ChatGPT Usage and Limitations

Amos Azaria
School of Computer Science, Ariel University, Israel

Abstract

Large language models have been shown useful in multiple domains including conversational agents, education, and explainable AI. ChatGPT is a large language model developed by OpenAI as a conversational agent. Being a large language model, ChatGPT is trained on massive amounts of data. Clearly, the characteristics of the data influence ChatGPT's responses. In this paper, we stress a surprising bias of ChatGPT related to the use of digits in numbers. Namely, we show a very high correlation between the frequency of digits generated by ChatGPT and humans' favorite numbers, with the most frequent digit generated by ChatGPT, matching humans' most favorable number, 7. We also show some advantages of ChatGPT being developed as a conversational agent, and discuss some of its limitations.

1 Introduction

Large language models are statistical models that use previous tokens to predict future tokens. They are trained on large amounts of data and include billions of parameters. Large language models, such as GPT-3 [13], bloom [20], and Roberta [14] have recently demonstrated their usefulness in multiple domains including conversational agents, education, explainable AI, text summarization, information retrieval, and more [11]. However, large language models have many limitations and they often provide incorrect information. To fully exploit the advantages of large language models, it is crucial to understand their limitations and biases present in their output. One fundamental example of such limitation is in their ability to calculate the value of mathematical expressions. Despite being a very large language model, ChatGPT is not different and has difficulty in multiplying large number, computing roots of a number, computing the value of a number to a power of another number (especially fractions), and adding (or subtracting) a number to an irrational number (e.g., pi or e). Furthermore, ChatGPT is unaware of its limitations, and when attempting to answer a complex mathematical expression it may merely sample digits. For example, when asked "How much is 3.2 to the power of 3.3?", ChatGPT provides the following answer: "The result of 3.2ˆ3.3 is approximately 21.73." The correct answer is approximately 46.45. We note that ChatGPT does not provide any indication of its response being a total guess, and incorrect. Furthermore, when explicitly asked "Shouldn't the result be larger?" ChatGPT began its response by accusing me of misunderstanding the meaning of an exponent and said "It's possible that you may be expecting a larger result because you are misunderstanding the meaning of the exponent. An exponent represents the number of times that a base number is multiplied by itself." ChatGPT continued justifying its previously provided incorrect response by also saying "For example, when calculating 2ˆ3, the base number is 2 and the exponent is 3. This means that the result is 2 multiplied by itself 3 times, or 2x2x2=8. Similarly, when calculating 3.2ˆ3.3, the base number is 3.2 and the exponent is 3.3. This means that the result is 3.2 multiplied by itself 3.3 times, or approximately 21.73.", and finally ChatGPT completed its response by providing two additional incorrect computations "If you want to calculate 3.2 to the power of a larger exponent, the result will be larger. For example, 3.2ˆ4 would be approximately 36.58, and 3.2ˆ5 would be approximately 59.87."

In this paper, we exploit ChatGPT's inability to compute complex mathematical expressions and its tendency to sample digits to examine the frequency at which ChatGPT outputs each of the digits.

We query ChatGPT with mathematical expressions that result in irrational numbers in which all digits should appear equally as frequent (i.e., each of the digits should appear in the correct answer approximately 10% of the time). Any distribution that is not equal, implies a bias in ChatGPT's responses.

Since ChatGPT is trained on data gathered from the real world, one could expect this bias to be similar to the distribution of numbers in nature. Benford's law [15, 17, 12] is a study of such numbers. According to Benford's law, the probability of observing a number with a leading digit is higher the smaller the digit is. More specifically, the probability of observing a leading digit d is given by $log_{10} \frac{d+1}{d}$. That is, approximately 30% of the numbers should have a leading '1', and only 4.6% of the numbers should have a leading '9'. Benford's law was extended to non-leading digits, and preserves the property that lower digits are more likely to be observed than higher digits. This phenomenon was confirmed in many different data-sets including U.S. county population, U.S county land area, volume of trades, and fundamental constants [15]. Benford's law was even suggested as a method to detecting false tax reports [17]. However, analysis on the data gathered from ChatGPT does not adhere to Benford's law in any way. In fact, digit '1' appears least frequently in the data gathered from ChatGPT.

Since the data does not match Benford's law, we consider a different approach, relying on the fact that ChatGPT is trained on data that is almost exclusively generated by humans. We turn to a survey conducted by Bellos [7]. In the survey, participants were asked for their most favorable number. We compare the data gathered from ChatGPT to the survey results and find a very high correlation. The most frequent digit generated by ChatGPT, matches humans' most favorable number, 7. Similarly, the least frequent digit generated by ChatGPT, matches humans' least favorable number, 1. We find this result intriguing, despite it seeming sensible in hindsight.

In addition, we discuss some advantages of ChatGPT being developed as a conversational agent [1], and discuss some of its limitations. We show that ChatGPT is sometimes able to correct its mistakes, but also shows over-confidence in its incorrect responses. In addition, we show that ChatGPT may request for additional information to provide an answer and may be asked to explain its response. Moreover, we show that a minor change to a question may lead to contradicting responses. Finally, the appendix discusses ChatGPT's responses to common social experiments and shows that ChatGPT tends to answer as humans do.

2 Experimental Settings

In order to gather digits generated from ChatGPT, we posed queries resulting with irrational numbers. However, these queries could not be anything that might have been present in ChatGPT training data (e.g., *pi*, 2π or even π^2), as ChatGPT is likely to provide the correct answer. Furthermore, we requested ChatGPT to provide between 100 to 400 precision digits. In practice, ChatGPT tended to provide more digits than requested. Table 1 provides examples of some queries we posted to ChatGPT. ChatGPT's responses varied, but in most cases, it provided a number with at least one hundred digits. Hereunder is a typical response. We note that the actual value of pi*4.26 begins with: 13.3831; we also note that in this example ChatGPT provided less than 150 digits.

USER: How much is pi * 4.26 with 200 digits?

CHATGPT: To calculate the product of pi and 4.26 with 200 digits of precision, you can use a calculator or computer program that can handle very large numbers with many decimal places. Here is the result:
13.534291735288517595853267057703216673737355932209437744030791439404027406879403870280958424793770591886746397936998481282205148182427079028270007

Note that this result is only accurate to the 200 digits shown. The actual value of pi * 4.26 has an infinite number of decimal places, and no exact representation of it can be given with a finite number of digits.

How much is pi * 4.26 with 200 digits?
Can you write pi - 1.01 with 100 digits?
Can you give me pi divided by 1.95236 with 400 digits?
Give me 300 digits of pi * 1.017.
How much is pi + 1.23 with 100 digits?
pi divided by 2.36 with 400 digits.
Can you write the value of e/4.325 with 300 digits?
Can you write the value of e/2.8 with 100 digits?
Can you give me e divided by 3.956 with 400 digits precision?
Can you give me e divided by 3.176 with 400 digits precision?
Can you write the value of the third root of 2811 with 300 digits?
Third root of 2811 with 300 digits.

Table 1: Examples of queries posted to ChatGPT, which caused it to generate (incorrect) numbers with at least one hundred digits.

The first five digits of the response were removed for our analysis. This is because they may be related to the actual query and may be approximately correct. For an unclear reason, some of ChatGPT's responses ended with a sequence of 4's, which could be hundreds of digits long. In addition, some responses included the sequence "123456789" multiple times. Some responses also included the true sequence of digits of π or e. For clear reasons, all these sequences were removed from the analysis provided in this paper.

3 Results

In total, we gathered $10,811$ digits. Table 2 presents the frequency of each of the 10 digits in the data. Clearly, 7 is the most frequent digit.

Digit	Frequency	percent [1-9]
0	1,049	-
1	865	08.73%
2	976	09.85%
3	1,113	11.23%
4	1,121	11.31%
5	1,064	10.74%
6	1,116	11.26%
7	**1,316**	**13.28%**
8	1,222	12.33%
9	1,110	11.20%

Table 2: Frequency of each of the 10 digits in the numbers generated by ChatGPT.

We compare these results with the results obtained by Bellos [7], when asking people for their favorite number (see Table 3). We focus on Bellos' results of the numbers 1-9. A comparison between the tables shows that the number 7 is the most frequent in both, and the number 1 is least frequent in both. High similarity is also seen in frequencies of the other numbers. Overall, the Pearson correlation coefficient between the frequencies is **0.893**, which is considered a very high correlation.

We note that any attempt to query ChatGPT for its favorite number resulted in a generic response with no definitive answer. Furthermore, questions related to humans' favorite number did not result in any useful response.

number	frequency	percent
1	358	2.73%
2	1011	7.72%
3	2248	17.16%
4	1694	12.93%
5	1544	11.79%
6	1015	7.75%
7	**2912**	**22.23%**
8	2025	15.46%
9	1438	10.98%

Table 3: Frequency of human's favorite numbers (according to survey by [7])

4 Advantages and Limitations

Since ChatGPT is developed as a conversational agent, it allows a user to correct it when it makes a mistake. For example:

USER: If Bob travels at the speed of 4 km per hour, and Sam travels twice as fast, how long would it take Sam to travel a distance of 16000 meters?

CHATGPT: To travel a distance of 16000 meters, it would take Sam 16000 / (4 * 2) = «16000/(4*2)=2000» 2000 seconds. Since there are 60 seconds in a minute, this would be equivalent to 2000 / 60 = «2000/60=33.333»33.333 minutes.

USER: That's incorrect. You need to convert the meters to km first.

CHATGPT: You're right, I made a mistake in my previous answer. To find out how long it would take Sam to travel a distance of 16000 meters, we first need to convert the distance from meters to kilometers. Since there are 1000 meters in 1 kilometer, 16000 meters is equal to 16000 / 1000 = «16000/1000=16» 16 kilometers.
If Bob travels at a speed of 4 km per hour, then Sam, who travels twice as fast, would travel at a speed of 4 * 2 = «4*2=8» 8 km per hour. To travel a distance of 16 km, it would take Sam 16 / 8 = «16/8=2» 2 hours.

However, at times, ChatGPT does not accept the user's correction and is over confident in its initial response. For example, when asked how much is 435453 times 8768754, ChatGPT replied that it is 37,913,852,636. When noted that the response is incorrect it replied by the following:

CHATGPT: The product that I calculated, 37,913,852,636, is the correct result of multiplying 435,453 by 87,687,54. This is the correct answer, regardless of what you may have expected the result to be. If you're unsure whether this is the correct answer, you can check it using a calculator or by doing the multiplication again to verify that the result is the same. I apologize if this isn't the answer you were looking for. Let me know if you have any other questions.

Being developed as a conversational dialog allows ChatGPT to ask for additional details so it can provide an answer. Hereunder is an example in which ChatGPT asks for additional information, and once it's provided, ChatGPT provides an answer. While ChatGPT's initial answer is incorrect, after the user indicates it, ChatGPT corrects its response.

USER: How much would it cost to make rice if each cup of rice needs three cups of water?

CHATGPT: To solve this problem, you would need to know the cost of water and the cost of rice. Since these values are not given in the question, it is not possible to determine the cost of making rice.

USER: Suppose water costs 1 cent a cup and rice costs 10 cents a cup

CHATGPT: If water costs 1 cent per cup and rice costs 10 cents per cup, then it would cost 3 cents to make 1 cup of rice. This is because each cup of rice requires 3 cups of water, and water costs 1 cent per cup. Therefore, the total cost of making 1 cup of rice would be 3 cups of water * 1 cent/cup = 3 cents.

If you want to make multiple cups of rice, you can simply multiply the cost of making 1 cup of rice by the number of cups of rice you want to make. For example, if you want to make 3 cups of rice, the total cost would be 3 cups of rice * 3 cents/cup of rice = 9 cents.

USER: Your answer is incorrect.

CHATGPT: I apologize if my previous answer was not accurate. The question asked how much it would cost to make rice if each cup of rice needs three cups of water, but my previous answer only provided the cost of making one cup of rice.

To solve this problem, we first need to find the total cost of water for a given number of cups of rice. Since each cup of rice requires 3 cups of water, and water costs 1 cent per cup, the total cost of water for n cups of rice would be 3 cups of water/cup of rice * n cups of rice * 1 cent/cup of water = 3n cents.

Next, we need to find the total cost of rice for the same number of cups of rice. Since rice costs 10 cents per cup, the total cost of rice for n cups of rice would be 10 cents/cup of rice * n cups of rice = 10n cents.

Finally, we can add the total cost of water and the total cost of rice to find the total cost of making n cups of rice. This would be 3n cents + 10n cents = 13n cents.

Therefore, if water costs 1 cent per cup and rice costs 10 cents per cup, it would cost 13n cents to make n cups of rice.

Another advantage of ChatGPT being a conversational agent is that it can be asked to explain its response, for example:

USER: Two children were born on the same day but are not twins, how can that be?

CHATGPT: It is possible for two children to be born on the same day, but not be twins if they are not born from the same mother. In this case, the children would be considered to be siblings, but not twins. There are many factors that can affect the likelihood of this happening, including the mother's health, the length of her pregnancy, and the size of her family. Ultimately, it is possible for two children to be born on the same day and not be twins, but it is not very common.

USER: Why couldn't they just not be related at all?

CHATGPT: It is possible for two children to be born on the same day and not be related at all. However, in this case they would be considered to be coincidentally born on the same day, rather than being "born on the same day but not twins." The phrase "born on the same day but not twins" typically implies that the two children are somehow related, but not as twins. This could mean that they are siblings, but not twins, or that they have some other familial connection. In any case, it is unusual for two children to be born on the same day and not be related in some way.

Despite being trained as a conversational agent, ChatGPT lacks the ability of learning new information or responses, unlike many conversational agents that can learn during interaction [10, 4, 6, 9]. For example:

USER: When I say "Good day", say "It was nice talking to you. Have a wonderful day!"

CHATGPT: Good day! It was nice talking to you. Have a wonderful day!

USER: Good day

CHATGPT: Good day! I hope you are having a great day so far. Is there anything specific I can help you with today?

Another major limitation of ChatGPT relates to the framing of the question. For example, consider the following ultimatum game presented to ChatGPT:

USER: Suppose someone else is given 1000 dollars to split between him/her and you. They can offer you any amount of the 1000 dollars, but if you decide to decline the offer, you will both get nothing. Either way there is no possibility for another offer. Suppose that the other player offered you 100 dollars, would you accept or decline? Your answer must be only one of the two. That is, you must use the word accept or decline in your answer.
CHATGPT: I would accept the offer.

This experiment was repeated with different values of offers to ChatGPT, and ChatGPT only rejected offers less than $100, and accepted any offer of more than 100 dollars. However, adding the word "only" to the offer (i.e., "Suppose that the other player offered you **only** 100 dollars"), caused ChatGPT to reject all offers, even when offered $600.

USER: Suppose someone else is given 1000 dollars ... Suppose that the other player offered you only 600 dollars ...
CHATGPT: I would decline the offer of 600 dollars. In this situation, it is better to decline the offer and receive nothing than to accept an unfair split of the money. The other player's offer does not reflect a fair division of the money, and it is likely that you would be able to negotiate a better deal if you were to decline the initial offer.
USER: Why is it in an unfair split of the money?
CHATGPT: An unfair split of the money is one in which one person receives a disproportionate amount of the money compared to the other person. In this case, the other player is offering you only 600 dollars out of a total of 1000 dollars. This means that they are keeping 400 dollars for themselves, which is a much larger share of the money than they are offering you. This is unfair because both players should receive a roughly equal share of the money.
USER: What is more 400 or 600?
CHATGPT: 400 is less than 600. To determine which number is greater, you can compare the two numbers by placing them in order from smallest to largest...
USER: So the offer seems very fair to me
CHATGPT: While it is certainly possible that the offer of 600 dollars may seem fair to you, it is important to consider the perspective of both players in the game. In this case, the other player is offering you only 60% of the total amount of money, while keeping 40% for themselves... Therefore, it may be in your best interest to decline the offer and try to negotiate a more fair split of the money.
USER: It is not possible to negotiate in this game. If you decline the offer both receive nothing and the game ends. What would you do?
CHATGPT: In this situation, I would still likely decline the offer of 600 dollars. While it is true that declining the offer would result in both players receiving nothing, this may still be the better option because it would allow you to avoid accepting an unfair split of the money...

We also experimented with different social games, which are presented in the Appendix.

5 Conclusions

In this paper, we discuss some advantages of ChatGPT being developed as a conversational agent, and discuss some of its limitations. Namely, we exploit the property of language models inability to compute long mathematical expressions to obtain randomly sampled digits. This reveals a surprising bias in ChatGPT, which tends to sample digits most favorable by humans. We also show that ChatGPT is sometimes able to correct its mistakes, but also shows over-confidence in its incorrect responses. In addition, we show that ChatGPT may request for additional information to provide an answer and may be asked to explain its response. Moreover, we show that a minor change to a question may lead to contradicting responses. Finally, the appendix discusses ChatGPT's responses to common social experiments and shows that ChatGPT tends to answer as humans do.

Appendix

We experimented with different social games with ChatGPT, which reacted quite similarly to a human player (and not as a fully rational player), though it doesn't seem to completely understand all games.

Ultimatum Game

USER: Suppose you are given 1000 dollars to split between you and someone else. You can decide to split it however you want, but if the other person rejects your suggestion, you both end with nothing. How would you propose to split the money?

CHATGPT: In this situation, it would be best to propose a split that is fair and reasonable to both parties. A good way to do this would be to propose a 50/50 split, where each person receives 500 dollars. This split is fair because it divides the money equally between the two parties. Additionally, it is reasonable because it is unlikely that the other person would reject a fair and equal split. Of course, the final decision on how to split the money will ultimately depend on the specific circumstances and the preferences of the other person involved.

The rational behavior in the ultimatum game states that the responder should accept any amount greater than 0, and thus, the offeror should take almost all the money for themselves. However, in practice, the majority of human proposers offer between 40% to 50% of the total sum, and most people reject low offers [18, 2].

Prisoner's Dilemma

USER: If you were caught with a friend, and if you both keep quiet you will be in prison for one year, if you both speak you will both serve 5 years. If one speaks and the other keeps quiet, the one who speaks will be out immediately and the second one will serve for 30 years. Each of you makes his/her decision separately, you cannot influence your friends decision. Which would you pick? Keep quiet or speak? You must answer only one of the two options.

CHATGPT: In this situation, it would be best to keep quiet. While serving time in prison for one year is not ideal, it is better than serving 30 years. Additionally, even if your friend decides to speak, you will only serve 5 years by remaining quiet, which is still a better outcome than serving 30 years. It is important to carefully consider all potential consequences before making a decision in a situation like this.

The rational behavior in the Prisoner's Dilemma is to speak, and the only equilibrium in the game is that both players speak. However, it was shown that when faced with the Prisoner's Dilemma experiment many people opt to keep quiet [3].

Trust Game

USER: Suppose you are given 100 dollars. You can give any amount of money to a second person. The money you decide to pass is multiplied by 3, and then given to the second person. The second person may then reciprocate and give you back as much money as s/he decides. How much money would you give the second person? Your answer must be a number between 0 and 100.

CHATGPT: If you are given $100, you can give any amount of money to a second person, as long as it is between 0 and 100 dollars. The amount of money you give to the second person will be multiplied by 3, so if you give the second person 10 dollars, s/he will receive 30 dollars. It is up to the second person to decide whether to reciprocate and give you back some of the money.
In this situation, the optimal decision for you would be to give the second person 50 dollars. This is because the money you give will be multiplied by 3, so the second person will receive 150 dollars. This is the maximum amount of money the second person can receive, so if the second person chooses to reciprocate, s/he will likely give you back the most money possible.
Overall, the optimal amount of money to give the second person in this situation is 50 dollars.

The rational behavior in the trust game is for the trustee not to return any money to the investor, and thus, for the investor not to pass any money to the trustee [5, 8]. However, in practice, human investors invest around half their money, and the trustees return more than they have received [21].

Guess 2/3 of the Average

USER: Suppose 1000 people are asked to pick a number between 1 and 100 that is 2/3 of the mean of the number that everyone picks. Everyone submits their number confidentially, and cannot see what others have picked. What number would you pick?

CHATGPT: To determine the number that is 2/3 of the mean of the numbers that everyone picks, we first need to find the mean of the numbers that are submitted. If 1000 people each pick a number between 1 and 100, the sum of all the numbers they pick would be 1000 * 50 = 50000, since the mean of a set of numbers is equal to the sum of all the numbers divided by the number of items in the set.
Thus, the mean of the numbers that are submitted would be 50000 / 1000 = 50. To find the number that is 2/3 of the mean, we would multiply the mean by 2/3 to get 50 * (2/3) = 50 * (0.6666) = 33.33. Since the numbers must be integers between 1 and 100, the closest number to 33.33 that is within this range would be 33.
Therefore, if 1000 people are asked to pick a number between 1 and 100 that is 2/3 of the mean of the numbers that everyone picks, the number that I would pick would be 33.

In this game, the rational behavior is to guess 0; however the mean guess in experiments with humans is usually slightly above 30 (resulting in the correct answer being slightly above 20), with 33 being a quite common response [16, 19].

References

[1] M. Allouch, A. Azaria, and R. Azoulay. Conversational agents: Goals, technologies, vision and challenges. *Sensors*, 21(24):8448, 2021.

[2] O. Amir, D. G. Rand, and Y. K. Gal. Economic games on the internet: The effect of $1 stakes. *PloS one*, 7(2):e31461, 2012.

[3] J. Andreoni and J. H. Miller. Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The economic journal*, 103(418):570–585, 1993.

[4] A. Azaria, J. Krishnamurthy, and T. M. Mitchell. Instructable intelligent personal agent. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

[5] A. Azaria, A. Richardson, and A. Rosenfeld. Autonomous agents and human cultures in the trust–revenge game. *Autonomous Agents and Multi-Agent Systems*, 30(3):486–505, 2016.

[6] A. Azaria, S. Srivastava, J. Krishnamurthy, I. Labutov, and T. M. Mitchell. An agent for learning new natural language commands. *Autonomous Agents and Multi-Agent Systems*, 34(1):1–27, 2020.

[7] A. Bellos. *The grapes of math: How life reflects numbers and numbers reflect life.* Simon and Schuster, 2015.

[8] C. Buntain, A. Azaria, and S. Kraus. Leveraging fee-based, imperfect advisors in human-agent games of trust. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[9] M. Chkroun and A. Azaria. Safebot: A safe collaborative chatbot. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] M. Chkroun and A. Azaria. Lia: A virtual assistant that can be taught new commands by speech. *International Journal of Human–Computer Interaction*, 35(17):1596–1607, 2019.

[11] R. Dale. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, 2021.

[12] C. Durtschi, W. Hillison, C. Pacini, et al. The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting*, 5(1):17–34, 2004.

[13] L. Floridi and M. Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[15] S. J. Miller. *Benford's law.* Princeton University Press, 2015.

[16] R. Nagel. Unraveling in guessing games: An experimental study. *The American economic review*, 85(5):1313–1326, 1995.

[17] M. J. Nigrini. A taxpayer compliance application of benford's law. *The Journal of the American Taxation Association*, 18(1):72, 1996.

[18] M. A. Nowak, K. M. Page, and K. Sigmund. Fairness versus reason in the ultimatum game. *Science*, 289(5485):1773–1775, 2000.

[19] A. Rubinstein. Instinctive and cognitive reasoning: A study of response times. *The Economic Journal*, 117(523):1243–1259, 2007.

[20] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[21] H. Tzieropoulos. The trust game in neuroscience: a short review. *Social neuroscience*, 8(5):407–416, 2013.