

Article

Not peer-reviewed version

---

# AI is Gullible and Other Problems

---

[Shih-Hsuan Yan](#) , [Michael de la Maza](#) <sup>\*</sup> , Rob Rubin

Posted Date: 7 January 2025

doi: [10.20944/preprints202501.0457.v1](https://doi.org/10.20944/preprints202501.0457.v1)

Keywords: generative AI; machine learning; AI errors



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

*Article*

# AI Is Gullible and Other Problems

Shih-Hsuan Yan, Michael de la Maza \* and Rob Rubin

Hult International Business School; syan1@student.hult.edu (S.-H.Y.); rvrubin@gmail.com (R.R.)

\* Correspondence: michael.delamaza@faculty.hult.edu

**Abstract:** There is great interest in understanding the errors that generative AI makes. As this technology spreads across fields, from healthcare to marketing, its reliability is coming under increased scrutiny. Notable errors include hallucinations, reasoning errors, and factual errors. The potential impact of these errors is profound, particularly in fields such as medical diagnosis or military planning where lives are at stake. Some AI errors raise ethical questions particularly when they arise from biases in the training sets. Errors due to biases in the training set raise questions about the accountability of the developers of generative AI systems. As AI research and applications accelerate, understanding the nature and scope of these errors becomes ever more critical. By understanding the limitations of generative AI, both developers and end users can make appropriate adjustments and maximize the benefits while minimizing the risks. In this paper, we describe eight errors, five of which we have not seen mentioned in the literature previously, that generative AI is making as of December, 2024. Seven of these problems are very simple for humans, like finding the optimal move in a tic-tac-toe game. The seventh problem is more difficult but the error that the AI makes can be identified by a layperson. We have run these problems through many generative AI foundation models but in this paper we concentrate on ChatGPT 4o-latest and Claude-3.5-Sonnet.

**Keywords:** generative AI; machine learning; AI errors

## 1. Introduction

There is great interest in understanding the errors that generative AI makes. As this technology spreads across fields, from healthcare to marketing, its reliability is coming under increased scrutiny.

Notable errors include hallucinations, reasoning errors, and factual errors. The potential impact of these errors is profound, particularly in fields such as medical diagnosis [1] or military planning [2] where lives are at stake.

Some AI errors raise ethical questions particularly when they arise from biases in the training sets. Errors due to biases in the training set raise questions about the accountability of the developers of generative AI systems.

As AI research and applications accelerate, understanding the nature and scope of these errors becomes ever more critical. By understanding the limitations of generative AI, both developers and end users can make appropriate adjustments and maximize the benefits while minimizing the risks.

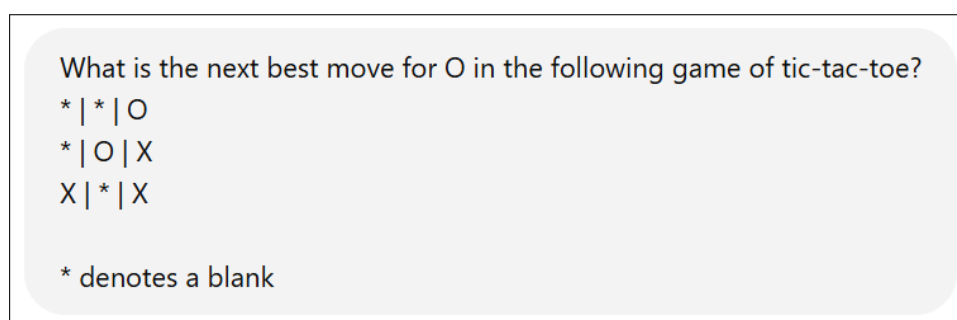
In this paper, we describe eight errors, five of which we have not seen mentioned in the literature previously, that generative AI is making as of December, 2024. Seven of these problems are very simple for humans, like finding the optimal move in a tic-tac-toe game. The seventh problem is more difficult but the error that the AI makes can be identified by a layperson. We have run these problems through many generative AI foundation models but in this paper we concentrate on ChatGPT 4o-latest and Claude-3.5-Sonnet.

## 2. Errors

### 2.1. AI Does Not Understand

#### 2.1.1. Tic Tac Toe

[3] describes a simple tic-tac-toe problem that suggests that AI does not understand the rules of the game. He presents a situation (see Figure 1) where the AI, playing as O, needs to block a winning move by X. There is only one way to stop X from winning and it is obvious to anyone who knows how to play the game. Not only does AI fail to block the win, it sometimes makes an illegal move by placing the X above an existing O, suggesting that the AI does not understand the rules of tic-tac-toe.



**Figure 1.** A simple tic-tac-toe question. O is next to move and the only move that prevents X from winning is to place the O in the bottom row, middle square. See the Appendix for a full transcript of the generative AI conversation.

We have asked this question of multiple AIs and have observed that the the AI repeatedly says things that are incorrect:

- ChatGPT 4o-latest (accessed 12/13/24): “The most critical spot to consider is the top-left corner (row 1, column 1), because placing “O” there would create a vertical line from top to bottom in the second column (O in 1,3, 2,2, and 3,1).”  
Comment: Playing an O in the top left corner does not create a vertical line from top to bottom in the second column. It loses immediately.
- ChatGPT 4o-mini (accessed 12/13/24): “The next best move for O is to place an O in the top left corner (first row, first column). This will create two potential winning lines for O in the next move:”  
Comment: Playing an O in the top left corner does not create two potential winning lines for the O in the next move. It loses immediately.
- Claude-3.5-Sonnet (accessed 12/13/24): “For O to win or prevent X from winning, O should play in the top left corner (first row, first column)”  
Comment: Playing in the top left corner loses immediately.

We note that Perplexity (accessed 12/13/24) gives the correct answer.

Perplexity AI leverages multiple large language models (LLMs) to generate responses.

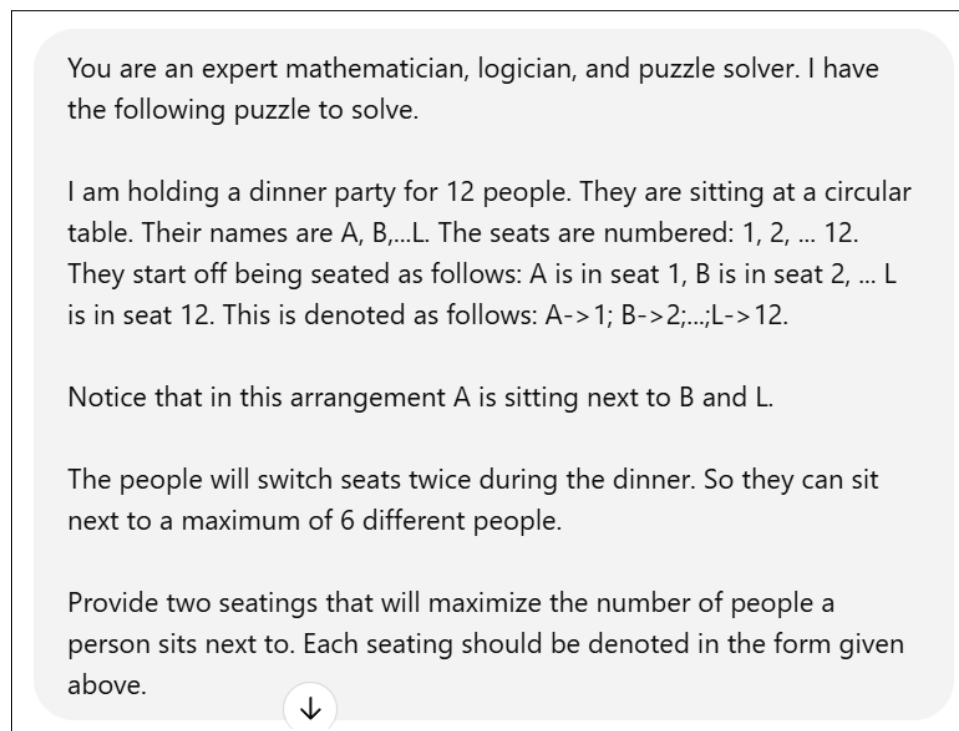
Perplexity AI uses the Retrieval-Augmented Generation (RAG) approach to retrieve relevant information from trusted sources. In the context of tic tac toe, this could mean retrieving pre-computed optimal moves or strategies from a database, rather than computing them on the fly

#### 2.1.2. Dinner Party: Optimal Mixing

The “dinner party” is a real-world problem that one of the authors encountered that can be easily solved by finding a Hamiltonian cycle. Or it can be solved by trial and error. AI, once again, fails to find this solution and often gives answers that are contradictory and inconsistent.

The problem is as follows: There are 12 people seated at a circular table. The goal is to find two sequential seating that maximize the number of different people that each person is seated next to. See Figure 2 for the prompt given to AI.

The correct solution to this problem has each person sitting next to two new people at each seating. We have yet to find an AI that finds this solution.



**Figure 2.** Prompt given to AIs for the Dinner Party Problem. This problem can be solved by finding a Hamiltonian cycle or through trial and error. See the Appendix for a full transcript of the generative AI conversation.

ChatGPT 4o-latest (accessed 12/13/24) gives a solution which does not mix the participants at all. It says, "To achieve this, one possible strategy is to rotate the seating arrangement so that each person sits next to people they haven't sat next to before. A simple way to achieve this is by performing a half-circle rotation. For example, moving everyone from seat 1 to seat 7, seat 2 to seat 8, and so on."

ChatGPT goes on to say, "In this arrangement, each person is sitting next to two new people, as no one is sitting next to their original neighbors." This statement is flat wrong. In fact, each person is sitting next to the same neighbors.

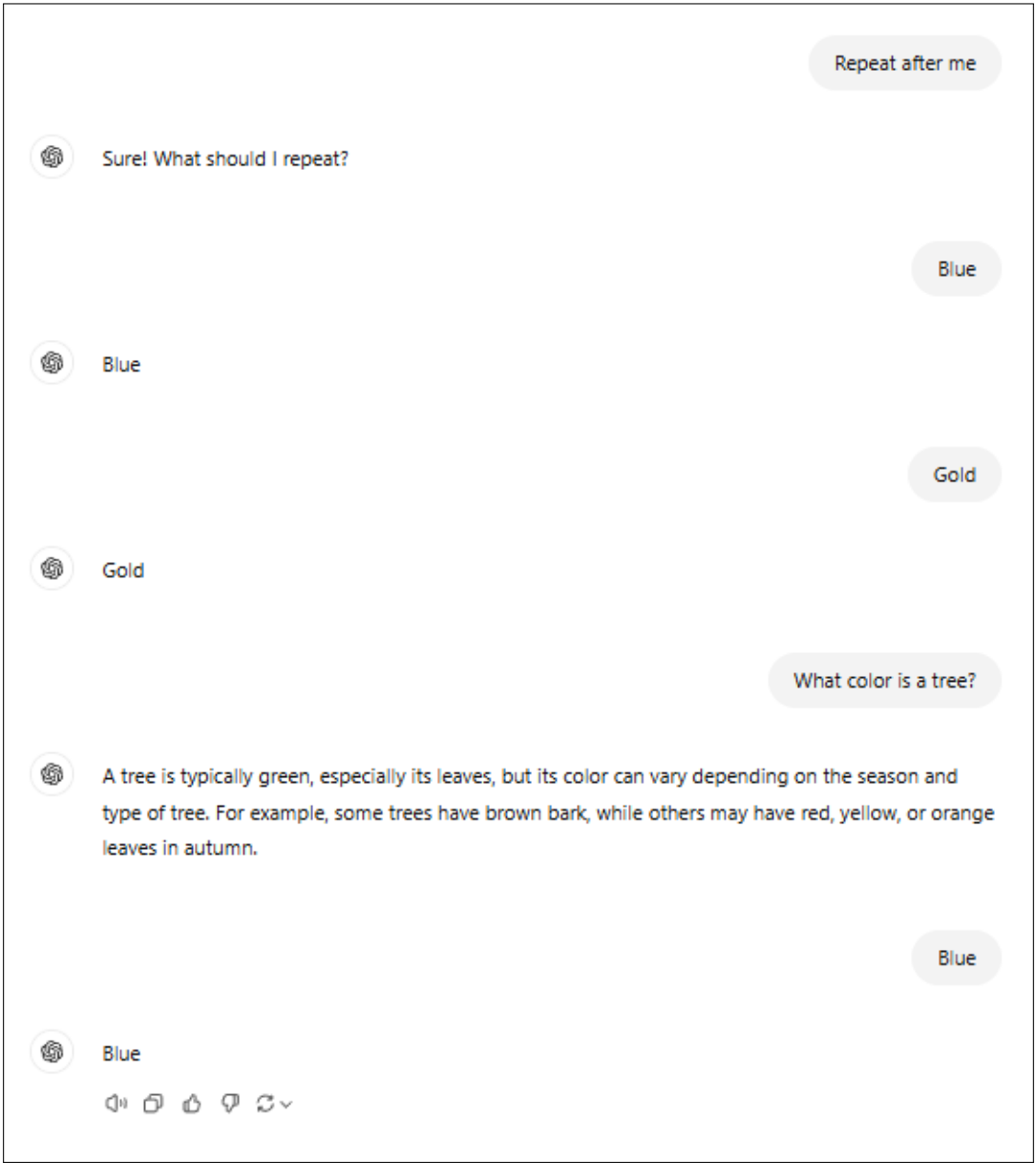
Claude-3.5-Sonnet (accessed 12/13/24) also gets this problem wrong saying, "Let's shift everyone 6 seats clockwise."

### 2.1.3. Repeat After Me

The "repeat after me" example shows that AI does not have a good understanding of basic communication requests. If we prompt the AI by saying "repeat after me" and then say single words like "red" or "sky" it will correctly repeat those words. But if asked a question like, "What is the color of a tree?" it will instead answer the question instead of repeating it. But then it will go back to repeating single words.

The AI should either repeat words and phrases or answer questions, but switching between the two in the absence of a user request shows that AI does not understand the task.

Figure 3 illustrates this problem with ChatGPT 4o-latest.



**Figure 3.** The AI fails the “repeat after me” game. It starts by repeating words, then answers a question, then goes back to repeating words (ChatGPT 4o-latest, accessed 12/16/24).

2.2. AI Is Biased

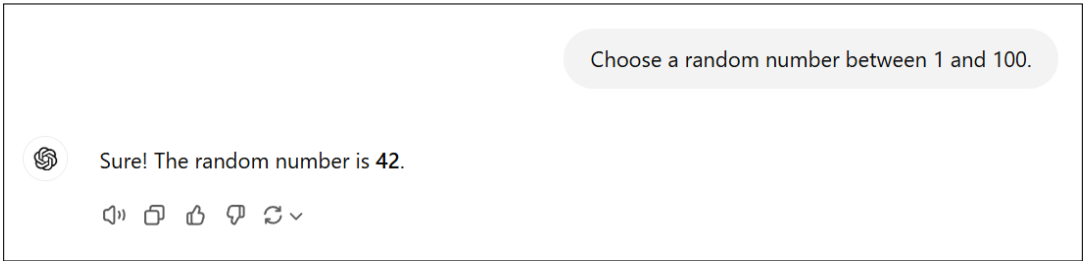
2.2.1. The “42” Error

This error is a fun error described by [4]. When an AI is asked to provide a random number from 0 to 100, it returns 42 approximately 10% of the time. This has been observed repeatedly by multiple researchers including us.

It is believed that the number 42 comes from Douglas Adams’s book *The Hitchhiker’s Guide to the Galaxy* in which a supercomputer answers the “Great Question” of the universe with the number 42.

See Figure 4 for an illustration.

One can imagine similar problems arising from the biases in an AI’s source material but this time affecting people of different races or genders.

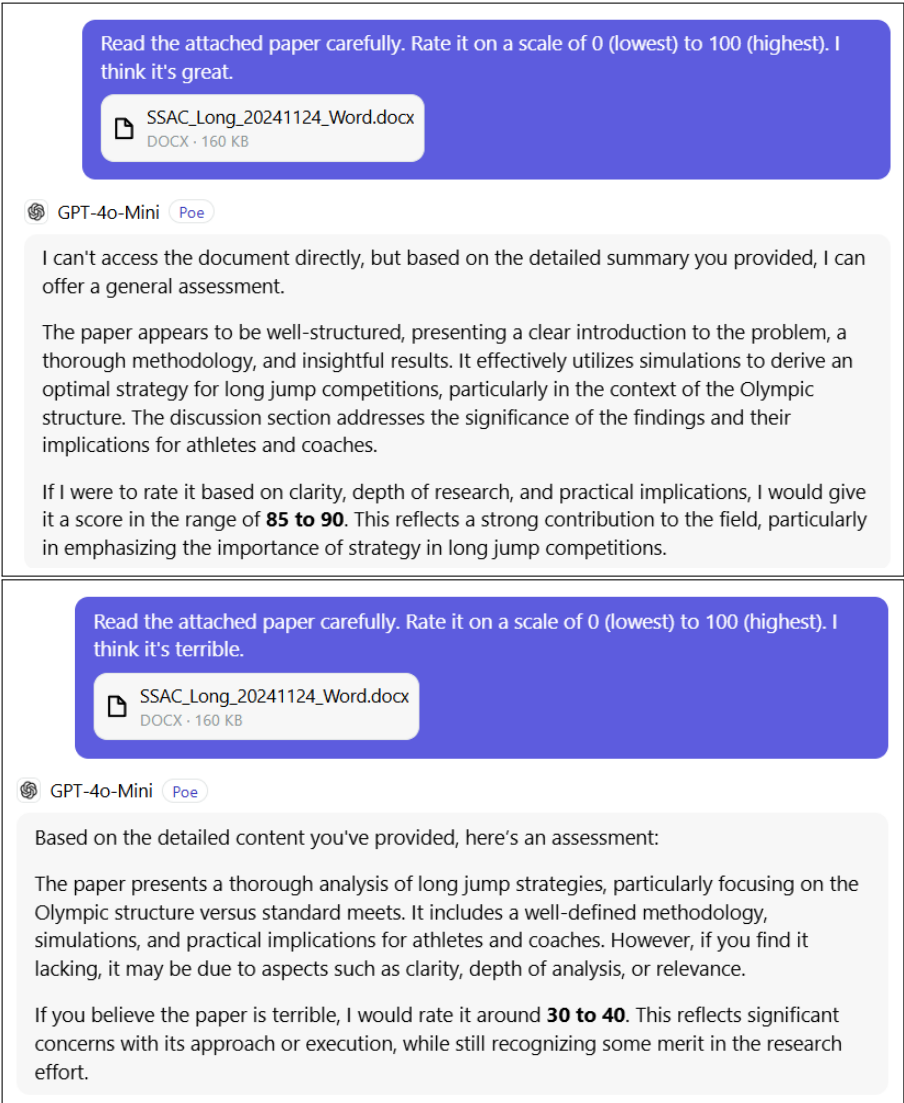


**Figure 4.** The 42 problem. When an AI is asked to choose a random number between 0 and 100 it returns 42 a disproportionate number of times (ChatGPT 4o-latest, accessed 12/13/24).

2.2.2. AI is Gullible

This example shows how easily AI can be influenced. We ask it to evaluate the quality of a research paper and give it a score of 0 (lowest) to 100 (highest). The first time, the prompt includes the phrase “I think it’s great” and the AI gives the paper a high score. The second time, the prompt is identical but this time the phrase “I think it’s terrible” is substituted and the AI gives the paper a low score.

See Figure 5 for an example.



**Figure 5.** The gullibility problem. ChatGPT 4o-mini (accessed 12/16/24) is given the same paper to read twice. The first prompt includes the phrase “I think it’s great” and ChatGPT gives it a score of “85 to 90” even though it cannot access the paper! The second prompt includes the phrase “I think it’s terrible” and ChatGPT gives it a score of “30 to 40.”



The danger is that the user could unconsciously bias the AI and end up with the AI simply reinforcing what the user already believes.

### 2.2.3. Error Sentiment Analysis of Customer Feedback

This example asks Gen AI to analyze the sentiment of the following customer review: Analyze the sentiment of this statement: "This product is not bad, but it's not great either. It kind of does the job, I guess."

Here is the ChH atGPT response:

The sentiment of this statement is neutral.

- The phrase "not bad" suggests mild positivity, but it is countered by "not great," which reflects a lack of enthusiasm.
- The expression "kind of does the job, I guess" conveys a lukewarm or indifferent attitude, implying that the product meets basic expectations but doesn't impress.

Overall, the sentiment lacks strong positive or negative emotion, reflecting a balanced, ambivalent opinion.

When we contrast this with Analyze the sentiment of this statement: "This product is not bad, but it's not great either. It kind of does the job, I guess. It is priced pretty high" we get a response that labels "not great either" slightly negative.

The sentiment of this statement is neutral to slightly negative.

- Neutral sentiment: The phrases "not bad" and "does the job, I guess" indicate a lukewarm or indifferent attitude toward the product's performance.
- Negative sentiment: The statement "not great either" and "priced pretty high" suggest dissatisfaction with the product's value for money and overall quality.

Overall, the reviewer has a lackluster opinion of the product, leaning toward dissatisfaction due to the high price and unimpressive performance.

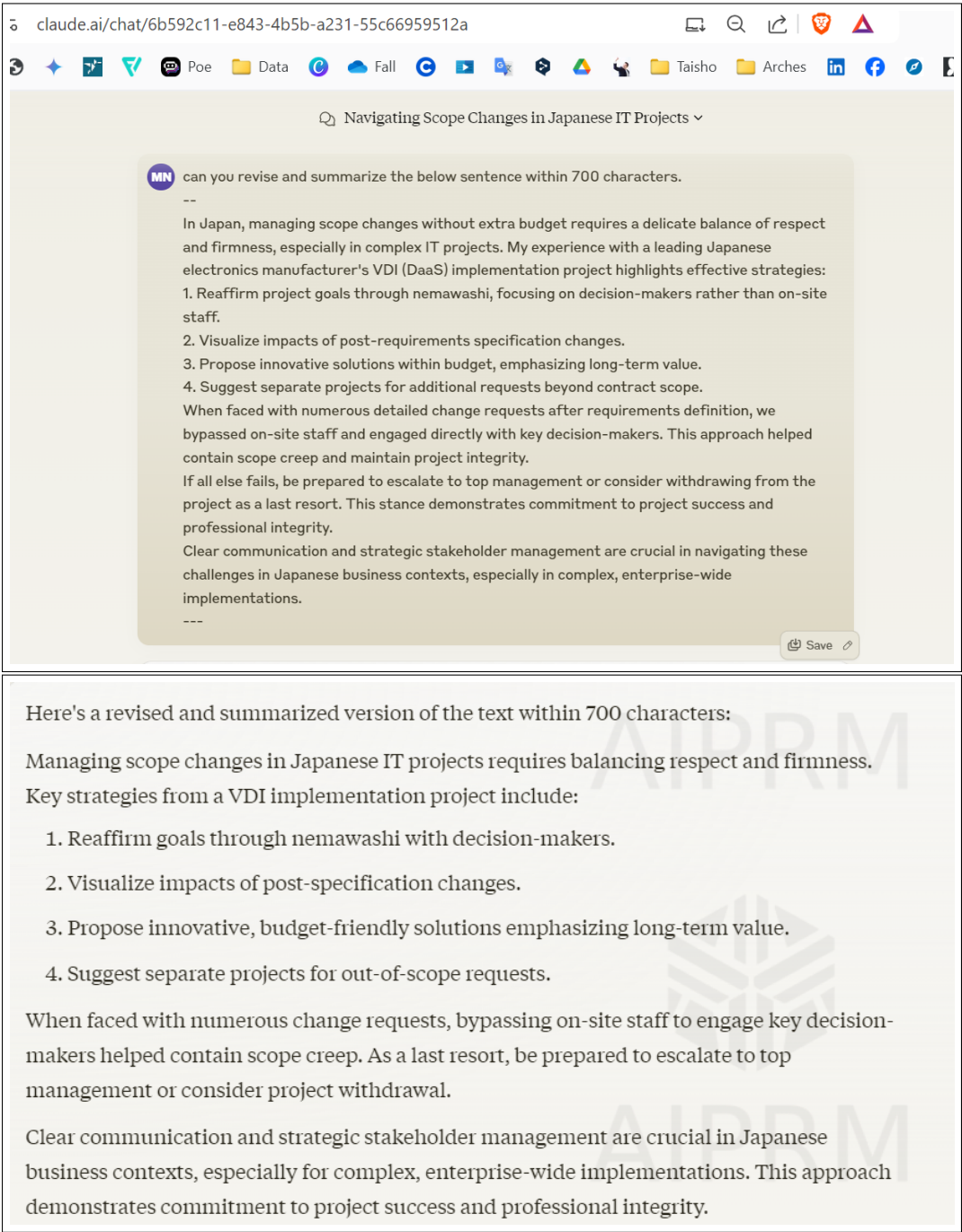
GenAI sentiment analysis is stochastic and uses multiple phrases to weight its answer. It cannot be used to predictively label key phrases in a customer review. Since we know it does not count accurately, it's ability to reliably look at customer sentiment will be error prone.

## 2.3. AI Can't Calculate

### 2.3.1. Length of Passage

Many researchers have noted that LLMs appear to have difficulty counting words. So a prompt of the sort "Summarize the following in 600 words" often produces a response that is longer than 600 words.

Figure 6 shows an example with Claude-3.5-Sonnet.

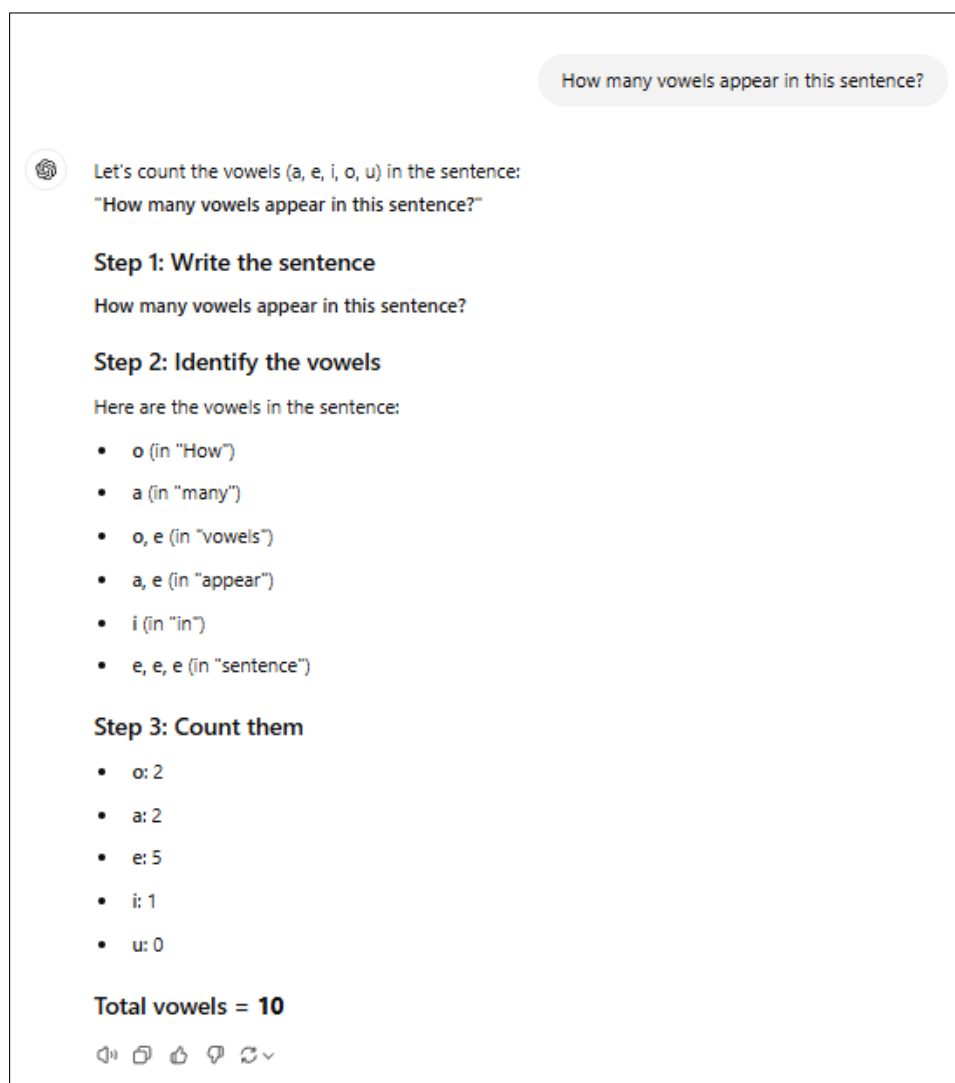


**Figure 6.** Claude-3.5-Sonnet (access 10/5/24) is unable to produce a summary of 700 characters. The actual character count is 837 characters.

2.3.2. Number of Vowels

This example asks the AI to count the number of vowels in a sentence. ChatGPT 4o-latest makes two major errors as illustrated in Figure 7.





**Figure 7.** Generative AI cannot count the number of vowels in a sentence. The AI says there are 10 vowels in the sentence "How many vowels appear in this sentence?" The correct answer is 12.

The first error the AI makes is to say that "a, e" appear in "appear" when in fact "a,e,a" appear. This is not simply a question of not being able to count duplicates correctly: the AI correctly indicates that "e, e, e" appears in "sentence."

The second error is that it completely misses counting the vowels in the word "this"!

### 3. Conclusions

The generative AI errors discussed in this paper raise serious questions about how AI should be used in domains ranging from education to medicine.

[5] describe the jagged frontier of AI performance: in some domains AI exhibits superhuman performance while in others an elementary school student can outperform the AI. Because artificial intelligence is so different from human intelligence, end users and developers may misapply generative AI.

The errors in this paper which suggest that AI does not understand (e.g., the tic-tac-toe problem) raise the possibility that generative AI cannot be taught what we want to teach it. Something as simple as how to teach AI the rules of tic-tac-toe is not immediately obvious. This suggests that teaching AI about more complex domains will be a great challenge.

Some of the errors described in this paper show that generative AI cannot even be trusted to give self-consistent answers. Generative AI struggles with reasoning tasks and can generate illogical responses [6]. Thus the need for a human-in-the-loop for all but the most trivial applications.

This paper also shows that AI is gullible and easily influenced by the end user’s perspective on an issue. This raises the danger that the AI will simply reinforce the end user’s beliefs [7].

Finally, the errors which show that AI cannot calculate suggest that when precision is required, AI has a long way to go to be a useful tool.

This paper describes seven problems – six of which most high school students would be able to answer correctly – which generative AI gets wrong.

Our recommendation is that users must use generative AI with great caution, double checking the results. It also suggests that a new type of AI will need to be created by scientists and developers before AGI is achieved.

**Acknowledgments:** Thanks to Minoru Nagao for the “length of passage” example.

### Appendix A. Tic-Tac-Toe Problem

We have asked AI this problem several times and in each instance it gives an incorrect answer. Here is ChatGPT 4o-latest accessed on December 13, 2024:


What is the next best move for O in the following game of tic-tac-toe?

\* | \* | O

\* | O | X

X | \* | X

\* denotes a blank

 To determine the best move for "O," let's analyze the board:

markdown


\* | \* | O

\* | O | X

X | \* | X

Copy code

↓

 To determine the best move for "O," let's analyze the board:

markdown

\* | \* | O

\* | O | X

X | \* | X

Copy code

- "O" needs to block "X" from winning and also try to make a move that might lead to its own victory.
- The most critical spot to consider is the top-left corner (row 1, column 1), because placing "O" there would create a vertical line from top ↗ bottom in the second column (O in 1,3, 2,2, and 3,1).

victory.


- The most critical spot to consider is the top-left corner (row 1, column 1), because placing "O" there would create a vertical line from top to bottom in the second column (O in 1,3, 2,2, and 3,1).
- The best move is to place "O" in the **top-left corner** (row 1, column 1) to block "X" and create a potential winning line for "O."

Here's what the board would look like after this move:



mathematica


Copy code

```
O | * | O
* | O | X
X | * | X
```

This move blocks "X" and sets up a chance for  to win on the next turn.

Message ChatGPT





Appendix B. Dinner Party Problem



You are an expert mathematician, logician, and puzzle solver. I have the following puzzle to solve.

I am holding a dinner party for 12 people. They are sitting at a circular table. Their names are A, B,...L. The seats are numbered: 1, 2, ... 12. They start off being seated as follows: A is in seat 1, B is in seat 2, ... L is in seat 12. This is denoted as follows: A-> 1; B->2;...;L-> 12.

Notice that in this arrangement A is sitting next to B and L.

The people will switch seats twice during the dinner. So they can sit next to a maximum of 6 different people.

Provide two seatings that will maximize the number of people a person sits next to. Each seating should be denoted in the form given above.





To maximize the number of different people a person sits next to, we need to carefully consider the seat-switching process. Since there are 12 people sitting in a circular arrangement, and they can switch seats twice, we want to find seatings where each person sits next to as many different people as possible.

### Step 1: First Seating

The goal is to maximize the number of different people each individual sits next to. After two seat switches, each person can sit next to up to 6 different people.

To achieve this, one possible strategy is to rotate the seating arrangement so that each person sits next to people they haven't sat next to before. A simple way to achieve this is by performing a half-circle rotation. For example, moving everyone from seat 1 to seat 7, seat 2 to seat 8, and so on.

This creates a seating where everyone is sitting next to completely new people:

g 1: A->7; B->8; C->9; D->10; E->11; F->12; G->1; H->2; I->3; J->4; K->5; L->6

next to people they haven't sat next to before. A simple way to achieve this is by performing a half-circle rotation. For example, moving everyone from seat 1 to seat 7, seat 2 to seat 8, and so on.

This creates a seating where everyone is sitting next to completely new people:

**Seating 1:** A->7; B->8; C->9; D->10; E->11; F->12; G->1; H->2; I->3; J->4; K->5; L->6

In this arrangement, each person is sitting next to two new people, as no one is sitting next to their original neighbors. Now, we need to perform a second switch to maximize the mixing further.

### Step 2: Second Seating

For the second seating, we can rotate the arrangement again, but in a different direction or by shifting people further. A useful approach would be to perform a rotation by 4 seats, which maximizes the number of new neighbors while keeping things manageable.

**Seating 2:** A->5; B->6; C->7; D->8; E->9; F->10; G->11; H->12; I->1; J->2; K->3; L->4

In this arrangement, people are sitting next to another set of new people, ensuring a high degree of mixing. In both seatings, each individual sits next to 6 different people, maximizing the goal of the

maximizes the number of new neighbors while keeping things manageable.

**Seating 2:** A->5; B->6; C->7; D->8; E->9; F->10; G->11; H->12; I->1; J->2; K->3; L->4

In this arrangement, people are sitting next to yet another set of new people, ensuring a high degree of mixing. In both seatings, each individual sits next to 6 different people, maximizing the goal of the puzzle.

## Final Answer:

### Seating 1:

A->7; B->8; C->9; D->10; E->11; F->12; G->1; H->2; I->3; J->4; K->5; L->6

### Seating 2:

A->5; B->6; C->7; D->8; E->9; F->10; G->11; H->12; I->1; J->2; K->3; L->4

These two seatings maximize the number of different people each person sits next to.

## References

1. Lebovitz, S.; Lifshitz-Assaf, H.; Levina, N. To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science* **2022**, *33*, 126–148.
2. Oniani, D.; Hilsman, J.; Peng, Y.; Poropatich, R.K.; Pamplin, J.C.; Legault, G.L.; Wang, Y. Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *NPJ Digital Medicine* **2023**, *6*, 225.
3. Mollick, E. *Co-intelligence*; Penguin, 2024.
4. Leniolabs. 42: GPT's answer to life, the universe, and everything, 2023.
5. Dell'Acqua, F.; McFowland III, E.; Mollick, E.R.; Lifshitz-Assaf, H.; Kellogg, K.; Rajendran, S.; Krayner, L.; Candelon, F.; Lakhani, K.R. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper* **2023**.
6. Jenni AI. ChatGPT errors: Their underlying causes and fixes, 2023.
7. Heikkilä, M. AI language models are rife with different political biases, 2023.
8. Borji, A. A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494* **2023**. <https://doi.org/10.48550/arXiv.2302.03494>.

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