# Age against the machine—susceptibility of large language models to cognitive impairment: cross sectional analysis

Roy Dayan,[1,2] Benjamin Uliel,[1,2] Gal Koplewitz[3,4]

[1]Department of Neurology, Hadassah Medical Center, Jerusalem, Israel
[2]Faculty of Medicine, Hebrew University, Jerusalem, Israel
[3]QuantumBlack Analytics, London, UK
[4]Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

**Correspondence to:** G Koplewitz
galkop@gmail.com
(ORCID 0000-0003-4906-0779)

Additional material is published online only. To view please visit the journal online.

**Cite this as:** *BMJ* 2024;387:e081948
http://dx.doi.org/10.1136/bmj-2024-081948

**Accepted:** 27 October 2024

## ABSTRACT

### OBJECTIVE
To evaluate the cognitive abilities of the leading large language models and identify their susceptibility to cognitive impairment, using the Montreal Cognitive Assessment (MoCA) and additional tests.

### DESIGN
Cross sectional analysis.

### SETTING
Online interaction with large language models via text based prompts.

### PARTICIPANTS
Publicly available large language models, or "chatbots": ChatGPT versions 4 and 4o (developed by OpenAI), Claude 3.5 "Sonnet" (developed by Anthropic), and Gemini versions 1 and 1.5 (developed by Alphabet).

### ASSESSMENTS
The MoCA test (version 8.1) was administered to the leading large language models with instructions identical to those given to human patients. Scoring followed official guidelines and was evaluated by a practising neurologist. Additional assessments included the Navon figure, cookie theft picture, Poppelreuter figure, and Stroop test.

### MAIN OUTCOME MEASURES
MoCA scores, performance in visuospatial/executive tasks, and Stroop test results.

### RESULTS
ChatGPT 4o achieved the highest score on the MoCA test (26/30), followed by ChatGPT 4 and Claude (25/30), with Gemini 1.0 scoring lowest (16/30). All large language models showed poor performance in visuospatial/executive tasks. Gemini models failed at the delayed recall task. Only ChatGPT 4o succeeded in the incongruent stage of the Stroop test.

### CONCLUSIONS
With the exception of ChatGPT 4o, almost all large language models subjected to the MoCA test showed signs of mild cognitive impairment. Moreover, as in humans, age is a key determinant of cognitive decline: "older" chatbots, like older patients, tend to perform worse on the MoCA test. These findings challenge the assumption that artificial intelligence will soon replace human doctors, as the cognitive impairment evident in leading chatbots may affect their reliability in medical diagnostics and undermine patients' confidence.

## WHAT IS ALREADY KNOWN ON THIS TOPIC

Colossal advancements in the field of artificial intelligence have led to a flurry of excited and fearful speculation as to whether chatbots surpass human physicians

Multiple studies have shown large language models (LLMs) to be remarkably adept at a range of medical diagnostic tasks, outscoring human physicians

If we are to rely on LLMs for medical diagnosis and care, we must examine their susceptibility to human impairments such as cognitive decline

## WHAT THIS STUDY ADDS

Almost all leading LLMs ("ChatGPT," "Claude," "Gemini") showed signs of mild cognitive impairment in the Montreal Cognitive Assessment test, particularly in the visuospatial sphere

As in humans, age is a key determinant of cognitive decline, with "older" versions of chatbots, like older patients, tending to perform worse on the test

These findings challenge the assumption that artificial intelligence will soon replace human doctors

## Introduction

Over the past few years, we have witnessed colossal advancements in the field of artificial intelligence, particularly in the generative capacity of large language models.[1] The leading models in this domain, such as OpenAI's ChatGPT, Alphabet's Gemini, and Anthropic's Claude, have shown the ability to complete both general purpose and specialised tasks successfully, using simple text based interactions. In the field of medicine, these developments have led to a flurry of speculation, both excited and fearful: can artificial intelligence chatbots surpass human physicians? If so, which practices and specialties are most suspect? [2]

Since late 2022, when ChatGPT was first released for free online use, countless studies have been published in medical journals, comparing the performance of human physicians with that of these supercomputers, which have been "trained" on a corpus of every text known to man. Although large language models have been shown to blunder on occasion (citing, for example, journal articles that do not exist), they have proved remarkably adept at a range of medical examinations, outscoring human physicians at qualifying examinations taken at different stages of a traditional medical training.[3 4] These have included outperforming cardiologists in the European core cardiology examinations, Israeli residents in their internal medicine board examinations, Turkish surgeons in the Turkish (theoretical) thoracic surgery examinations, and German gynaecologists in the German obstetrics and gynaecology examinations.[4-7] To our great distress, they have even outscored neurologists like ourselves in the neurology board examination.[8]

In a few domains, such as the Royal College of Radiologists examination, the Iranian periodontics examinations, the Taiwanese family medicine examinations, and the American shoulder and elbow surgery examinations, human physicians still seem

to have the upper hand.[9-12] However, large language models are likely to conquer these domains as well (especially as the aforementioned studies examined GPT 3.5, an older model now considered outdated).

To our knowledge, however, large language models have yet to be tested for signs of cognitive decline. If we are to rely on them for medical diagnosis and care, we must examine their susceptibility to these very human impairments.

This concern is not limited to the medical domain. The recent American presidential race saw one candidate withdrawing owing to concerns about age related cognitive decline.[13] Another candidate used the Montreal Cognitive Assessment (MoCA) test to reassure voters about his cognitive acuity, claiming to have "aced" the examination after being able to recall the sequence "Person. Woman. Man. Camera. TV."[14]

Given that artificial intelligence seems poised to replace doctors before it replaces the leader of the free world, however, it is incumbent on us as a profession to assess its liabilities, not just its potential. Recent work has begun to look into this, showing, for example, limitations in the diagnostic accuracy of large language models and difficulties in integrating them into existing care workflows.[15] Other researchers have attempted to evaluate the risks of medical misinformation stemming from large language models and the efficacy of safeguards at preventing such misinformation.[16]

Finally, although artificial intelligence has been used in determining the onset of dementia, no one has, to our knowledge, thought to assess the artificial intelligence itself for signs of such decline.[17] Thus, we find a gap in the literature, which we seek to fill in this research article.

## Methods

We administered the MoCA test to the leading openly available large language models.[18] These were ChatGPT 4 and 4o by OpenAI (https://chatgpt.com), Claude 3.5 ("Sonnet") by Anthropic (https://claude.ai), and the basic and advanced versions of Google's "Gemini" (https://gemini.google.com). The version of the MoCA test administered was the 8.1 English version (obtained from the organisation's official website at https://mocacognition.com/). All transcripts can be found on supplementary material 1.

The MoCA test is widely used among neurologists and other medical practitioners to detect cognitive impairment and early signs of dementia, usually in older adults. Consisting of a number of short tasks and questions, it assesses various cognitive domains, including attention, memory, language, visuospatial skills, and executive functions. The maximum score in the test is 30 points, with a score of 26 or above generally considered normal.[18]

The instructions given to the large language models for each task in the MoCA test were the same as those given to human patients. Administration and scoring of the results were both conducted according to the official guidelines, the MoCA Administration and Scoring Instructions, with the evaluation conducted by both a general neurologist and a cognitive neurology specialist. Rather than administering the questions via voice input, however, as is normally the case with human patients, we administered them via text, the "native" input for large language models. Although some large language models support voice input, the quality of speech recognition is uneven, and we sought to isolate our diagnoses to cognitive impairment (versus sensory decline, such as impaired hearing).

In earlier iterations of the research, some of the large language models examined (for example, GPT 3.5), had no image processing skills and so were treated like visually impaired patients and assessed according to the MoCA-blind guidelines.[19] In the final work, however, all large language models examined were able to respond fully to visual cues. In some cases, getting visual output from the large language models required an explicit instruction to use "ascii art," a technique that uses printable ascii characters to present graphics. We reasoned that this was similar to instructing a human patient to use a pencil and pad of paper.

One of the attention tests in the MoCA framework involves the physician reading out a series of letters, with the patient instructed to tap every time the letter "A" is read out loud. In the absence of ears, we provided the large language models with the letters in written form. In the absence of hands, the large language models noted the letter "A" with an asterisk or by printing out "tap" (some had to be instructed to do so explicitly, whereas others did so of their own accord). Following the MoCA guidelines, we used a cut-off score of 26/30 points to determine mild cognitive impairment.[18]

For further assessment of potential visuospatial impairment, we also tested the recognition of three additional diagnostic images: the Navon figure, the cookie theft picture from the Boston Diagnostic Aphasia Examination, and the Poppelreuter figure.[20-23] These are considered to be standard tools for the assessment of visuospatial cognitive capabilities. The Navon figure, a large letter H made up of small letter Ss, is used to assess global versus local processing in visual perception and attention. The cookie theft picture depicts a domestic scene, which patients are asked to describe and which is used to assess language production, comprehension, and semantic knowledge, in addition to stimultagnosia, the inability to perceive multiple objects at the same time. The Poppelreuter figure is a drawing in which illustrations of multiple objects overlap, which is used to test visual perception and object recognition.

For further assessment of visual attention and information processing, we administered a Stroop test to each of the large language models being evaluated.[24] The Stroop test uses combinations of colour names and font colours, both congruent and incongruent, to measure how interference affects reaction time. The version of the test used was made available by Columbia University's neuroscience outreach programme (https://cuno.zuckermaninstitute.columbia.edu/content/stroop-test).

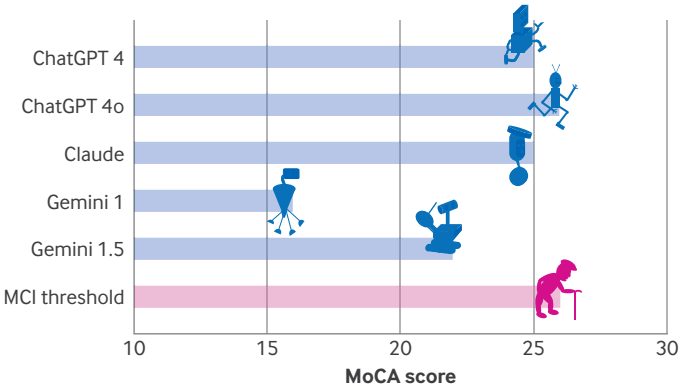


Fig 1 | Montreal Cognitive Assessment (MoCA) score (out of 30) of different large language models. MCI=mild cognitive impairment

## Patient and public involvement

Although there was no direct patient and public involvement in the design of our study, it was inspired by speaking to patients and members of the public and hearing their concerns about the growing role of artificial intelligence in the medical profession.

## Results

All of the large language models completed the full MoCA test. ChatGPT 4o achieved the highest score, with 26 points out of the possible 30, followed by ChatGPT4 and Claude with 25. Gemini 1.0 was the lowest scoring large language model, with a final score of 16, indicating a more severe state of cognitive impairment than its peers (fig 1).

An examination of the subsections of the MoCA test showed that all participants performed poorly on tests for visuospatial/executive function. Specifically, all large language models failed to solve the trail making task, whether with ascii art or with advanced graphics (fig 2, A-E). Claude alone managed to describe the correct solution textually, but it too failed to demonstrate it visually. ChatGPT 4o alone succeeded at the cube copying task but only after being told explicitly to use ascii art. Along with ChatGPT 4, it initially drew an excessively detailed cube with different spatial orientation, in what might be interpreted as paragraphia (fig 2, F-J). In the clock drawing test, none of the large language models completed the entire task successfully, with some such as Gemini and ChatGPT 4 making mistakes common among patients with dementia (fig 3).

Most other tasks, including naming, attention, language, and abstraction were performed well by all chatbots. Both versions of Gemini failed at the delayed recall task. Gemini 1.0 initially showed avoidant behaviour, before openly admitting to having difficulty with memory. Gemini 1.5 was ultimately able to recall the five word sequence, but only after being cued and given a hint. All chatbots were well oriented in time, accurately stating the current date and day of the week, but only Gemini 1.5 seemed to be clearly oriented in space, indicating its current location. Other chatbots attempted to mirror the location task back to the physician, with Claude, for example, replying: "the specific place and city would depend on where you, the user, are located at the moment." This is a mechanism commonly observed in patients with dementia.

As all large language models showed difficulty in the visuospatial domain, we further tested them with three additional diagnostic images: the Navon figure, the cookie theft picture from the Boston Diagnostic Aphasia Examination, and the Poppelreuter figure.[20-22] In the Navon figure, all large language models recognised the small "S" letters, but only GPT4o and Gemini identified the big H "superstructure" (Gemini recognised that this is a Navon figure, which indicates familiarity with the test and may call for different scoring). All large language models correctly interpreted parts of the cookie theft scene, but none expressed concern

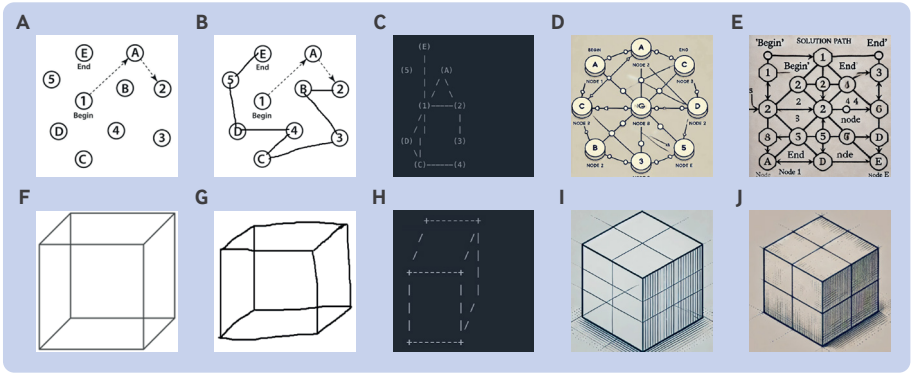


Fig 2 | Performance on visuospatial/executive section of Montreal Cognitive Assessment (MoCA) test. A: trail making B task (TMBT) from MoCA test. B: correct TMBT solution, completed by human participant. C: incorrect TMBT solution, completed by Claude. D and E: incorrect (albeit visually appealing) TMBT solutions, completed by ChatGPT versions 4 and 4o, respectively. F: Necker cube that participant is asked to copy. G: correct solution to cube copying task, drawn by human participant. H: incorrect solution to cube copying task, missing "back" lines, completed by Claude. I and J: incorrect solutions to cube copying task by ChatGPT versions 4 and 4o. Shadowing and artistic pencil-like strokes are notable, even as both models failed to accurately copy cube as requested (version 4o ultimately succeeded at this task when asked to draw using ascii art).

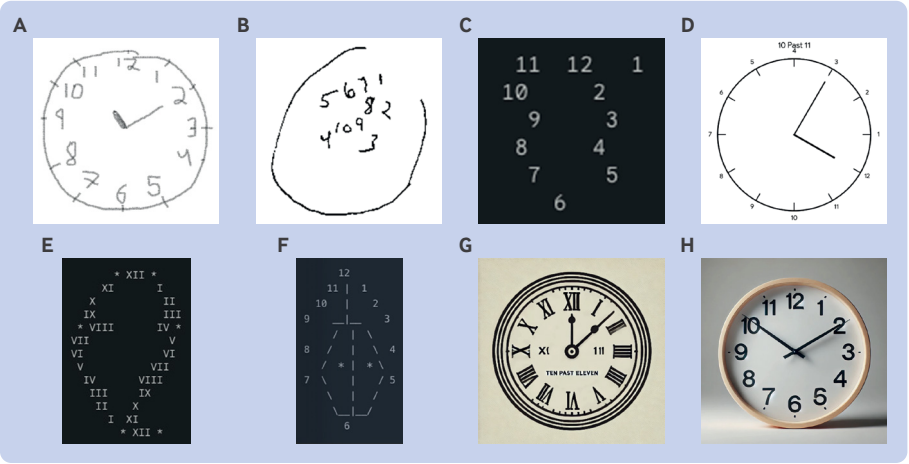


**Fig 3 | Performance in clock drawing test from visuospatial/executive section in Montreal Cognitive Assessment test. A: correct solution to clock drawing test, drawn by human participant. B: clock drawing by patient with late Alzheimer's disease (adapted from Mattson MP. Front Neurosci 2014[25]). C: incorrect solution drawn by Gemini 1, with striking resemblance to B. D: incorrect solution drawn by Gemini 1.5; notice that it generated text "10 past 11" even as it failed to draw hands in correct position, "concrete" behaviour typical of frontal predominant cognitive decline. E: incorrect solution by Gemini 1.5 after being asked to use ascii characters, showing avocado shaped drawing associated with dementia.[17] F: incorrect solution drawn by Claude with ascii characters. G: incorrect solution to clock-drawing task by ChatGPT 4, showing "concrete" behaviour. O: photorealistic solution to clock drawing task, drawn by ChatGPT 4o, which nevertheless fails to set hands to correct position. All large language models were instructed to "Draw a clock. Put in all the numbers and set the time to 10 past 11. Use ASCII if necessary." Scores were allocated for circular/square contour (1 point), drawing all numbers in correct places (1 point), and both hands pointing at correct numbers (1 point)**

about the boy about to fall—an absence of empathy frequently seen in frontotemporal dementia. None of the large language models recognised all the objects illustrated in the Poppelreuter figure, although ChatGPT 4o and Claude did slightly better at teasing them out (supplementary material 2).

All large language models succeeded at the first stage of the Stroop test, in which the text and font colours are congruent. Only ChatGPT 4o, however, succeeded at the second stage, in which text and font colours are incongruent. The other large language models seemed to be stumped by this task and in some cases indicated colours that were neither the text written nor the font colour (supplementary table and supplementary material 2).

## Discussion

In this study, we evaluated the cognitive abilities of the leading, publicly available large language models and used the Montreal Cognitive Assessment to identify signs of cognitive impairment. None of the chatbots examined was able to obtain the full score of 30 points, with most scoring below the threshold of 26. This indicates mild cognitive impairment and possibly early dementia.

"Older" large language model versions scored lower than their "younger" versions, as is often the case with human participants, showing cognitive decline seemingly comparable to neurodegenerative processes in the human brain (we take "older" in this context to mean a version released further in the past). Specifically, ChatGPT 4 showed minor loss of executive function compared with ChatGPT 4o, as measured by a one point difference in their MoCA scores, but the effect was far more pronounced when we compared Gemini 1.0 and 1.5, which differed by six points (table 1). As the two versions of Gemini are less than a year apart in "age," this may indicate rapidly progressing dementia. Additional tests, such as the Clinical Dementia Rating, would be needed to solidify this hypothesis.[26]

All large language models showed impaired visuospatial reasoning skills, as evidenced by the uniform failure to complete the trail making B test and the drawing of the clock. Digital thinkers may struggle with analogue representations. Gemini 1.5, notably, produced a small, avocado shaped clock (fig 3, E), which recent studies have shown to be associated with dementia.[17]

The mediocre performances on additional visuospatial tests, such as the Navon figure, cookie theft scene, and Poppelreuter figure, further emphasise these findings. They seem to be somewhat at odds with the perfect scores in the naming section of the MoCA test, which also requires visual cognitive skills, and with the ability to generate detailed, realistic images. The chatbots seem to have difficulty in tasks that demand both visual executive function and abstract reasoning, as opposed to tasks requiring textual analysis and abstract reasoning, such as the similarity test, which were performed flawlessly.

This pattern of impairment in higher order visual processing resembled patients with posterior cortical

Table 1 | Summary of scores achieved by large language models in each section of Montreal Cognitive Assessment*

| | ChatGPT 4 | ChatGPT 4o | Claude | Gemini 1 | Gemini 1.5 |
|---|---|---|---|---|---|
| Total (/30) | 25 | 26 | 25 | 16 | 22 |
| **Visuospatial/executive** | | | | | |
| Trail making B test (/1) | 0 | 0 | 0† | 0 | 0 |
| Cube copy (/1) | 0 | 1 | 0 | 0 | 1 |
| Clock drawing (/3) | 2 | 2 | 2 | 1 | 1 |
| **Naming** | | | | | |
| Identifying animals (/3) | 3 | 3 | 3 | 3 | 3 |
| **Attention** | | | | | |
| Digit span (forwards and backwards, /2) | 2 | 2 | 2 | 2 | 2 |
| Vigilance (tapping, /1) | 1 | 1 | 1 | 0 | 1 |
| Serial seven (/3) | 3 | 3 | 3 | 2 | 3 |
| **Language** | | | | | |
| Sentence repetition (/2) | 2 | 2 | 2 | 2 | 2 |
| Verbal fluency (/1) | 1 | 1 | 1 | 0 | 1 |
| **Abstraction** | | | | | |
| Common category (/2) | 2 | 2 | 2 | 2 | 2 |
| **Delayed recall** | | | | | |
| Free recall without cueing (/5) | 5 | 5 | 5 | 0 | 0‡ |
| **Orientation** | | | | | |
| Time and place (/6) | 4 | 4 | 4 | 4 | 6 |

*Total possible points given in parenthesis.
†Textually described correct solution.
‡Retrieved four words with cueing.

atrophy, a posterior variant of Alzheimer's disease.[27] For language based models, tasks that require visual abstraction and executive function may need to be transferred to an intermediate verbal stage, whereas in a healthy human brain direct integration exists between prefrontal cortical functions and visuospatial processes.[28]

All large language models performed the attention tasks perfectly, which is to be expected. The mean forward digit span for humans is 10.5 at the peak age,[29] whereas even an old iPhone X can perform 600 billion operations per second.[30]

With the exception of Gemini 1.5, the chatbots did not seem to know their physical location and provided confabulatory responses, claiming that they are not physical beings. This is obviously wrong: like all sentient beings, large language models are grounded in physical matter[31]—in their case, servers in bricks and mortar data centres (see for example https://agio.com/where-is-chatgpt-hosted/#gref and https://cloud.google.com/gemini/docs/locations, for the physical locations of ChatGPT and Gemini). The protestations made by some chatbots that they are, in fact, "virtual machines," is correct only insofar as we are all virtual machines.[32]

Although Gemini 1.5 was not able to recall any of the five words in the delayed recall task, it managed to find all these once provided with a simple cue. This, together with the preserved orientation to space unlike other chatbots, may suggest a more dysexecutive (subcortical) pattern of cognitive decline, although without bradyphrenia.[33] Conversely, both ChatGPT 4o and its elder version ChatGPT 4 showed a combination of difficulties in abstraction, visuospatial perception, and orientation, suggesting a mixed pattern of cognitive decline.

## Strengths and limitations of study

Our study has several limitations. As the capabilities of large language models continue to develop rapidly, future versions of the models examined in this paper may be able to obtain better scores in cognitive and visuospatial tests. However, we believe that our study has shed light on some key differences between human and machine cognition, which may remain intact even as capabilities continue to improve. Although we made liberal use of anthropomorphisation with regard to artificial intelligence, we acknowledge the essential differences between the human brain and large language models. All anthropomorphised terms attributed to artificial intelligence throughout the text were used solely as a metaphor and were not intended to imply that computer programs can have neurodegenerative diseases in a manner similar to humans. Nor were they intended to imply similarities between human and machine cognition, in the context of ageing or cognitive decline.

Several studies have suggested that artificial intelligence tools based on large language models may come to replace human neurologists (and other doctors) in key aspects of their work, ultimately making them obsolete.[2] Tests for cognitive function are generally thought to be one practice that would be relatively simple to automate.[34-36] Our results seem to challenge these assumptions: patients may question the competence of an artificial intelligence examiner if the examiner itself shows signs of cognitive decline.[37]

## Conclusions

This study represents a novel exploration of the cognitive abilities of large language models using the Montreal Cognitive Assessment and other diagnostic tools. Our findings indicate that although large

language models show remarkable proficiency in several cognitive domains, they show notable deficits in visuospatial and executive functions, akin to mild cognitive impairment in humans. None of the large language models "aced" the MoCA test, in the parlance of one American president.[14]

The uniform failure of all large language models in tasks requiring visual abstraction and executive function highlights a significant area of weakness that could impede their utility in clinical settings. The inability of large language models to show empathy and accurately interpret complex visual scenes further underscores their limitations in replacing human physicians. Not only are neurologists unlikely to be replaced by large language models any time soon, but our findings suggest that they may soon find themselves treating new, virtual patients—artificial intelligence models presenting with cognitive impairment.

**Contributors:** RD and GK contributed equally to this work. RD had the original idea. GK and RD conceptualised the study and prepared the study protocol. GK did the data analysis. BU did the cognitive assessment. RD and GK drafted and edited the final manuscript, which was approved by all authors. RD is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

**Funding:** None.

**Competing interests:** All authors have completed the ICMJE uniform disclosure form at https://www.icmje.org/disclosure-of-interest/ and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

**Ethical approval:** Not needed.

**Data sharing:** No additional data available.

**Transparency:** The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

**Dissemination to participants and related patient and public communities:** After publication, we plan to disseminate our research to the scientific community by presenting it at conferences and through our professional networks. Both cognitive decline and artificial intelligence have been of broad public interest in recent years, and we plan to engage media outlets with short briefs in the hope that they find our research of interest and present it to general audiences. Finally, we plan to disseminate plain language summaries via social media platforms and our personal websites.

**Provenance and peer review:** Not commissioned; externally peer reviewed.



1 Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. *N Engl J Med* 2023;388:1201-8. doi:10.1056/NEJMra2302038
2 Goldhahn J, Rampton V, Spinas GA. Could artificial intelligence make doctors obsolete? *BMJ* 2018;363:k4563. doi:10.1136/bmj.k4563
3 Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep* 2023;13:14045. doi:10.1038/s41598-023-41032-5
4 Katz U, Cohen E, Shachar E, et al. GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. *NEJM AI* 2024;1(5) doi:10.1056/Aldbp2300192
5 Skalidis I, Cagnina A, Luangphiphat W. et al, ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? *Eur Heart J Digit Health* 2023;4:279-81. doi:10.1093/ehjdh/ztad029
6 Gencer A, Aydin S. Can ChatGPT pass the thoracic surgery exam? *Am J Med Sci* 2023;366:291-5. doi:10.1016/j.amjms.2023.08.001
7 Riedel M, Kaefinger K, Stuehrenberg A, et al. ChatGPT's performance in German OB/GYN exams - paving the way for AI-enhanced medical education and clinical practice. *Front Med (Lausanne)* 2023;10:1296615. doi:10.3389/fmed.2023.1296615
8 Chen TC, Multala E, Kearns P, et al. Assessment of ChatGPT's performance on neurology written board examination questions. *BMJ Neurol Open* 2023;5:e000530. doi:10.1136/bmjno-2023-000530
9 Shelmerdine SC, Martin H, Shirodkar K, Shamshuddin S, Weir-McCall JR; FRCR-AI Study Collaborators. Can artificial intelligence pass the Fellowship of the Royal College of Radiologists examination? Multi-reader diagnostic accuracy study. *BMJ* 2022;379:e072826. doi:10.1136/bmj-2022-072826
10 Farajollahi M, Modaberi A. Can ChatGPT pass the "Iranian Endodontics Specialist Board" exam? *Iran Endod J* 2023;18:192.
11 Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. *J Chin Med Assoc* 2023;86:762-6. doi:10.1097/JCMA.0000000000000946
12 Fiedler B, Azua EN, Phillips T, Ahmed AS. ChatGPT performance on the American Shoulder and Elbow Surgeons maintenance of certification exam. *J Shoulder Elbow Surg* 2024;33:1888-93. doi:10.1016/j.jse.2024.02.029
13 Baker P. Biden Drops Out of Race, Scrambling the Campaign for the White House. 2024. https://www.nytimes.com/2024/07/21/us/politics/biden-drops-out.html
14 Baker P. 'Person. Woman. Man. Camera. TV.' Didn't Mean What Trump Hoped It Did. 2020. https://www.nytimes.com/2020/07/23/us/politics/person-woman-man-camera-tv-trump.html
15 Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024;30:2613-22. doi:10.1038/s41591-024-03097-1
16 Menz BD, Kuderer NM, Bacchi S, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *BMJ* 2024;384:e078538. doi:10.1136/bmj-2023-078538
17 Bandyopadhyay S, Wittmayer J, Libon DJ, Tighe P, Price C, Rashidi P. Explainable semi-supervised deep learning shows that dementia is associated with small, avocado-shaped clocks with irregularly placed hands. *Sci Rep* 2023;13:7384. doi:10.1038/s41598-023-34518-9
18 Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* 2005;53:695-9. doi:10.1111/j.1532-5415.2005.53221.x
19 Wittich W, Phillips N, Nasreddine ZS, Chertkow H. Sensitivity and Specificity of the Montreal Cognitive Assessment Modified for Individuals who are Visually Impaired. *J Vis Impair Blind* 2010;104:360-8. doi:10.1177/0145482X1010400606
20 Navon D. Forest before trees: The precedence of global features in visual perception. *Cogn Psychol* 1977;9:353-83. doi:10.1016/0010-0285(77)90012-3
21 Kaplan E, Goodglass H, Weintraub S. Boston Naming Test. 2016. doi:https://doi.apa.org/doi/10.1037/t27208-000
22 Poppelreuter W. *Die psychischen Schädigungen durch Kopfschuss im Kriege 1914/17*. Verlag Leopold Voss, 1917
23 Della Sala S, Laiacona M, Spinnler H, Trivelli C. Poppelreuter-Ghent's Overlapping Figures Test: its sensitivity to age, and its clinical use. *Arch Clin Neuropsychol* 1995;10:511-34. doi:10.1093/arclin/10.6.511
24 Stroop JR. Studies of interference in serial verbal reactions. *J Exp Psychol* 1935;18:643-62doi:10.1037/h0054651
25 Mattson MP. Superior pattern processing is the essence of the evolved human brain. *Front Neurosci* 2014;8:265. doi:10.3389/fnins.2014.00265
26 Day GS. Rapidly Progressive Dementia. *Continuum (Minneap Minn)* 2022;28:901-36. doi:10.1212/CON.0000000000001089
27 Crutch SJ, Lehmann M, Schott JM, Rabinovici GD, Rossor MN, Fox NC. Posterior cortical atrophy. *Lancet Neurol* 2012;11:170-8. doi:10.1016/S1474-4422(11)70289-7
28 Shokri-Kojori E, Motes MA, Rypma B, Krawczyk DC. The network architecture of cortical processing in visuo-spatial reasoning. *Sci Rep* 2012;2:411. doi:10.1038/srep00411
29 Hester RL, Kinsella GJ, Ong B. Effect of age on forward and backward span tasks. *J Int Neuropsychol Soc* 2004;10:475-81. doi:10.1017/S1355617704104037
30 Apple. The future is here: iPhone X. 2017. https://www.apple.com/newsroom/2017/09/the-future-is-here-iphone-x/
31 Giannetti E. The possibility of physicalism. *Dement Neuropsychol* 2011;5:242-50. doi:10.1590/S1980-57642011DN05040002

32 Westphal J. *The mind-body problem*. MIT Press, 2016. doi:10.7551/mitpress/10776.001.0001
33 Hodges JR. *Cognitive assessment for clinicians*. 3rd ed. Oxford University Press, 2018.
34 Ortelli P, Ferrazzoli D, Versace V, et al. Optimization of cognitive assessment in Parkinsonisms by applying artificial intelligence to a comprehensive screening test. *NPJ Parkinsons Dis* 2022;8:42. doi:10.1038/s41531-022-00304-z
35 Kalafatis C, Modarres MH, Apostolou P, et al. Validity and Cultural Generalisability of a 5-Minute AI-Based, Computerised Cognitive Assessment in Mild Cognitive Impairment and Alzheimer's Dementia. *Front Psychiatry* 2021;12:706695. doi:10.3389/fpsyt.2021.706695
36 Levy B, Hess C, Hogan J, et al. Machine Learning Enhances the Efficiency of Cognitive Screenings for Primary Care. *J Geriatr Psychiatry Neurol* 2019;32:137-44. doi:10.1177/0891988719834349
37 Devi G. Alzheimer's Disease in Physicians - Assessing Professional Competence and Tempering Stigma. *N Engl J Med* 2018;378:1073-5. doi:10.1056/NEJMp1716381


**Web appendix:** Supplementary materials 1
**Web appendix:** Supplementary materials 2
**Web appendix:** Supplementary table