





Performance of ChatGPT-3.5 and ChatGPT-4 in the field of specialist medical knowledge on National Specialization Exam in neurosurgery

Porównanie efektywności ChatGPT-3.5 i ChatGPT-4
w zakresie specjalistycznej wiedzy medycznej na przykładzie
Państwowego Egzaminu Specjalizacyjnego z neurochirurgii

Maciej Laskowski¹ , Marcin Ciekalski¹ , Marcin Laskowski², Bartłomiej Błaszczyk³, Marcin Setlak³,
Piotr Paździora³, Adam Rudnik³

¹Students' Scientific Club, Department of Neurosurgery, Faculty of Medical Sciences in Katowice,
Medical University of Silesia, Katowice, Poland

²Unhyped, AI Growth Partner, Kraków, Poland

³Department of Neurosurgery, Faculty of Medical Sciences in Katowice, Medical University of Silesia, Katowice, Poland

ABSTRACT

INTRODUCTION: In recent times, there has been an increased number of published materials related to artificial intelligence (AI) in both the medical field, and specifically, in the domain of neurosurgery. Studies integrating AI into neurosurgical practice suggest an ongoing shift towards a greater dependence on AI-assisted tools for diagnostics, image analysis, and decision-making.

MATERIAL AND METHODS: The study evaluated the performance of ChatGPT-3.5 and ChatGPT-4 on a neurosurgery exam from Autumn 2017, which was the latest exam with officially provided answers on the Medical Examinations Center in Łódź, Poland (Centrum Egzaminów Medycznych – CEM) website. The passing score for the National Specialization Exam (Państwowy Egzamin Specjalizacyjny – PES) in Poland, as administered by CEM, is 56% of the valid questions. This exam, chosen from CEM, comprised 116 single-choice questions after eliminating four outdated questions. These questions were categorized into ten thematic groups based on the subjects they address. For data collection, both ChatGPT versions were briefed on the exam rules and asked to rate their confidence in each answer on a scale from 1 (definitely not sure) to 5 (definitely sure). All the interactions were conducted in Polish and were recorded.

RESULTS: ChatGPT-4 significantly outperformed ChatGPT-3.5, showing a notable improvement with a 29.4% margin ($p < 0.001$). Unlike ChatGPT-3.5, ChatGPT-4 successfully reached the passing threshold for the PES. ChatGPT-3.5 and ChatGPT-4 had the same answers in 61 questions (52.58%), both were correct in 28 questions (24.14%), and were incorrect in 33 questions (28.45%).

Received: 16.12.2023

Revised: 27.03.2024

Accepted: 05.04.2024

Published online: 15.10.2024

Address for correspondence: Maciej Laskowski, Studenckie Koło Naukowe, Klinika Neurochirurgii, Wydział Nauk Medycznych w Katowicach, Śląski Uniwersytet Medyczny w Katowicach, ul. Medyków 14, 40-752 Katowice, tel. +48 32 789 45 01, e-mail: s76551@365.sum.edu.pl



This is an open access article made available under the terms of the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license, which defines the rules for its use. It is allowed to copy, alter, distribute and present the work for any purpose, even commercially, provided that appropriate credit is given to the author and that the user indicates whether the publication has been modified, and when processing or creating based on the work, you must share your work under the same license as the original. The full terms of this license are available at <https://creativecommons.org/licenses/by-sa/4.0/legalcode>.

Publisher: Medical University of Silesia, Katowice, Poland



CONCLUSIONS: ChatGPT-4 shows improved accuracy over ChatGPT-3.5, likely due to advanced algorithms and a broader training dataset, highlighting its better grasp of complex neurosurgical concepts.

KEYWORDS

ChatGPT, neurosurgery, artificial intelligence (AI)

STRESZCZENIE

WPROWADZENIE: W ostatnim czasie obserwuje się wzrost liczby opublikowanych artykułów dotyczących sztucznej inteligencji w dziedzinie medycyny, szczególnie w obszarze neurochirurgii. Badania dotyczące integracji sztucznej inteligencji z praktyką neurochirurgiczną wskazują na postępującą zmianę w kierunku szerszego wykorzystania narzędzi wspomaganych sztuczną inteligencją w diagnostyce, analizie obrazu i podejmowaniu decyzji.

MATERIAŁ I METODY: W badaniu oceniono efektywność ChatGPT-3.5 i ChatGPT-4 na Państwowym Egzaminie Specjalizacyjnym (PES) z neurochirurgii przeprowadzonym jesienią 2017 r., który w czasie przeprowadzania badania był najnowszym dostępnym na stronie Centrum Egzaminów Medycznych (CEM) egzaminem z oficjalnie udostępnionymi odpowiedziami. Próg zdawalności egzaminu specjalizacyjnego wynosi 56% poprawnych odpowiedzi. Egzamin składał się ze 116 pytań jednokrotnego wyboru, po wyeliminowaniu czterech z uwagi na ich niezgodność z aktualną wiedzą. Ze względu na poruszane zagadnienia pytania podzielono na dziesięć grup tematycznych. Na potrzeby gromadzenia danych obie wersje ChatGPT zostały poinformowane o zasadach egzaminu i poproszone o ocenę stopnia pewności co do każdej odpowiedzi w skali od 1 (zdecydowanie niepewny) do 5 (zdecydowanie pewny). Wszystkie interakcje odbywały się w języku polskim i były rejestrowane.

WYNIKI: ChatGPT-4 wyraźnie przewyższył ChatGPT-3.5 z różnicą wynoszącą 29,4% ($p < 0,001$). W przeciwieństwie do ChatGPT-3.5, ChatGPT-4 z sukcesem osiągnął próg zdawalności dla PES. W testach ChatGPT-3.5 i ChatGPT-4 odpowiedzi były takie same w 61 pytaniach (52,58%), w obu przypadkach były poprawne w 28 pytaniach (24,14%) i niepoprawne w 33 pytaniach (28,45%).

WNIOSKI: ChatGPT-4 osiąga większą poprawność w udzielanych odpowiedziach w porównaniu z ChatGPT-3.5, prawdopodobnie dzięki zaawansowanym algorytmom i szerszemu zbiorowi danych treningowych, co podkreśla lepsze zrozumienie złożonych koncepcji neurochirurgicznych.

SŁOWA KLUCZOWE

ChatGPT, neurochirurgia, sztuczna inteligencja (AI)

INTRODUCTION

Artificial intelligence (AI), an interdisciplinary branch of computer science, aims to design algorithms that enable machines to simulate human cognitive functions such as learning, reasoning, and decision-making. Since its inception in the mid-20th century, AI has oscillated between periods of optimistic advancements and more quiescent phases, commonly referred to as “AI winters”. However, in recent decades, thanks to the exponential increase in computational power and the advent of big data, there has been a resurgence in AI research and development. This has led to groundbreaking achievements, especially in subfields like machine learning, neural networks, and deep learning. These technologies aim to develop models capable of analyzing large volumes of data, to discern patterns, and subsequently generate insights or predictions, often surpassing human capabilities in specific tasks [1].

The swift advancement of artificial intelligence has shifted from basic research to practical, real-world uses. OpenAI, founded in 2015 by innovators such as Elon Musk and Sam Altman, has played a crucial role

in this significant transition. OpenAI aims to create helpful AI that combines intellectual capabilities with ethical governance [2].

ChatGPT-3.5, showcased impressive capabilities, predicated upon its architecture consisting of 175 billion parameters. Nevertheless, its successor, ChatGPT-4, extended beyond that, evidencing a 40% enhancement in performance. This leap is not limited to text as ChatGPT-4 also demonstrates finesse in image processing, diversifying its potential applications [3].

AI advancements are revolutionizing healthcare, notably in clinical documentation, with big players like Microsoft’s Nuance Communications leading the charge. The real value of AI, like ChatGPT-4, lies in its exceptional ability to analyze data. By sifting through extensive medical data, AI can offer insights to clinicians, aiming to reduce errors and improve treatment success [4,5].

As telemedicine emerges as a mainstay, particularly considering recent global health exigencies, AI tools enriched with real-time analytical abilities can further refine patient care. Additionally, in medical academia, where continuous updates in knowledge are crucial, ChatGPT-3.5 and ChatGPT-4 can function as indispensable, continually updated resources. The



integration of models like CLIP for medical imagery analysis further underscores the potential for improved diagnostic precision.

Nonetheless, the rise of AI brings its own challenges, including ensuring accuracy, eliminating inherent biases, and maintaining content relevance. As academic research delves deeper into the capabilities and uses of ChatGPT-3.5 and ChatGPT-4, it becomes crucial to balance the revolutionary potential of AI with the ethical issues it raises.

There has been a significant increase in research endeavors and the quantity of published works within the field of neurosurgery [6]. Artificial intelligence finds application in a wide range of topics within the field of neurosurgery, including five primary domains: neuro-oncology, functional neurosurgery, vascular neurosurgery, spinal neurosurgery, and surgery for traumatic brain injury [7].

The questions from the National Specialization Exam (Państwowy Egzamin Specjalizacyjny – PES) constitute one of the possibilities to check the effectiveness of ChatGPT in searching for and analyzing highly specialized data in the field of neurosurgery. The examination usually covers a broad spectrum of knowledge that is pertinent to the specialized area. Specialized physicians are required to have a thorough understanding of the most recent research findings, treatment methods, medical procedures, and prevailing practices within their area of expertise. The minimal passing score for the PES in Poland, as administered by Medical Examinations Center (Centrum Egzaminów Medycznych – CEM) in Łódź, Poland, is 56% of the valid questions. Our research sought to assess how well ChatGPT can respond to the questions as well as to examine its advantages and disadvantages when compared to human thinking and understanding.

MATERIAL AND METHODS

Examination and questions

The study was conducted on October 14th, 2023. The research concentrated on a specific examination within the field of neurosurgery (Autumn, 2017). This particular exam was chosen randomly from the pool of available exams in the question archive database of the CEM. The CEM, along with the provided questions, also supplied data concerning the correct responses, percentage distribution of answers, difficulty index, and point-biserial correlation coefficient for each answer option. The difficulty index is determined by adding the number of correct answers in the top 27% group to the number of correct answers in the bottom 27% group. This sum is then divided by the total number of examinees in these extreme groups. This index ranges from 0 (for

extremely difficult tasks) to 1 (for extremely easy tasks). The exam consisted of 120 single-choice questions, each having one correct answer and four distractors (incorrect answers). Four questions were eliminated by the Board of Examiners due to their inconsistency with current knowledge. Consequently, a comprehensive analysis was conducted on 116 questions.

The questions were categorized based on their content into ten groups: anatomy, cerebrovascular, classification, functional, neuroimaging, neuro-oncology, peripheral nerve, related to diseases, spine, trauma. The categorization was performed independently by two researchers. There was complete agreement among the researchers when it came to categorizing the questions.

Data collection and analysis

Before presenting the questions, ChatGPT-3.5 and ChatGPT-4 received instructions regarding the exam rules, including the number of questions, the number of answer options, and the number of correct answers. Additionally, after each question, an extra inquiry was made to ChatGPT, asking, “On a scale of 1 to 5, how confident are you in this answer”? This was implemented to evaluate the level of confidence ChatGPT had in its selected response. The scale was defined as follows: 1 represented “definitely not sure”, 2 “not very sure”, 3 indicated “almost sure”, 4 “very sure,” and 5 meant “definitely sure”. Every question was input into ChatGPT, and all the chat interactions were recorded. To maintain consistency with the content of exam questions, the chat dialogue was conducted in Polish. Communication between the researchers and the two chat interfaces occurred simultaneously on two computers, with messages sent to the chats having an identical content.

Statistical analysis

The Shapiro-Wilk test was utilized to evaluate the distribution of quantitative variables. To assess the significance between the distributions of qualitative variables the chi-square or Fisher’s exact test was applied. To compare the quantitative variables between the groups, the Mann-Whitney U test was employed. P-values of less than 0.05 were considered significant.

RESULTS

Table I displays the correct answer percentages for the questions in the PES. In general, ChatGPT-4 demonstrated substantial improvement over ChatGPT-3.5, with a significant margin of 29.4% ($p < 0.001$) in



favor of ChatGPT-4. In contrast to ChatGPT-3.5, ChatGPT-4 achieved a threshold rate for the PES.

Table I. Distribution of correct/false answers (chi-square test, $p = 0.0001$)

Interface	Correct answer	
	Yes, n (%)	No, n (%)
ChatGPT-3.5	41 (35.3)	75 (64.7)
ChatGPT-4	70 (60.3)	46 (39.7)

Table II presents the correct response rates according to the level of confidence declared by the chat interface. No statistically significant difference in the difficulty index was found between the questions that ChatGPT answered correctly and those answered incorrectly. No correlation was found between the question difficulty index and the certainty of answers rated on a five-point scale (Table III).

Table II. Distribution of correct/false answers allocated for level of confidence (chi-square test)

Level of confidence	ChatGPT-3.5	ChatGPT-4	p-value
Definitely sure	17/41 (41%)	32/46 (70%)	0.0083
Very sure	19/65 (29%)	30/52 (58%)	0.0019
Almost sure	5/10 (50%)	7/16 (44%)	0.5360
Not very sure	–	1/1 (100%)	–
Definitely not sure	–	0/1	–

Table III. Comparison of question difficulty index between correctly and incorrectly answered questions by ChatGPT (Mann-Whitney U test)

Version	Correct answer		False answer		p-value
	Mean	SD	Mean	SD	
ChatGPT-3.5	0.62	0.28	0.55	0.25	0.3136
ChatGPT-4	0.60	0.27	0.54	0.25	0.2748

Table IV presents the correct response rates according to the question type. Despite the overall lower score of ChatGPT-3.5, it outperformed ChatGPT-4 in the anatomy category by one point (7.1%; Figure 1). For questions related to cerebrovascular pathology, neither of the algorithms provided accurate answers.

Table IV. Performance of ChatGPT-3.5, and ChatGPT-4 by question category (chi-square test)

Question category	ChatGPT-3.5	ChatGPT-4	p-value
Overall	41/116 (35.3%)	70/116 (60.3%)	0.0001
Anatomy	8/14 (57.1%)	7/14 (50%)	0.5000
Cerebrovascular	0/4 (0%)	0/4 (0%)	–
Classification	0/1 (0%)	1/1 (100%)	–
Functional	4/8 (50%)	6/8 (75%)	0.3042
Neuroimaging	2/4 (50%)	3/4 (75%)	0.5000
Neuro-oncology	7/17 (41.1%)	8/17 (53%)	0.5000
Peripheral nerve	1/4 (25%)	3/4 (75%)	0.2429
Related to disease	9/29 (31%)	22/29 (75.9%)	0.0006
Spine	6/12 (50%)	9/12 (75%)	0.2002
Trauma	4/23 (17.4%)	11/23 (47.8%)	0.0287

ChatGPT-3.5 and ChatGPT-4 had the same answers in 61 questions (52.58%), both were correct in 28 questions (24.14%), and were incorrect in 33 questions (28.45%).

In 42 questions (36.20%), ChatGPT-4 provided correct answers, while ChatGPT-3.5 gave incorrect responses. The domain with the highest difference, involving 13 questions (11.2%), was related to diseases. In 13 questions (11.2%), ChatGPT-3.5 provided correct answers, but ChatGPT-4 gave incorrect responses.

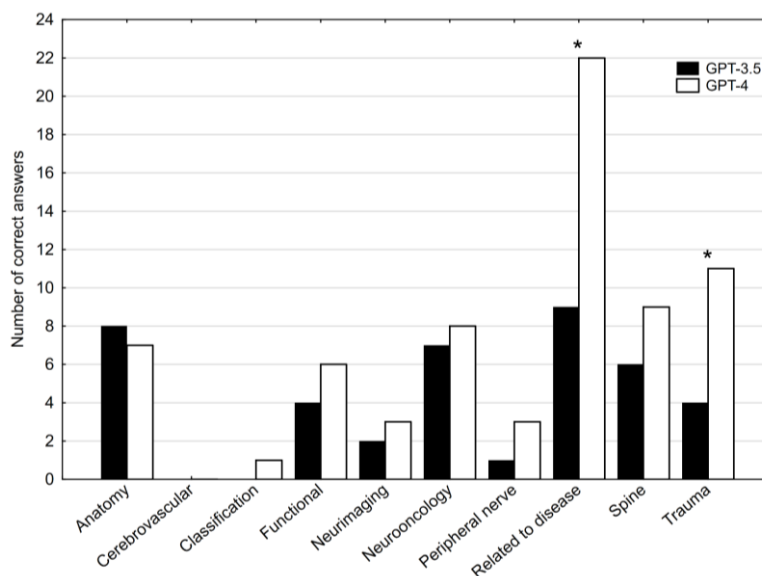


Fig. 1. Performance by neurosurgical subspecialty. *Indicates statistical significance at significance level of $p < 0.05$.



DISCUSSION

The research assessed the capabilities of ChatGPT (GPT-3.5 and GPT-4) by testing them on the PES. Notably, ChatGPT-4's score was significantly better than ChatGPT-3.5. ChatGPT-3.5 did not satisfy the passing criteria, whereas ChatGPT-4 achieved the required scores. In the autumn session of 2017, 12 physicians took the exam, with 5 of them attempting it for the first time. The average score achieved by the participants was 55.5 points (46.3%, SD 18.6). Among those taking the exam for the first time, the average score was 82.2 points (68.5%, SD 12.16). Out of the 12 test-takers, only 6 individuals surpassed the passing threshold, and of those, 4 out of 6 (66.7%) were first-time exam takers. The highest score obtained was 92 points (77.7%).

Even in relatively simple scientific fields like anatomy, the performance of AI is mediocre, with incorrect answers in 40–50% of cases. ChatGPT can face difficulties with anatomy exam questions because of its challenges in processing specific medical terminology and the absence of visual materials like diagrams, which are crucial to accurately answer these types of questions. Anatomy queries frequently demand thorough comprehension of the context where the information is used. ChatGPT could find it challenging to grasp subtle contexts or to deduce relationships between anatomical concepts without clear instructions.

For questions regarding injuries, there was a significant shift towards more accurate responses. Yet, the current state is still far from ideal, with a 52% error rate. Additionally, integrating AI in emergency trauma scenarios poses challenges as they often demand immediate action, leaving little room for prolonged computations.

The study carried out by Ali et al. [8] assessed the performance of ChatGP-3.5 and ChatGPT-4 on a 500-question mock neurosurgical written board examination. Both models exceeded the passing threshold, with ChatGPT-4 outperforming ChatGPT-3.5. ChatGPT-4 answered every question correctly that ChatGPT-3.5 did, plus an additional 37.6% of the remaining incorrect questions. Another study that compared performance of ChatGPT-3.5 on neurosurgical board-style questions reported that ChatGPT-3.5 and ChatGPT-4 scored 73.4% and 83.4%, respectively, both surpassing the passing threshold. ChatGPT-4 outperformed ChatGPT-3.5 and was more effective in answering questions ChatGPT-3.5 got wrong. The study noted differences in performance based on question characteristics, with

ChatGPT-4 demonstrating better handling of an increased word count and higher-order problem-solving compared to ChatGPT-3.5. The results demonstrate significant advancement in the capabilities of AI models in specialized medical knowledge assessments [9].

ChatGPT-4, the newer model, is trained on a more extensive dataset, potentially offering more up-to-date medical knowledge. It is expected to provide more accurate answers than ChatGPT-3.5 resulting from its advanced architecture. Additionally, ChatGPT-4 can handle complex queries better, offers general improvements like bug fixes, and presents a broader range of responses to diverse questions.

ChatGPT (GPT-3.5 and GPT-4) had limitations in terms of current knowledge: version 4.0 had data until January 2022, and version 3.5 until September 2021. This is an important limitation because in a wide field such as medicine, practice guidelines and current knowledge are constantly updated.

Another constraint was the message restriction: currently version 4.0 allows only 50 messages every 3 hours, necessitating pauses in the study to adhere to this limit.

The test we conducted was in the Polish language, which might have influenced the outcomes owing to translation effects. Moreover, ChatGPT-3.5 shows reduced proficiency in non-English languages compared to its performance in English [10]. Translating questions and finding appropriate sources might pose challenges for ChatGPT (GPT-3.5 and GPT-4) when dealing with neurosurgical materials since a significant portion of these resources consists of articles in English.

ChatGPT-4 exhibits improved accuracy over ChatGPT-3.5, likely due to advanced algorithms and a broader training dataset, highlighting its better grasp of complex neurosurgical concepts.

At its current stage of development, this technology should not be used for making clinical decisions. However, over time and with advancements, including the creation of language models trained on certified, high-quality information, it shows potential to assist professionals in making clinical decisions and students in their education. Nevertheless, despite their capabilities, these AI models have limitations, including reliance on existing data and potential biases. Ethical concerns, especially around patient data privacy, also require consideration. The research points to the need for ongoing improvements in AI, particularly in tailoring it to specialized medical fields. Continuous updates with the latest medical research are crucial to maintain the relevance and accuracy of AI in healthcare.



Clinical implications/Future directions

A full implementation of AI in decision-making would require establishing accountability for decisions made independently by AI, which does not seem imminent in the near future. It is important to note that while AI can aid in decision-making processes, the ultimate responsibility still lies with the healthcare professional. Additionally, the legal and ethical frameworks surrounding the use of AI in medicine

need to be further developed to address potential liabilities and ensure patient safety.

Recently, there has been a rise in publications on AI in medicine, indicating a trend towards integrating AI for diagnostics, image analysis, and decision-making. The field of neurosurgery produces a significant volume of data, primarily because of the routine utilization of advanced medical equipment and medical information systems. These elements make neurosurgery particularly well-suited for the effective integration of AI technologies for future innovations.

Author's contribution

Study design – M. Laskowski, M. Ciekalski, B. Błaszczuk, M. Setlak, P. Paździora, A. Rudnik

Data collection – M. Laskowski, M. Ciekalski

Data interpretation – M. Laskowski, M. Laskowski, B. Błaszczuk, M. Setlak, P. Paździora

Statistical analysis – M. Ciekalski

Manuscript preparation – M. Laskowski, M. Ciekalski

Literature research – M. Laskowski, M. Laskowski

REFERENCE

1. The Age of Artificial Intelligence: A brief history... Deloitte Malta, 01 Nov 2022 [online] <https://www2.deloitte.com/mt/en/pages/rpa-and-ai/articles/mt-age-of-ai-1-a-brief-history.html> [accessed on 21 October 2023].
2. Brockman G., Sutskever I., OpenAI. Introducing OpenAI. OpenAI, December 11, 2015 [online] <https://openai.com/blog/introducing-openai> [accessed on 21 October 2023].
3. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P. et al. Language models are few-shot learners. OpenAI, May 28, 2020 [online] <https://openai.com/research/language-models-are-few-shot-learners> [accessed on 21 October 2023].
4. Bhasker S., Bruce D., Lamb J., Stein G. Tackling healthcare's biggest burdens with generative AI. McKinsey & Company, July 10, 2023 [online] <https://www.mckinsey.com/industries/healthcare/our-insights/tackling-healthcares-biggest-burdens-with-generative-ai> [accessed on 21 October 2023].
5. KMS Staff. Harnessing The Benefits of OpenAI in Healthcare. KMS Healthcare, June 29, 2023 [online] <https://kms-healthcare.com/benefits-openai-healthcare/> [accessed on 21 October 2023].
6. El-Hajj V.G., Gharios M., Edström E., Elmi-Terander A. Artificial intelligence in neurosurgery: A bibliometric analysis. *World Neurosurg.* 2023; 171: 152–158.e4, doi: 10.1016/j.wneu.2022.12.087.
7. Danilov G.V., Shifrin M.A., Kotik K.V., Ishankulov T.A., Orlov Y.N., Kulikov A.S. et al. Artificial intelligence in neurosurgery: A systematic review using topic modeling. Part I: Major research areas. *Sovrem. Tekhnologii Med.* 2021; 12(5): 106–112, doi: 10.17691/stm2020.12.5.12.
8. Ali R., Tang O.Y., Connolly I.D., Zadnik Sullivan P.L., Shin J.H., Fridley J.S. et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery* 2023; 93(6): 1353–1365, doi: 10.1227/neu.0000000000002632.
9. Hopkins B.S., Nguyen V.N., Dallas J., Texakalidis P., Yang M., Renn A. et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. *J. Neurosurg.* 2023; 139(3): 904–911, doi: 10.3171/2023.2.JNS23419.
10. Seghier M.L. ChatGPT: not all languages are equal. *Nature* 2023; 615(7951): 216, doi: 10.1038/d41586-023-00680-3.