BEYOND THE HYPE: A CRITICAL EVALUATION OF CHATGPT'S CAPABILITIES FOR MATHEMATICAL CALCULATIONS

Ewuradjoa Mansa Quansah[1,2]

[1]Department of Strategic and International Management, Graduate School of Management Saint Petersburg University, Saint Petersburg, Russian Federation
[2]International Business School, Teesside University, Middlesbrough, United Kingdom

ABSTRACT

As generative AI systems like ChatGPT gain popularity, empirical analysis is essential to evaluate capabilities. This study investigates ChatGPT's skills for mathematical calculations through controlled experiments. Tests involving counting numbers, finding averages, and demonstrating Excel methods reveal inconsistencies and errors, indicating lack of true contextual understanding. While ChatGPT can provide solutions, its reasoning shows gaps versus human cognition. The results provide concrete evidence of deficiencies, complementing conceptual critiques. Findings caution against over-reliance on generative models for critical tasks and highlight needs to advance reasoning and human-AI collaboration. This analysis contributes to AI literature by urging continued progress of AI, so technologies like ChatGPT can be deployed safely and responsibly.

KEYWORDS

ChatGPT, Artificial intelligence, AI, Generative AI, large language-based models, experiment

1. INTRODUCTION

In the contemporary landscape of digital transformation, large language-based models have experienced a remarkable surge in adoption. Artificial intelligence (AI) has initiated a profound transformation in our approach to work across all facets of life [1]. These transformative waves transcend industries, impacting everything from education and healthcare to the broader spectrum of professional endeavours[2]. Notably, language models like ChatGPT, Model AI, Bard, Replica AI, among others, are rapidly ascending in prominence, with professionals across diverse fields swiftly integrating these technologies into their work.

The advantages of deploying large language models and generative AIs are substantial, as emphasized by a growing body of literature [3], [4]. These benefits encompass heightened operational efficiency, cost savings through reduced reliance on frequent chatbot updates, and the facilitation of creative brainstorming, among various others. Generative AI tools, including ChatGPT, demonstrate the capacity to handle complex tasks [5], expanding their utility across a spectrum of professional functions.

David C. Wyld et al. (Eds): AIBD, MLSC, ACSTY, NATP, CCCIoT, SVC, SOFE, ITCSS -2024
pp. 87-95, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140407

However, in tandem with the immense traction garnered by these technologies, critical voices have emerged, elucidating limitations and concerns.Baidoo-Anu and Owusu Ansah[5]underscore some of the technological constraints. Despite the widespread adoption of generative AI tools since their release in 2022, an array of scholars harnesses these tools to enhance teaching, learning, and research, while professionals employ them to address everyday work challenges [4].

Yet, amid the myriad advantages of ChatGPT and its peers, it is vital to acknowledge the presence of limitations and concerns regarding their application in diverse contexts. Over-reliance on these language models without conducting a rigorous critical analysis of their outputs may result in unintended consequences. Therefore, this study seeks to address two essential research questions: first, what are the pitfalls associated with generative AI tools like ChatGPT? Second, how can these pitfalls be effectively mitigated within the realm of professional work?

In making these inquiries, we aim to contribute to the field of information systems research by shedding light on the inherent limitations of utilizing generative AI tools, like ChatGPT. Our objective is to enhance awareness, prompt cautious and vigilant use, and deter over-dependence on these tools. Ultimately, we intend to guide professionals and scholars in navigating the transformative landscape of AI with a nuanced understanding of both its potential and its limitations.

2. LITERATURE REVIEW

Recent advances in artificial intelligence have led to the rapid development of large language models like ChatGPT. ChatGPT was developed by OpenAI, and it is based on the GPT-3.5 architecture. The release date for GPT-3 was in June 2020, and there have been subsequent versions and improvements since then[3], [4], [6]. Since its launch, ChatGPT has quickly gained popularity for its ability to generate remarkably human-like text on a wide range of topics [7]–[9].

ChatGPT is powered by a machine learning technique called transformers and trained on vast datasets scraped from the internet[10]–[12]. This allows it generate responses to natural language prompts with coherence and relevance, though it lacks human reasoning. ChatGPT and similar large language models are referred to as generative AI[9], [13].

Several studies have highlighted the potential benefits of ChatGPT and generative AI in education, business, and healthcare. These include democratizing access to information, increasing productivity through automation, and supporting creativity [4], [5], [8], [11].However, significant risks and limitations have also been identified.

Key risks of generative AI include potential to spread misinformation, copyright/plagiarism issues, security vulnerabilities, and embedded biases[11], [14]–[16] For example, ChatGPT can generate highly convincing but false content, posing a threat of 'weaponized misinformation' if misused[17], [18]. Regarding bias, NLP models like ChatGPT reflect problematic assumptions and stereotypes from training data [19].

In terms of limitations, ChatGPT lacks factual knowledge past 2021 and a grounded sense of truth or ethics[9]. Its capabilities are constrained by training data, lacking common sense and causal reasoning abilities[20], [21]. As such, outputs may be logically incoherent, useless, or harmful if not critically assessed[22].

These risks necessitate caution and ethical considerations when deploying generative AI in real world settings[11], [14]. Scholars emphasize the need for transparency, human oversight, and addressing harmful biases during development [23]–[25]. Overall, while the technology shows promise, more research is needed to develop robust, beneficial applications.

3. METHODOLOGY & RESULTS

3.1. Research Design

To address the research questions posed an experimental method was used. The experiment involved a simple yet sophisticated calculation of a problem of different mathematical problems using ChatGPT 3.5. This version was selected due to its widespread availability and use among individuals.

3.2. Method and Results

The experimental procedure began with a basic mathematical calculation, specifically counting the numbers within a given range. The initial command provided to ChatGPT was:

"count numbers in range
4+4+3+4+5+3+2+3+2+2+2+2+1+1+1+3+2+2+2+1+2+2+2+2+2+1+2+2+2+2+2+2+2"

ChatGPT's response was as follows:

"There are 32 numbers in the given list:
4, 4, 3, 4, 5, 3, 2, 3, 2, 2, 2, 2, 1, 1, 1, 3, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2"

However, a manual count indicated that there were 33 numbers, not 32 as suggested by the software.

To verify the accuracy of ChatGPT's feedback, we issued the command "are you sure?" and received the following response:

"Apologies, I made a mistake in my previous response. Let me recount the numbers for you:
There are 31 numbers in the given list:
4, 4, 3, 4, 5, 3, 2, 3, 2, 2, 2, 2, 1, 1, 1, 3, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2"

A subsequent command, "wait what is happening?", elicited the following response from ChatGPT:

"I apologize for the confusion. It seems there was a misunderstanding in the previous responses. Let me clarify:
To count the numbers in the given range without summing them up, we count how many numbers are listed:
4+4+3+4+5+3+2+3+2+2+2+2+1+1+1+3+2+2+2+1+2+2+2+2+2+1+2+2+2+2+2+2+2
There are 32 numbers in the given list. I apologize for any confusion caused. If you have any further questions or need assistance, feel free to ask."

The experiment was then elevated to a higher level of complexity, involving finding the average of a set of numbers.

To calculate the average, a series of random numbers was chosen, and the command "find the average" was issued, as demonstrated here:

"55+29+200+44+153+404+216+96+60+144+84+99+53+150+0+34+39+94+72+117+293+152+
121+128+128+111+34+39+36+77+41+14+18
find average"

ChatGPT computed the average by first summing the numbers and then dividing the sum by the total count of numbers. The result provided by ChatGPT was:

"To find the average of a set of numbers, you need to add up all the numbers and then divide the sum by the total count of numbers in the set. Let's do the calculations:

Sum of the numbers:
55 + 29 + 200 + 44 + 153 + 404 + 216 + 96 + 60 + 144 + 84 + 99 + 53 + 150 + 0 + 34 + 39 + 94 + 72 + 117 + 293 + 152 + 121 + 128 + 128 + 111 + 34 + 39 + 36 + 77 + 41 + 14 + 18 = 2758

Total count of numbers in the set: 33

Now, let's calculate the average:
Average = Sum of the numbers / Total count of numbers
Average = 2758 / 33 ≈ 83.58

So, the average of the given set of numbers is approximately 83.58."

This process was repeated with various sets of random numbers.

To ensure result accuracy, we cross-verified ChatGPT's responses using Excel and manual calculations. Additionally, we tested ChatGPT's ability to perform calculations in an Excel-like manner.

The results revealed occasional inconsistencies in the data interpretation by ChatGPT.

To calculate the average of the numbers in Excel, follow these steps:

1. Open your Excel spreadsheet.

2. Enter the numbers in one column, for example, in column A starting from cell A1.

A1: 3
A2: 3
A3: 2
A4: 2
A5: 2
A6: 2
A7: 3
A8: 2
A9: 3
A10: 3
A11: 4
A12: 3
A13: 2
A14: 3
A15: 3
A16: 3

A17: 2
A18: 3
A19: 1
A20: 2
A21: 2
A22: 2
A23: 2
A24: 2
A25: 3
A26: 2
A27: 2
A28: 2
A29: 3
A30: 3
A31: 2
A32: 3
A33: 2

3. Now, delete cell A34.

4. In a cell where you want the average to appear (for example, in cell B1), use the AVERAGE function to calculate the average of the numbers in column A. Type the following formula and press Enter:

B1: =ROUND(AVERAGE(A1:A33), 2)

Excel will display the average of the numbers in the specified range rounded to two decimal places. In this case, the average is approximately 2.39 (rounded to two decimal places).

However, ChatGPT misinterpreted the result and mistakenly began deleting the last number, which is a problematic error. A corrected approach is as follows:

To find the average of the numbers in the given range in Excel, follow these steps:

1. Open your Excel spreadsheet.

2. Enter the numbers in one column, for example, in column A starting from cell A1.

A1: 4
A2: 2
A3: 3
A4: 4
A5: 3
A6: 2
A7: 2
A8: 2
A9: 2
A10: 2
A11: 3
A12: 3
A13: 2
A14: 3

A15: 3
A16: 3
A17: 2
A18: 2
A19: 2
A20: 2
A21: 4
A22: 4
A23: 1
A24: 4
A25: 4
A26: 2
A27: 3
A28: 2
A29: 2
A30: 2
A31: 2

3. Now, delete cell A32.

4. In a cell where you want the average to appear (for example, in cell B1), use the AVERAGE function to calculate the average of the numbers in column A. Type the following formula and press Enter:

B1: =ROUND(AVERAGE(A1:A31), 2)

Excel will display the average of the numbers in the specified range rounded to two decimal places. In this case, the average is approximately 2.58 (rounded to two decimal places).

4. DISCUSSION

The experimental results reveal several important limitations and inconsistencies in ChatGPT's ability to perform basic mathematical calculations. While ChatGPT was often able to reach the right conclusion after multiple tries and clarification questions, its thought process showed a lack of true contextual understanding.One clear pattern was ChatGPT's tendency to make careless mistakes that required correction, such as miscounting the number of items in a list. This aligns with other studies showing AI systems can make obvious errors that humans would not [26], [27]. The need for clarification on simple tasks indicates issues with sound logical reasoning.Additionally, when asked to demonstrate Excel calculations, ChatGPT struggled to provide fully coherent explanations. It often omitted key steps or incorrectly described the approach, suggesting flawed mental models of the required procedures. This matches existing research indicating gaps between how humans and AIs solve problems [11]. ChatGPT's inconsistent performance shows risks of over-relying on it for tasks requiring accuracy. Findings reinforce that while generative AI can provide reasonable responses, it lacks robust understanding [9]. More concerningly, on real-world applications like financial analysis, medical diagnosis, or legal work, mistakes could have serious consequences.

Overall, the experiment reveals meaningful functionality gaps in ChatGPT's skills. While narrow AI systems can surpass human capabilities on specific tasks, general intelligence necessitates strong contextual reasoning and judgment [19]. Current generative models remain far from achieving broad human competencies. Caution is warranted in deploying them in impactful real-world settings.

5. RESEARCH CONTRIBUTIONS

This research makes several valuable contributions to the emerging body of literature on generative AI systems like ChatGPT. First, it provides direct experimental evidence of limitations in ChatGPT's ability to perform basic mathematical calculations. While other studies have speculated on deficiencies in reasoning, this work offers concrete examples of inconsistencies and errors. Second, the methodology demonstrates an approach to rigorously testing AI capabilities through controlled experiments. Assessing sophisticated models like ChatGPT requires creative techniques beyond standard benchmarks. The variety of mathematical tests and validation methods used here could inform future research aiming to characterize strengths versus shortcomings.Additionally, the findings contribute to ongoing discussions around the responsible development and deployment of generative AI. Demonstrating specific cases where ChatGPT fails complements ethical and philosophical arguments for caution adopting these technologies. Together, empirical studies and conceptual analyses will lead to safer innovation.Finally, this research adds to knowledge about interaction design for human-AI collaboration. The results show ChatGPT can arrive at solutions when challenged, hinting at the value of interfaces that allow back-and-forth clarification. More work is needed on how systems and users can cooperatively overcome inherent limitations.In a field marked by hype and speculation, this study provides grounding evidence on generative AI's capabilities. The contributions set the stage for further research into mitigating risks and developing beneficial applications of this powerful technology.

6. RESEARCH IMPLICATIONS

The findings from this study have several key implications both for the field of AI research and for professionals employing tools like ChatGPT. First, the results underscore the need for continued progress in robustness and reasoning for generative models. While capabilities are rapidly advancing, there remain substantial gaps compared to human cognition that limit reliability. Researchers must continue developing transparent and generalizable AI systems.For those utilizing ChatGPT and similar technologies, this work highlights the importance of maintaining human oversight and not fully outsourcing critical tasks. Professionals in fields like finance, medicine, and law need awareness that generative tools may provide seemingly sensible outputs that are logically incoherent or factually incorrect. Failing to independently validate conclusions drawn from AI assistance could lead to mistakes with serious consequences.More broadly, businesses and organizations should focus adoption of ChatGPT on use cases where occasional errors will not cause significant harm. Generative models can be very useful for brainstorming ideas, summarizing information, or prototyping - but lack the reasoning for advising high-stakes decisions without human supervision. Responsible leaders will proactively seek to maximize benefits while mitigating risks.At a societal level, this study contributes evidence to inform policies around AI transparency and algorithm audits. As generative models become more prevalent, regulators must balance innovation and ethical imperatives. Understanding specific limitations, like those surfaced in this work, will enable nuanced governance.

7. CONCLUSION

In conclusion, this experimental research highlights key weaknesses in ChatGPT's ability to reliably perform basic mathematical calculations. The results reveal inaccuracies and logical gaps indicating the technology remains far from human-level intelligence. Generative AI shows immense promise, but more development is required before these tools can be completely trusted for critical applications without oversight. By empirically demonstrating specific deficiencies, this work provides valuable grounding for future advancement and adoption. Research must continue exploring safe implementation of AI technologies in order to realize benefits while avoiding harmful mistakes or misuse of generative models.

REFERENCES

[1] M. Ruffolo, ‘The Role of Ethical AI in Fostering Harmonic Innovations that Support a Human-Centric Digital Transformation of Economy and Society’, in *Harmonic Innovation*, vol. 282, F. Cicione, L. Filice, and D. Marino, Eds., in Lecture Notes in Networks and Systems, vol. 282. , Cham: Springer International Publishing, 2022, pp. 139–143. doi: 10.1007/978-3-030-81190-7_15.

[2] A. Zarifhonarvar, ‘Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence’, *Available SSRN 4350925*, 2023.

[3] J. Deng and Y. Lin, ‘The benefits and challenges of ChatGPT: An overview’, *Front. Comput. Intell. Syst.*, vol. 2, no. 2, pp. 81–83, 2022.

[4] S. Sok and K. Heng, ‘ChatGPT for education and research: A review of benefits and risks’, *Available SSRN 4378735*, 2023, Accessed: Nov. 04, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4378735

[5] D. Baidoo-Anu and L. Owusu Ansah, ‘Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning (January 25, 2023)’, *Available SSRN 4337484*, 2023.

[6] ChatGPT, ‘ChatGPT’. Accessed: Nov. 04, 2023. [Online]. Available: https://chat.openai.com

[7] S. S. Gill and R. Kaur, ‘ChatGPT: Vision and challenges’, *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 262–271, 2023.

[8] A. Haleem, M. Javaid, and R. P. Singh, ‘An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges’, *BenchCouncil Trans. Benchmarks Stand. Eval.*, vol. 2, no. 4, p. 100089, 2022.

[9] P. P. Ray, ‘ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope’, *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Jan. 2023, doi: 10.1016/j.iotcps.2023.04.003.

[10] MIT Technology Review, ‘ChatGPT is everywhere. Here’s where it came from’, MIT Technology Review. Accessed: Nov. 05, 2023. [Online]. Available: https://www.technologyreview.com/2023/02/08/1068068/chatgpt-is-everywhere-heres-where-it-came-from/

[11] Y. K. Dwivedi *et al.*, ‘Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy’, *Int. J. Inf. Manag.*, vol. 71, p. 102642, Aug. 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[12] C. Raffel *et al.*, ‘Exploring the limits of transfer learning with a unified text-to-text transformer’, *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[13] A. Abid, M. Farooqi, and J. Zou, ‘Persistent Anti-Muslim Bias in Large Language Models’, in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,* Virtual Event USA: ACM, Jul. 2021, pp. 298–306. doi: 10.1145/3461702.3462624.

[14] R. Bommasani *et al.*, ‘On the Opportunities and Risks of Foundation Models’. arXiv, Jul. 12, 2022. Accessed: Nov. 05, 2023. [Online]. Available: http://arxiv.org/abs/2108.07258

[15] OECD, ‘Initial policy considerations for generative artificial intelligence’, OECD Artificial Intelligence Papers 1, Sep. 2023. doi: 10.1787/fae2d1e6-en.

[16] M. Vartak, ‘Six Risks Of Generative AI’, Forbes. Accessed: Nov. 05, 2023. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2023/06/29/six-risks-of-generative-ai/

[17] E. Bell, ‘A fake news frenzy: why ChatGPT could be disastrous for truth in journalism’, *The Guardian,* Mar. 03, 2023. Accessed: Nov. 05, 2023. [Online]. Available: https://www.theguardian.com/commentisfree/2023/mar/03/fake-news-chatgpt-truth-journalism-disinformation

[18] J. J. Koplin, ‘Dual-use implications of AI text generation’, *Ethics Inf. Technol.*, vol. 25, no. 2, p. 32, May 2023, doi: 10.1007/s10676-023-09703-z.

[19] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □’, in *Proceedings of the 2021 ACM Conference on*

Fairness, Accountability, and Transparency, Virtual Event Canada: ACM, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[20] B. Marr, ‘The Top 10 Limitations Of ChatGPT’, Forbes. Accessed: Nov. 05, 2023. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2023/03/03/the-top-10-limitations-of-chatgpt/

[21] H. Santos, K. Shen, A. M. Mulvehill, M. Kejriwal, and D. L. McGuinness, ‘A Theoretically Grounded Question Answering Data Set for Evaluating Machine Common Sense’, *Data Intell.*, pp. 1–29, Oct. 2023, doi: 10.1162/dint_a_00234.

[22] M. Abedi, I. Alshybani, M. R. B. Shahadat, and M. Murillo, ‘Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education’, *Qeios*, 2023, Accessed: Nov. 05, 2023. [Online]. Available: https://www.qeios.com/read/MD04B0

[23] W. H. Organization, ‘Ethics and governance of artificial intelligence for health: WHO guidance’, 2021, Accessed: Nov. 05, 2023. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/341996/9789240029200-eng.pdf

[24] R. Rodrigues, ‘Legal and human rights issues of AI: Gaps, challenges and vulnerabilities’, *J. Responsible Technol.*, vol. 4, p. 100005, Dec. 2020, doi: 10.1016/j.jrt.2020.100005.

[25] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, ‘Towards a standard for identifying and managing bias in artificial intelligence’, *NIST Spec. Publ.*, vol. 1270, no. 10.6028, 2022, Accessed: Nov. 05, 2023. [Online]. Available: https://view.ckcest.cn/AllFiles/ZKBG/Pages/264/c914336ac0e68a6e3e34187adf9dd83bb3b7c09f.pdf

[26] A. Szabo, ‘ChatGPT is a breakthrough in science and education but fails a test in sports and exercise psychology’, *Balt. J. Sport Health Sci.*, vol. 1, no. 128, pp. 25–40, 2023.

[27] A. Szabo, ‘ChatGPT fails a test’, 2023, Accessed: Nov. 05, 2023. [Online]. Available: https://osf.io/ks365/download

© 2024 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.