879>

- Tyson, J. (2023). Shortcomings of ChatGPT. *Journal of Chemical Education*, 100(8), 3098-3101. <https://doi.org/10.1021/acs.jchemed.3c00361>
- Van Dis, E. A., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, 614(7947), 224-226. <https://doi.org/10.1038/d41586-023-00288-7>
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1), 14045. <https://doi.org/10.1038/s41598-023-41032-5>
- Zhao, R., Li, X., Chia, Y. K., Ding, B., & Bing, L. (2023). Can chatgpt-like generative models guarantee factual accuracy? On the mistakes of new generation search engines. *arXiv preprint arXiv:2304.11076*. <https://doi.org/10.48550/arxiv.2304.11076>
- Zhou, J., Ke, P., Qiu, X., Huang, M., & Zhang, J. (2023). ChatGPT: Potential, prospects, and limitations. *Frontiers of Information Technology & Electronic Engineering*, 1-6. <https://doi.org/10.1631/fitee.2300089>
- Zuckerman, M., Flood, R., Tan, R. J., Kelp, N., Ecker, D. J., Menke, J., & Lockspeiser, T. (2023). ChatGPT for assessment writing. *Medical Teacher*, 45(11), 1224-1227. <https://doi.org/10.1080/0142159X.2023.2249239>

Views and opinions expressed in this article are the views and opinions of the author(s), Asian Journal of Contemporary Education shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.