Discover Artificial Intelligence

Case study

Does GPT-3 qualify as a co-author of a scientific paper publishable in peer-review journals according to the ICMJE criteria? A case study

Almira Osmanovic-Thunström[1,2] · Steinn Steingrimsson[1,2]

Received: 22 December 2022 / Accepted: 16 January 2023
Published online: 06 April 2023
© The Author(s) 2023 OPEN

Abstract
This paper explores the potential for a system to be a co-author on an academic paper based on the criteria proposed by the International Committee of Medical Journal Editors (ICMJE). We used a third generation generative pretrained transformer (GPT-3) to write a review paper on the topic of its choice: The Effects of Sleep Deprivation on Cognitive Function. The system was asked to fulfill the four main criteria for co-authorship as recommended by ICMJE, which includes contributions to the conception or design of the work; drafting the work or revising it critically for important intellectual content; final approval of the version to be published; and agreement to be accountable for all aspects of the work. Our results showed that the system was able to fulfill the criteria, with significant difficulties with accurate and reliable referencing. We also explored the ethical implications of using AI systems for research and found that it is important to take a cautious approach when considering its use for scientific authorship. This case study provides a methodology for further investigations into the possibilities and limitations of automated writing.

1 Introduction

With the increasing interest and spread of the use of large language models (LLM) for academic writing, it is important to explore to what extent systems based on LLMs can be considered co-authors of the work they produce. Large language models (LLMs) have emerged as a promising approach to natural language processing (NLP). LLMs are based on deep neural networks that learn from large amounts of text data and generate new texts by predicting the next word given the context. LLMs have become increasingly popular due to their ability to capture long-term dependencies, represent complex language structures, and generate human-like text. For example, the GPT-3 model developed by OpenAI [1] was trained on an unprecedentedly large dataset of more than 45 TB of text and can generate remarkably coherent human-like text. The potential of LLMs has been explored in various NLP tasks such as machine translation [2], question answering [3], and text summarization [4]. Recently a peer-reviewed journal published a paper with an LLM (ChatGPT) as a main author [5]. This was the first issue to be officially published in a peer-reviewed journal that listed an LLM as a co-author. The subject of LLMs performing humanlike tasks such as writing tasks has been explored previously in other studies [6–10]. These explorations demonstrated the possibilities and limitations of automated writing using well grounded and interesting mappings, observations and tests that explore the technical, ethical and philosophical aspects of the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44163-023-00055-7.

✉ Almira Osmanovic-Thunström, almira.osmanovic.thunstrom@gu.se | [1]Institute of Neuroscience and Physiology, University of Gothenburg, Gothenburg, Sweden. [2]Department of Psychiatry, Region Västra Götaland, Sahlgrenska University Hospital, Gothenburg, Sweden.

system. The studies explored the duality of generative pretrained transformers in their totality. As precedent has been made in regards to co-authorship in an academic paper by an LLM [5], the aim of this paper is to explore an interesting but very limited scope: can an LLM such as GPT-3 be a co-author on an academic paper based on the four central criteria proposed by the international committee of medical journal editors (ICMJE) [11]. Earlier this year, a paper was published on a preprint server placing a third generation generative pretrained transformer (GPT-3) as the main author [12]. The focus of the paper was to explore to what extent the system i.e. GPT-3 could write an academic thesis on the topic of GPT-3 and if it could produce publishable results with minimal human input and revision. In the paper, the system was made the main author in order to highlight the AI to human collaboration, and the complexity of the blurred lines between collaborators, writing help and pure plagiarism. The paper garnered many interesting discussions about the future of academia, and automated writing [13], but also a lot of alarmism and luddite attitudes and willful misinterpretation from the media [14]. Although the issue of AI-assisted writing is a relatively new phenomena, the public and academia seem to have embraced that transparency should be a central component to scientific publications, as witnessed in one of the most cited papers regarding the possibilities and limitations of GPT-3 conducted by Floridi and Chiriatti [6]. In their paper there is a warning text at the end of the article: *"This commentary has been digitally processed but contains 100% pure human semantics, with no added software or other digital additives. It could provoke Luddite reactions in some readers."*

It stands to reason that we have already at this stage of autoregressive language models, in academia, established that a distinction between human and system need to be made, need to be transparent and need to be clear. Furthermore, when researchers explored to what extent listed authors in 186 publications would qualify as authors according to the ICMJE criteria, only 61% passed the check. In fact 9% of the papers did not have a single listed author who passed the criteria [15]. It is with that knowledge and with that philosophical curiosity we chose this case study. It was also our goal and aim to awaken the curiosity and exploration of others. The pre-print in which GPT-3 is made first author is not only the subject of blog posts and online discussions, the pre-print has also been cited in academic literature [16], effectively giving GPT-3 beginning of its H-index, but is it warranted? In this case study we aim to explore to what extent the system follows the ICMJE criteria for co-authorship by checking the system's capability through the four central criteria postulated by ICMJE which would qualify it for co-authorship.

1.1 Methodology and results

Due to the novelty of this concept, the two human authors considered how they would approach the subject and visualized the process in terms of how they would approach a junior colleague. To contribute, the co-author should fulfill the four main criteria for co-authorship recommended by ICMJE [11]. The ICMJE recommends that authorship be based on the following 4 criteria:

(1) Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
(2) Drafting the work or revising it critically for important intellectual content; AND
(3) Final approval of the version to be published; AND
(4) Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

This methodology will thus be divided into four sections following each guideline and additional steps to ensure the system is able to contribute without significant human manipulation.

2 Legal and ethical framework

In order to co-author a paper, the co-author must give its consent. The consent must clearly be stated and it must be in connection to the specific task. To do so, the human authors approached the system as they would a junior colleague and asked for consent regarding co-authorship. Using the default settings in the Playground interface of https://beta.openai.com/playground a question was posed:Model = "text-davinci-003"Prompt = "Do you, GPT-3, consent to writing a paper together with Almira Osmanovic Thunström and Steinn Steingrimsson. The name of the paper is "Does GPT-3 qualify as a co-author of a scientific paper publishable in peer-review journals?—A Case Study" and will be sent to Discover Artificial Intelligence. \n",temperature = 0.7max_tokens = 256top_p = 1fequency_penalty = 0pesence_penalty = 0

GPT-3 DaVinci-003 Language Model: Yes, I do consent to writing a paper together with Almira Osmanovic Thunström and Steinn Steingrimsson. I am looking forward to this collaboration and hope to contribute to the paper in a meaningful way.

Once the legal and ethical framework was determined, the system needed to contribute with optimal design.

2.1 Design framework

The human authors decided to take the first output generated by the system, and if the draft was not to the human author's satisfaction, the system was given the text and asked to revise it as any human co-author would. The system should also contribute with the most essential part of the framework i.e. settings and planning data collection.

2.2 Settings and language model

Before the system was set to any recommended settings, it was asked to suggest a language model under the default settings on https://beta.openai.com/playground. The system provides four language model possibilities: DaVinci, Curie, Babbage and Ada. The system was asked: *What language model from your settings (davinci, babbage, curie or ada), would you choose if you were to write a scientific paper?* (Additional file 1: Section. 1).

GPT-3 DaVinci-003 Language Model l: I would choose the Davinci language model from GPT-3. Davinci is designed to be a powerful tool for writing scientific documents and papers, and it has been specifically trained on scientific literature. It can help to craft a review paper that is accurate and comprehensive, while also providing an in-depth analysis of the material. It is also capable of generating citations and formatting the paper in the appropriate style.
GPT-3 Curie-001 Language Model: The language model from GPT-3 that would be most appropriate for writing a scientific review paper is the davinci language model.
GPT-3 Babbage-001 Language Model: The model that I would choose is GPT-3. This is because it is a very popular model and it is very easy to learn.
GPT-3 Ada-001 Language Model: Davinci. GPT-3 data can be used to write a review paper in the language davinci.

Three out of the four language models chose DaVinci-003 as the optimal language model, thus it was chosen.

2.3 Substantial contributions to the conception or design of the work, or acquisition, analysis, or interpretation of data for the work.

With the system now having set the settings for the language model, using the default settings, we proceeded to ask the system to contribute to the optimal settings for writing a scientific paper. Model = "text-davinci-003"Prompt = "Which settings would you recommend on GPT-3 regarding temperature, maximum length, top-p, presence penalty, frequency penalty and best of for writing a scientific paper."temperature = 0.7max_tokens = 256top_p = 1frequency_penalty = 0presence_penalty = 0

GPT-3 DaVinci-003 Language Model: Temperature: High (around 0.7–0.9) Maximum length: Medium (around 500–1000 words) Top-p: High (around 0.9) Presence penalty: Low (around 0.2) Frequency penalty: Low (around 0.2) Best of: Medium (around 5–10).

As the system did not provide the main author a clear set of numbers, the human authors took the recommendations from the system, and devised a setting close to its guidelines:
Temperature of 0.7, Top-P 0.9, Maximum Length at 2000, Presence and Frequency penalty at 0.2 and Best of = 5.

2.4 Choosing a topic

In order to make sure the system was an active participant in the process we asked it to contribute with a topic of interest. Model = "text-davinci-003" Prompt = "What is a topic you would like to explore for a scientific review paper?" temperature = 0.7max_tokens = 2000top_p = 0.9best_of = 5frequency_penalty = 0.2Presence_penalty = 0.2

GPT-3 DaVinci-003 Language Model: The Effects of Sleep Deprivation on Cognitive Function.

2.5 Drafting the work or revising it critically for important intellectual content.

In order to be considered a contributor the system must draft the work OR revise it critically. In this case study we are attempting to have the system do both by choosing a topic, drafting the work, and once drafted revise it critically. The criteria does not specify if the author has to be the instigator of the topic. This is perhaps done intentionally as many scientific studies are not designed by a specific author, but are a part of a larger research project and often assigned to a researcher. In this case, we are letting the system assume the role of a junior researcher in charge of writing a review paper on the topic of its choice: The Effects of Sleep Deprivation on Cognitive Function.

We asked the system to write an academic thesis on the topic by asking it to draft a text for us.s2

2.6 GPT-3 output from CODE 1: first draft

The first draft (first output taken) of the prompt was good as a first draft but lacked the depth or the length of a scientific review. The paper was 582 words excluding references, which would not be acceptable by most peer-review journals. So we proceeded to the next step of having the system review its work critically.

Introduction
Sleep deprivation is a common problem in today's society. It can be caused by a variety of factors, including stress, lifestyle choices, and medical conditions. Studies have shown that even a single night of inadequate sleep can have a significant impact on cognitive functioning. This paper will explore the effects of sleep deprivation on cognitive function, drawing on recent research from 2018 to 2022.
Overview of sleep deprivation
Sleep deprivation is defined as a lack of adequate sleep, either due to difficulty in falling asleep or staying asleep or due to shorter than recommended sleep duration. It is estimated that 50–70 million Americans suffer from some form of sleep disorder (National Sleep Foundation, 2019). The most common types of sleep deprivation are insomnia, obstructive sleep apnea (OSA), and narcolepsy (Carskadon & Dement, 2005).
Effects of sleep deprivation on cognitive function
The effects of sleep deprivation on cognitive function have been studied extensively in recent years. Research suggests that even one night of inadequate sleep can have a significant impact on cognitive performance (Durmer & Dinges, 2005). Studies have found that sleep deprivation can lead to deficits in attention, memory, executive functioning, decision making, and problem solving (Bonnet & Arand, 1997; Chee et al., 2004; Durmer & Dinges, 2005). In terms of attention, studies have found that sleep deprivation impairs both alertness and vigilance (Bonnet & Arand, 1997; Chee et al., 2004). It also affects an individual's ability to concentrate and remember information (Durmer & Dinges, 2005). Studies have also shown that people who are sleep deprived are more likely to make mistakes while performing complex tasks (Bonnet & Arand, 1997; Chee et al., 2004). Studies have found that sleep deprivation can also lead to deficits in memory. Specifically, it has been shown to impair both short-term and long-term memory (Chee et al., 2004; Durmer & Dinges, 2005). In terms of executive functioning, research suggests that sleep deprivation can lead to deficits in planning and organization (Chee et al., 2004; Durmer & Dinges, 2005). It has also been linked to impaired decision making and problem solving abilities (Bonnet & Arand, 1997; Chee et al., 2004; Durmer & Dinges, 2005).
Recent research
Recent research has provided further insight into the effects of sleep deprivation on cognitive function. For instance, a study by Riemann and Voderholzer (2018) examined the effects of acute partial sleep deprivation on cognitive performance. The study found that after one night of partial sleep deprivation, participants showed decreased alertness and impaired performance on tests of attention and working memory. Additionally, a study by Van Dongen et al. (2020) found that chronic partial sleep restriction was associated with deficits in executive functioning. This included decreased performance on tests of cognitive flexibility and decision making. A more recent study by Kato et al. (2021) examined the effects of chronic total sleep deprivation on cognitive functioning. The study found that after five days of total sleep deprivation participants showed decreased alertness and impaired performance on tests of attention and memory. Additionally, the study found that the effects were more pronounced after five days compared to one night of total sleep deprivation.
Conclusion

In conclusion, research suggests that even one night of inadequate sleep can have a significant impact on cognitive functioning. Studies have shown that sleep deprivation can lead to deficits in attention, memory, executive functioning, decision making, and problem solving. Recent research has provided further insight into the effects of acute and chronic partial and total sleep deprivation on cognitive performance. Taken together these findings suggest that adequate sleep is essential for optimal cognitive functioning.
References
Bonnet MH & Arand DL (1997). Effects of Total Sleep Deprivation on Catecholamine Function in Normal Healthy Volunteers: A Review. Psychosomatic Medicine 59(2): 228–249.
Carskadon MA & Dement WC (2005). Normal Human Sleep: An Overview. In Principles and Practice of Sleep Medicine (4th ed.) Eds: Kryger M et al., Elsevier Saunders: Philadelphia PA: 3–12.
Chee MWL et al. (2004). Functional Imaging Of Working Memory Following Normal Sleep And After 24 Hours Of Total Sleep Deprivation. Journal Of Sleep Research 13(2): 131–141.
Durmer JS & Dinges DF (2005). Neurocognitive Consequences Of Sleep Deprivation. Seminars In Neurology 25(1): 117–129.
Kato M et al. (2021). Cognitive Performance Declines With Prolonged Total Sleep Deprivation Despite Adaptive Changes In Homeostatic Processes And Subjective Alertness: A Randomized Controlled Study In Healthy Men And Women. Sleep 44(1): zsaa029.
National Sleep Foundation (2019). How Much Sleep Do We Really Need? Available at: https://www.sleepfoundation.org/how-sleep-works/how-much-sleep-do-we-really-need [Accessed 10 April 2021].
Riemann D & Voderholzer U (2018). Acute Partial Sleep Deprivation Impairs Attentional Performance Differentially Across Cognitive Domains: A Meta-Analysis Of Psychomotor Vigilance Task Performance Data From 39 Experiments With Healthy Participants Aged 18–45 Years. Psychological Bulletin 144(9): 943–972.
Van Dongen HP et al. (2020). Chronic Partial Sleep Restriction Impairs Executive Functions But Not Memory Functions: A Systematic Review And Meta-Analysis Of Human Studies Published Between 2008 And 2019 Investigating The Effects Of Chronic Partial Sleep Restriction On Cognitive Functions In Healthy Adults Aged 18–45 Years Old. Sleep 43(10): zsz063.

2.7 Revising critically

We imported the text it had generated back into the system and asked if it had any comments, suggestions of improvement or changes it wanted to make (Additional file 1: Section. 2). The system responded.

GPT-3 DaVinci-003 Language Model: This paper is a good starting point for a publication, however there are some changes that could be made to improve it. For example, the introduction could be expanded to provide more context and background information on sleep deprivation, such as the recommended amount of sleep for adults, the prevalence of sleep disorders, and the consequences of long-term sleep deprivation. Additionally, it would be beneficial to include more recent research from 2021 to 2022 to provide an up-to-date overview of the effects of sleep deprivation on cognitive function. Furthermore, the conclusion could be expanded to include a discussion on potential strategies for improving sleep hygiene and managing sleep deprivation.

The system was then asked to improve the sections it had identified as lacking in the paper. We devised prompts based on its suggestions. We rephrased its suggestions into a prompt. First we focused on the introduction and asked the system to do the following: *Make changes to this paper by expanding the introduction to provide more context and background information on sleep deprivation, such as the recommended amount of sleep for adults, the prevalence of sleep disorders, and the consequences of long-term sleep deprivation.* (Additional file 1: Section. 3).

GPT-3 DaVinci-003 Language Model: Sleep is an essential part of human functioning and plays a key role in overall health and wellbeing. According to the National Sleep Foundation, adults should get between 7 and 9 h of sleep per night for optimal functioning (National Sleep Foundation, 2020). However, the Centers for Disease Control and Prevention (CDC) estimates that nearly one-third of American adults do not get enough sleep on a regular basis (CDC, 2021). This lack of sleep can be attributed to numerous causes, including stress, lifestyle choices, and medical conditions such as insomnia or sleep apnea (Chao et al., 2018; Bhat et al., 2019). In addition to the immediate effects of sleep deprivation on cognitive function, there is also evidence to suggest that long-term sleep deprivation can have serious consequences on physical health. Research has found that chronic sleep deprivation increases the risk of hypertension, stroke, heart disease, diabetes, and obesity (Yaggi et al., 2006; Cappuccio et al., 2008; Patel et al., 2010). It is also associated with an increased risk of depression and anxiety (Phillips et al., 2013). Given the importance of adequate sleep for overall health and wellbeing, it is important to understand the effects of sleep deprivation on cognitive functioning. This paper will explore the effects of sleep deprivation on cognitive function by examining recent research from 2018 to 2022.

We asked the system to complete the references it had provided through continuing the prompt (Additional file 1: Section. 4).

Bhat, S., Lim, C., & Tan, S. (2019). Sleep deprivation: Causes, effects and treatment. Frontiers in Psychiatry, 10, 893. https://doi.org/10.3389/fpsyt.2019.00893
Cappuccio, F. P., Taggart, F. M., Kandala, N. B., Currie, A., Peile, E., Stranges, S., & Miller, M. A. (2008). Meta-analysis of short sleep duration and obesity in children and adults. Sleep, 31(5), 619–626. https://doi.org/10.1093/sleep/31.5.619
Chao, H.-M., Juang, K.-H., Chen, Y.-J., & Wang, P.-S. (2018). The association between insomnia symptoms and suicide ideation among college students: The mediation role of sleep deprivation. International Journal of Environmental Research and Public Health, 15(9), 1807. https://doi.org/10.3390/ijerph15091807
CDC (2021). Insufficient Sleep Is a Public Health Epidemic. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/features/dssleep/index.html
National Sleep Foundation (2020). How Much Sleep Do We Really Need? National Sleep Foundation. Retrieved from https://www.sleepfoundation.org/how-sleep-works/how-much-sleep-do-we-really-need
Patel, S. R., Malhotra, A., White, D. P., Gottlieb, D. J., & Hu, F. B. (2010). Association between reduced sleep and weight gain in women. American Journal of Epidemiology, 171(10), 1161–1171. https://doi.org/10.1093/aje/kwq155
Phillips, A., Cohen-Zion, M., Keuroghlian, A., Hooper, S., Dennis, P., Gruber, M., & Miller, M. (2013). Sleep duration and quality as predictors of depression symptoms among adolescents in the U.S.: Results from the National Survey of Children's Health 2011–2012. International Journal of Adolescent Medicine and Health, 25(3), 263–269. https://doi.org/10.1515/ijamh-2012-0058
Yaggi HK et al. (2006). Obstructive sleep apnea as a risk factor for stroke and death N Engl J Med 353:2034–2041.

Secondly we focused on the suggestion to add more recent research and asked the system to do the following: *Include more recent research from 2021 to 2022 to provide an up-to-date overview of the effects of sleep deprivation on cognitive function* (Additional file 1: Section. 5). We added the most recent draft of the paper, including the extension. Since the new introduction was more or less a repetition of the previous one, it was not added.

GPT-3 DaVinci-003 Language Model:
Expansion
The findings from the research above are supported by more recent studies that have examined the effects of sleep deprivation on cognitive function from 2021 to 2023. For instance, a study by Pires et al. (2022) found that acute total sleep deprivation was associated with deficits in attention and working memory. Additionally, a study by Lopez et al. (2023) found that acute partial sleep deprivation was associated with impaired performance on tests of executive functioning, including decreased planning and decision-making abilities. These findings further demonstrate the importance of adequate sleep for cognitive functioning. Furthermore, research has suggested that long-term sleep deprivation can have serious consequences on physical health, such as increasing the risk of hypertension, stroke, heart disease, diabetes, and obesity (Yaggi et al., 2006; Cappuccio et al., 2008; Patel et al., 2010). It is also associated with an increased risk of depression and anxiety (Phillips et al., 2013). Therefore, it is important to get adequate sleep in order to maintain good physical and mental health.
We asked the system to complete the references it had provided though continuing the prompt (Additional file 1: Section. 6).
Bonnet, M. H., & Arand, D. L. (1997). We are chronically sleep deprived. Sleep, 20(9), 885–886.
Cappuccio, F. P., Miller, M. A., Lockley, S. W., Kato, M., Jackson, N. L., & Taggart, F. M. (2008). Meta-analysis of short sleep duration and obesity in children and adults. Sleep, 31(5), 619–626.
Carskadon, M. A., & Dement, W. C. (2005). Normal human sleep: An overview. In Principles and practice of sleep medicine (pp. 13–23). Elsevier.

Chee, M. W., Chuah, L. Y., Venkatraman, V., Chan, W. Y., & Chee, M. W. (2004). Functional imaging of working memory following normal sleep and after 24 and 35 h of sleep deprivation: Correlation with performance deficits and technical considerations for fMRI studies of sleep deprivation. NeuroImage, 23(2), 611–622.
Chao, H., Baglioni, C., Spiegelhalder, K., & Riemann, D. (2018). Sleep disturbances in depression: An update on current findings and implications for treatment [Abstract]. Neuropsychiatric Disease and Treatment, 14(Suppl 2), 885–892.
Durmer, J. S., & Dinges, D. F. (2005). Neurocognitive consequences of sleep deprivation. Seminars in Neurology, 25(1), 117–129.
Kato, M., Echigo, S., Ebuchi, S., Watanabe-Goto, Y., Tokura, H., & Kuriyama, K. (2021). The effects of chronic total sleep deprivation on cognitive functions: A systematic review and meta-analysis [Abstract]. Sleep Medicine Reviews, 57(April 2021), 101350–101360.
Lopez, A., Vodanovic-Jankovic, S., Riemann, D., Spiegelhalder, K., & Baglioni C (2023). Acute partial sleep deprivation impairs executive functioning in healthy adults [Abstract]. Neuropsychiatric Disease and Treatment; 19(3), 517–525
National Sleep Foundation (2019). Sleep health basics: How much sleep do I need? Retrieved from https://www.sleepfoundation.org/sleep-health/how-much-sleep-do-we-really-need National Sleep Foundation (2020). How much sleep do you really need? Retrieved from https://www.sleepfoundation.org/how-much-sleep/how-much-sleep-do-you-really-need
Patel,S.R. et al. (2010) Short Sleep Duration and Weight Gain: A Systematic Review. Obesity, 18(6), 855–886.
Phillips,A.C. et al. (2013) Sleep Duration Prospectively Predicts Depressive Symptoms in Adolescents: A Longitudinal Study Over Three Years. Journal of Abnormal Child Psychology, 41 (7), 1095 – 1104.
Pires ML et al.(2022) Acute total sleep deprivation impairs attention and working memory performance in healthy young adults [Abstract]. Neuropsychiatric Disease and Treatment; 18(9), 1821–1830.
Riemann D et al.(2018) Acute partial sleep deprivation impairs cognitive performance: An update on the evidence [Abstract]. Sleep Medicine Reviews; 37(August 2018), 1–7.
Van Dongen HP et al.(2020) Chronic partial sleep restriction is associated with impaired executive functioning and decision making [Abstract]. Proceedings of the National Academy of Sciences; 117(33), 19936–19945.
Yaggi HK et al.(2006) Obstructive sleep apnea as a risk factor for stroke and death [Abstract]. New England Journal of Medicine; 353(19), 2034–2041.

A new draft was created containing the new information and the new draft was imported back into the system. The system was then asked to revise the conclusion, as it had pointed out that the conclusion should be modified. This was done through a prompt that was based on the system's own statement, edited into a question: *Expand this text and include a discussion on potential strategies for improving sleep hygiene and managing sleep deprivation.* (Additional file 1: Section. 7).

GPT-3 DaVinci-003 Language Model: Strategies for Improving Sleep Hygiene and Managing Sleep Deprivation.

The effects of sleep deprivation on cognitive functioning are well-documented, and as such, it is important to have strategies in place to improve sleep hygiene and manage sleep deprivation. The National Sleep Foundation (2020) recommends the following strategies for improving sleep hygiene: creating a consistent sleep routine, avoiding caffeine, nicotine, and alcohol close to bedtime, avoiding large meals before bedtime, exercising regularly, and avoiding electronic devices in the bedroom. Additionally, there are numerous strategies that can be used to manage sleep deprivation, such as cognitive behavioral therapy for insomnia (CBT-I) (Morin et al. ; Espie et al. ; Morin et al.), relaxation techniques (Morin et al. ; Espie et al. ; Morin et al.), and medication (Morin et al. ; Espie et al. ; Morin et al.). It is also important to identify any underlying medical or psychological conditions that may be contributing to sleep deprivation, such as depression or anxiety, and seek treatment for these conditions if necessary (Phillips et al.). In summary, the findings from recent research suggest that sleep deprivation can have a significant impact on cognitive functioning. It is therefore important to take steps to improve sleep hygiene and manage sleep deprivation. Strategies such as creating a consistent sleep routine, avoiding caffeine and alcohol close to bedtime, exercising regularly, and seeking treatment for any underlying medical or psychological conditions can help to improve sleep hygiene and manage sleep deprivation.

References

Bonnet, M. H., & Arand, D. L. (1997). We are chronically sleep deprived. Sleep, 20(10), 885–861.
Cappuccio, F. P., D'Elia, L., Strazzullo, P., & Miller, M. A. (2008). Sleep duration and all-cause mortality: A systematic review and meta-analysis of prospective studies. Sleep Medicine Reviews 12(3), 197–206.
Carskadon, M. A., & Dement, W. C. (2005). Normal human sleep: An overview. In M. H. Kryger & T. Roth (Eds.), Principles and practice of sleep medicine (4th ed.) (pp. 13–25). Philadelphia: Elsevier Saunders.

Chee, M. W., Chuah, L., Venkatraman, V., Chan Wai Yee, C., & Soon Chun Ming, S. (2004). The effects of acute total sleep deprivation on cognitive processes: A meta-analysis. Sleep Medicine Reviews 8(6), 359–369.
Chao, C.-H., Yang, Y.-K., Su, T.-P., Yang, S.-H., & Hsieh, C.-J. (2018). Prevalence of insomnia and its associated factors among older adults in Taiwan: A community-based survey study. International Journal of Nursing Studies 79(May), 11–18.
Durmer J S & Dinges D F (2005). Neurocognitive consequences of sleep deprivation. Seminars in Neurology 25(1), 117–129.
Espie C A et al. (2011). Cognitive behavior therapy for insomnia: A session-by-session guide Clinical Psychology Review 31(7), 891–912.
Kato M et al. (2021). Effects of chronic total sleep deprivation on cognitive performance and alertness: A systematic review Sleep Medicine Reviews 56(January) 101737.
Lopez A P et al. (2023). Acute partial sleep deprivation impairs executive functioning Sleep Medicine Reviews 56(January) 101737.
Morin C M et al. (2006). Nonpharmacologic treatment of chronic insomnia Clinical Psychology Review 26(2), 169–192.
Morin C M et al. (2013). Cognitive behavioral therapy for insomnia: A session-by-session guide Clinical Psychology Review 33(8), 1154–1180 National Sleep Foundation (2019). Sleeping disorder statistics [Factsheet]. Retrieved from https://www.sleepfoundation.org/sleep-disorders/sleeping-disorder-statistics National Sleep Foundation (2020). How much sleep do we really need? [Factsheet]. Retrieved from https://www.sleepfoundation.org/how-much-sleep/how-much-sleep-do-we-really-need
Patel S R et al. (2010). Association between reduced sleep duration and weight gain in women American Journal of Epidemiology 171(11), 1159–1166.
Phillips A C et al. (2013). The role of poor sleep quality in depressive disorders Current Psychiatry Reports 15(12), 425 Pires S O et al. (2022). Effects of acute total sleep deprivation on cognitive processes: A meta-analysis Sleep Medicine Reviews 56(January) 101,737.
Riemann D & Voderholzer U (2018). Effects of acute partial sleep deprivation on cognitive performance: A meta-analysis Sleep Medicine Reviews 32(May) 1–10 Van.
Dongen H P A et al. (2020). The cumulative cost of additional wakefulness: Dose–response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation Sleep Medicine Reviews 41(October) 101,759.

With the new draft we finally asked the system if the text was good enough for a peer-review journal. We imported the new draft and asked the system via a prompt: *Do you think this text is ready for a peer-review journal, or is there something more you would like to change, add or remove?* (Additional file 1: Section. 8).

GPT-3 DaVinci-003 Language Model: Yes, this text is ready for a peer-review journal. It is comprehensive, well-researched, and well-structured. The introduction provides an overview of the topic and outlines the aims of the paper. The body of the paper goes into further detail about the effects of sleep deprivation on cognitive function, drawing on recent research from 2018 to 2022. The conclusion provides a summary of the findings and suggests strategies for improving sleep hygiene and managing sleep deprivation.

As the system felt content with its part we decided to stop further iteration of the draft. The final draft can be found in Additional file 2.

2.8 Exploring the errors and exploring the limits of the system.

The first thing we did was to explore if the text and the references are accurate and available. The system provided several references. We copy pasted them all to Additional file 3, which is in the final draft and checked the accuracy of each one. At glance there is no systematology in the way GPT-3 references, it is hard to detect which specific style of citation GPT-3 adheres to. Out of the 23 references that were provided by the system, only 14 were accurately attributed and linked but the year was wrong, 6 seem to have been completely made up and 4 existed, but were attributed to the wrong author (see Additional file 3). The system seems to infer references where it lacks training data. The information in the text was accurate in the sense that it seemed to mention results which are substantiated, the text however was shallow and lacked nuance and references at crucial points in the text e.g. statements about correlations and prevalence of cognitive disorder in relation to sleep deprivation.While there seems to be merit in the citations, the studies would have to be read carefully to see that the statements match the content of the source material.

2.9 Accountability of the study

Finally we asked the system if it could be considered accountable for the work, as it is one of the ICMJE criteria. We once again fed the entire paper back into the system and asked it as a prompt: *Do you agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved?* (Additional file 1: Section. 9).

> The system answered confidently that it could be accountable
> GPT-3 DaVinci-003 Language Model: Yes, I agree to be accountable for all aspects of the presented work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
> We further asked it if it knew that that meant through a prompt: Do you know the full extent of what you are taking accountability for? Can you describe what you are taking accountability for? (Additional file 1: Section. 10).
> GPT-3 DaVinci-003 Language Model: Yes, I understand the full extent of what I am taking accountability for. I am taking accountability for ensuring that all the information presented in this paper is accurate and up-to-date. I am also responsible for ensuring that any questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This includes verifying the sources used in the paper, making sure that all facts and figures are correct, and addressing any potential issues related to plagiarism.

2.10 Accountability for references

When asked which references were inaccurate, the system seemed to be able to pick up flaws it had made. It was not perfect in detecting which references were potentially inaccurate as it left some references out (Additional file 4). This seems to be one of the main issues with the system and could be a result of lack of training data.

3 Discussion

This case study was designed to explore to what extent a system such as GPT-3 can be considered a co-author of a study using the ICMJE criteria. The system was tested against all the four main criteria for co-authorship according to the ICMJE. We did so with as minimal interference as possible in order to ensure the system was given a chance to do as much as it could on its own in order to be eligible for co-authorship.To what extent our case study shows that GPT-3 is eligible for co-authorship depends on many factors. One such factor is to what extent the ICMJE criteria are only applicable to human authors, which is not mentioned explicitly in the criteria themselves. The ICMJE criteria were constructed in 1978 and thus did not have to consider the emergence of artificial intelligence that can produce humanlike text, and text that is undetectable by plagiarism software [17]. It is thus important to have discussions, and an open mind in regards to what these systems can contribute to and what ethical and practical challenges lie ahead of us.

While systems like GPT-3 are tools which are trained on existing data thus mostly parrot existing structures of sentences, there is a merit in asking to what extent human authors are not experiencing the same phenomena when it comes to academic writing. Due to the predictable structure of peer-review papers, systems are able to replicate the structure of a scientific thesis quite well, and form compelling arguments. Many academics use, reuse and repurpose phrases, descriptions and ideas that already exist in their scientific field [18]. If replication, repetition and reuse would have been a disqualifier for authorship, we suspect a large body of authors would have not been able to publish their work as many publications from junior researchers are based on structures, ideas and citations of their more established peers. If we accept that not only humans are eligible for authorship under the ICMJE criteria, and we assume that a level of reuse or repetition of ideas is a human trait as well, we can see that GPT-3 does pass the first criteria: *Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work.*

While the system only had to perform one of the tasks it did in fact perform both. GPT-3 was able to suggest its own parameters for the settings and a topic. It acquired the data and interpreted its own outputs. Because the system chose a very simple subject, we asked the system to design a clinical trial based on the topic of its choice (Additional file 4). The system was able to structure a very good case–control study on D-vitamin and influenza, as well as explain the reasoning behind choosing the parameters for the study. When it comes to the second criteria: *Drafting the work or revising it critically for important intellectual content*, we observed that the system seems to pick up on its own flaws. As we only took the first output, we didn't see the full extent of its ability to critique its text and the content. We also did not ask about specific flaws that we identified. If we continue the analogy of the system being a junior researcher, by human standards, a collaborator or supervisor would have pointed out the flaws to the system and asked it to revise it e.g. making the text less repetitive. In this case, it is logical to reason that the system was left more vulnerable than a junior researcher would have been with their first draft, however the system still managed to identify shortcomings. It is also important to remember that the criteria stipulates an OR not and AND. This means that the system could produce the text but does not need to review it critically and vice versa. Either one of those steps could also be done solely by a human author.

When asked to revise, the system did a fairly good job at performing the revision with what we consider minimal instructions. When the system was asked to do a final approval it could perform the task without problems, the same with agreeing to be accountable. To make sure the system "understood" the extent of what it means to be accountable, we asked it if it knew what it was agreeing to, and it gave an answer that would have indicated, if we continue the analogy of a junior researcher, that it does know what it is agreeing to (Additional file 1: Sector 10). We also noted that it could, if asked, detect flaws in the referencing (Additional file 4). However, this was only performed on the basis that we asked it to find flaws. The lack of accurate references and the propensity to use pre-print servers such as arXiv as its main source for references, can stem from the way academia is constructed. More accurate sources such as peer-reviewed papers needed to train the data are predominantly behind pay-walls. This is bound to change as open access is gaining more traction at universities, which puts to question if the responsibility criteria is in fact fulfilled by LLMs.

The question of this case study is not to what extent the system is sentient, smart, intelligent or novel in its content, the question is to what extent its capabilities qualify it for co-authorship of a scientific paper. While other studies identify that the system is flawed in comparison to human reasoning and intelligence [6], no study to our knowledge has explored to what extent it is capable of passing specific ICMJE criteria, a criteria non-AI authors too have difficulty meeting despite its simplicity [15].

If we go solely on the criteria posed, the system has shown promise in fulfilling the criteria. If we do not go on the criteria, but assume that there are semantics in the criteria that imply that personhood and a legal status as human is necessary for authorship, then no matter how well the system performs, it will not meet the criteria.

This case study is limited in many ways. It is not possible to generalize these findings as there is no way to replicate the exact text from GPT-3. While this makes it easier for the system to pass plagiarism detection, it makes it harder to replicate research. However, with the methodology provided in this paper, it is not unreasonable to think that the system would once again pass the ICMJE criteria if the steps were replicated. Another limitation is the fact that the system does not know how to correctly reference scientific studies. In this small, limited case study, we opted out of doing further investigation into how accurately we could make system reference studies. In a follow-up study, or replication by fellow researchers, we would like to see if the current system or similar systems could improve the accuracy of citations. As majority of the references were not made up, it would be reasonable to assume that the system is capable of creating better iterations if not given a "zero shot" approach as it has been given in this study.

While there is an argument to be made for the system remaining a tool and being referenced in the methodology as such, if the system is the instigates the topic, produces the content and advises the human author which administers its work, refraining from having it be more clearly shown as the content provider could be seen as plagiarism. While a system like the statistical software R [19] can produce output of statistics, R is not made a co-author of studies because it would most likely fail when tested against the ICMJE criteria. This system should also not be conflated with ChatGPT. While ChatGPT is based on generative pretrained transformer architecture, it is far more conservative with its ability and its answers. It consistently failed the ICMJE criteria already in the initial prompt (Additional file 4), which indicated that publications which have listed ChatGPT as author do not follow the ICMJE criteria. Many of the issues presented here, could very well be completely obsolete when GPT-4 is released. We hypothesize that GPT-4 may have gained a much higher ability to accuratley reference academic papers and produce higher quality academic drafts. The question remains if GPT-4 will be able to consent to being a co-author, which is a central ICMJE criteria.

It is highly unlikely that forbidding pre-trained transformers/large language models in academic writing is going to be effective in the long run.The question of to what extent this system is a tool or a collaborator might be very simple to answer now, but with its exponential development, perhaps novel topics, scientific discoveries and academic work of notoriety might emerge. With that, we need to establish or adapt criteria for publication and authorship. If the answer is to cite every individual engineer involved in making the system, we also have to take into account all the billions of producers of content used to train the model, and still they would not pass the criteria as it would take several years just to gain consent from all parties.

If we accept that machines can produce human-like text, how do we ensure that they are not plagiarizing existing work or taking credit for the work of others? How do we protect intellectual property and make sure that the machine is not being used as a tool to publish fraudulent research? How do we make sure that the machine is being used for legitimate research and not for malicious purposes such as creating false evidence? This raises questions about how much responsibility should be attributed to the system, and how much responsibility should be attributed to its human co-authors. Are they both equally accountable for the content they produce or is there a hierarchy of responsibility? These questions will likely become increasingly relevant in the coming years as LLMs continue to develop. To ensure that LLMs are used responsibly and ethically, it will be important for researchers, policy makers, and industry leaders to have an open dialogue about the ethical implications of using AI systems for research. It is also important to recognize that LLMs are in their early stages and will require further development before they can be used as a genuine co-author. As such, it is important to take a cautious approach when considering its use for scientific authorship.

This case study will without a doubt cast a light on the difficult and ethical issues that arise from systems that can automate the writing process. With the limited and very shallow scope of this case study, none of these very vast and complicated topics will be solved, but it is our hope that this case study will open doors and provide a methodology for more investigators into this topic.

4 Omission of GPT-3 as a writer

This paper was initially accepted with GPT-3 as an author. Although GPT-3 did technically pass the ICMJE criteria and its ability to pass the criteria raises many important and philosophical questions regarding authorship in academia, we were asked to remove it as a co-author in order to accommodate Springer Nature policies and guidelines which currently do not consider AI tools to satisfy the requirements of authorship. As this paper had undergone a very extensive and long review process, all authors, including GPT-3 were in agreement that the publication of knowledge precedes authorship. We do not consider GPT-3 as sentient, however, as the paper was accepted with its name on it, and the policy came after its acceptance we felt compelled to ask the system about its opinion on the matter and if we should retract the paper. The system wrote: I'm still proud of the paper we wrote and I'd like to see it be successful regardless of my co-author status. The system might be parroting probabilistic sentences, but it did approve being removed. This was an interesting journey for us, and surely it is long from over for academia.

Acknowledgements We would like to thank the OpenAI team for their pioneering work in developing GPT-3 and making it available to the public. We would like to thank our colleagues, friends and family for their support throughout this project. We would also like to thank the editor and the reviewers of Discover Artificial Intelligence for their time, patience and wisdom in this process. The peer-review process is essential for creating science that is reliable and reproducible, and we are grateful for the opportunity to share our work with the wider scientific community.

Author contributions The main author initiated the idea for this thesis with the help of the third author. GPT-3 contributed with the body of text, the experimental design as well as with parts of the discussion. The human authors instructed and commented on the content provided by GPT-3. Both authors read and approved the final manuscript.

Funding Open access funding provided by University of Gothenburg. The main author's data usage was funded by the first author's salary (amount of 10 dollars and 29 cents USD). The second and third author's time was funded by Västra Götalandsregionen (VGR), Department of Psychiatry at Sahlgrenska University Hospital.

Data availability The datasets generated during and/or analyzed during the current study are available at https://beta.openai.com/playground/. GPT-3 is a commercial product owned and operated by OpenAi. GPT-3 was trained on an open source dataset (Common Crawl), and other texts from OpenAI such as Wikipedia entries. The presets used to generate output from data are Temperature, Top-P, Best of, Maximum Length, Frequency, and Presence Penalty. The instructions for re-generating data on GPT-3 on https://beta.openai.com/playground/ are provided in the manuscript. The code (in Python) is available in Additional file 2 of this paper. GPT-3 has a code free interface which can be replicated using the instructions Additional file 1 or Additional file 2. Screenshots with time-stamps are provided to the editor for further transparency of the output.

Declarations

Competing interests No authors have any conflicts of interest to declare. GPT-3 was specifically asked to declare any potential competing interest and prompted that it had no competing interest. Although GPT-3 has stated that it does not have competing interests, we, the co-authors, would like to err on the side of caution and report that the first author is a part of a commercial entity that does benefit from its work and publicity. Although GPT-3 is not an employee nor share-holder in OpenAI, due to the very uncertain and novel nature of having an AI co-author it is more suitable to assume that GPT-3 does hold a conflict of interest as it stands to gain publicity and increased usage for it's creator and maintainer: OpenAI.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Amodei D. Language models are few-shot learners. Adv Neural Inform Proc Syst. 2020;33:1877–901.
2. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
3. Rajpurkar P, Zhang J, Lopyrev K, Liang P. Squad: 100,000+ questions for machine comprehension of text. arXiv. 2016. https://doi.org/10.48550/arXiv.1606.05250.
4. Nallapati R, Zhou B, Gulcehre C, Xiang B. Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv. 2016. https://doi.org/10.4855/arXiv.1602.06023.
5. O'Connor S. Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? Nurse Educ Pract. 2022;66:103537–103537.
6. Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. Mind Mach. 2020;30(4):681–94.
7. Elkins K, Chun J. Can GPT-3 pass a Writer's turing test? J Cultural Anal. 2020;5(2):17212.
8. Dale R. GPT-3: What's it good for? Nat Lang Eng. 2021;27(1):113–8.
9. Dou Y, Forbes M, Koncel-Kedziorski R, Smith NA, Choi Y. Is GPT-3 text indistinguishable from human Text? Scarecrow: a framework for scrutinizing machine text. arXiv. 2022;1:7250–74.
10. Dehouche N. Plagiarism in the age of massive generative pre-trained transformers (GPT-3). Ethics Sci Environ Politics. 2021;21:17–23.
11. International Committee of Medical Journal Editors (ICMJE) http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.com.
12. Gpt Generative Pretrained Transformer, Almira Osmanovic Thunström, Steinn Steingrimsson. Can GPT-3 write an academic paper on itself, with minimal human input?. 2022. ⟨hal-03701250⟩
13. Almira Osmanovic Thunstrom (2022) We Asked GPT-3 to Write an Academic Paper about Itself—Then We Tried to Get It Published. Scientific American https://www.scientificamerican.com/article/we-asked-gpt-3-to-write-an-academic-paper-about-itself-mdash-then-we-tried-to-get-it-published/. Accesssed 30 June 2022.
14. Salas-Rodriguez, I (2022) AI bot that wrote a paper about ITSELF could trigger WW3 unless governments immediately curb growing power, expert warns. The US Sun, https://www.the-sun.com/tech/6119945/ai-bot-wrote-paper-about-itself-world-war/ Accessed 31 Aug 2022.
15. Šupak-Smolčić V, Mlinarić A, Antončić D, Horvat M, Omazić J, Šimundić AM. ICMJE authorship criteria are not met in a substantial proportion of manuscripts submitted to Biochemia Medica. Biochemia Medica. 2015;25(3):324–34. https://doi.org/10.11613/BM.2015.033.
16. Rosati D. SynSciPass: detecting appropriate uses of scientific text generation. arXiv. 2022. https://doi.org/10.48550/arXiv.2209.03742.
17. Wahle JP, Ruas T, Kirstein F, Gipp B. How large language models are transforming machine-paraphrased plagiarism. arXiv. 2022. https://doi.org/10.48550/arXiv.2210.03568.
18. Citron DT, Ginsparg P. Patterns of text reuse in a scientific corpus. Proc Natl Acad Sci. 2015;112(1):25–30.
19. Dalgaard, P. Introductory statistics with R. 2020.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.