



Can AI pass the written European Board Examination in Neurological Surgery? - Ethical and practical issues[☆]

Felix C. Stengel^a, Martin N. Stienen^a, Marcel Ivanov^b, María L. Gandía-González^c, Giovanni Raffa^d, Mario Ganau^e, Peter Whitfield^f, Stefan Motov^{a,*}

^a Department of Neurosurgery & Spine Center of Eastern Switzerland, Kantonsspital St. Gallen & Medical School of St.Gallen, St. Gallen, Switzerland

^b Royal Hallamshire Hospital, Sheffield, United Kingdom

^c Hospital Universitario La Paz, Madrid, Spain

^d Division of Neurosurgery, BIOMORF Department, University of Messina, Messina, Italy

^e Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom

^f South West Neurosurgery Centre, Plymouth, United Kingdom

ARTICLE INFO

Handling Editor: Dr W Peul

Keywords:

Neurosurgery board examination

Artificial intelligence

Chat gpt

Bing

Bard

EANS

Board-certification

ABSTRACT

Introduction: Artificial intelligence (AI) based large language models (LLM) contain enormous potential in education and training. Recent publications demonstrated that they are able to outperform participants in written medical exams.

Research question: We aimed to explore the accuracy of AI in the written part of the EANS board exam.

Material and methods: Eighty-six representative single best answer (SBA) questions, included at least ten times in prior EANS board exams, were selected by the current EANS board exam committee. The questions' content was classified as 75 text-based (TB) and 11 image-based (IB) and their structure as 50 interpretation-weighted, 30 theory-based and 6 true-or-false. Questions were tested with Chat GPT 3.5, Bing and Bard. The AI and participant results were statistically analyzed through ANOVA tests with Stata SE 15 (StataCorp, College Station, TX). P-values of <0.05 were considered as statistically significant.

Results: The Bard LLM achieved the highest accuracy with 62% correct questions overall and 69% excluding IB, outperforming human exam participants 59% ($p = 0.67$) and 59% ($p = 0.42$), respectively. All LLMs scored highest in theory-based questions, excluding IB questions (Chat-GPT: 79%; Bing: 83%; Bard: 86%) and significantly better than the human exam participants (60%; $p = 0.03$). AI could not answer any IB question correctly.

Discussion and conclusion: AI passed the written EANS board exam based on representative SBA questions and achieved results close to or even better than the human exam participants. Our results raise several ethical and practical implications, which may impact the current concept for the written EANS board exam.

1. Introduction

The EANS board examination was first introduced as an oral assessment in 1983 (Ljubljana). Since 1992, a more formal approach with a written and an oral part of the exam has been established and participants have since been granted the European Diploma in Neurosurgery. In October 2015, the EANS and UEMS Section of Neurosurgery transformed the existing exam into the European Board Examination in Neurological Surgery to further increase its importance and general

recognition. Residents who pass both parts of the exam are named nowadays Fellow of the European Board of Neurological Surgery (FEBNS) (EANS, 2023). In its current form, the exam represents a comprehensive assessment with a broad scope of applied knowledge based on information from basic science, cranial and spinal surgery. Part I of the exam consists of 100 single best answer (SBA) questions, which explore the candidates' ability to interpret case vignettes, read radiological, histopathological, and clinical images, and assess their theoretical and analytic knowledge (Stienen et al., 2016). The passing mark

[☆] A study of the EANS Young Neurosurgeons Committee (YNC), the Task Force for Emerging Technologies and Innovations in Neurosurgery (ETIN), the European Board of Neurological Surgery (EBNS) Exam Committee and the Ethico-Legal Committee.

* Corresponding author. Department of Neurosurgery & Spine Center of Eastern Switzerland, Kantonsspital St. Gallen & Medical School of St.Gallen, Rorschacher Str. 95, CH-9007, St. Gallen, Switzerland.

E-mail address: stefan.motov@kssg.ch (S. Motov).

<https://doi.org/10.1016/j.bas.2024.102765>

Received 27 November 2023; Received in revised form 28 January 2024; Accepted 12 February 2024

Available online 13 February 2024

2772-5294/© 2024 The Authors. Published by Elsevier B.V. on behalf of EUROSPINE, the Spine Society of Europe, EANS, the European Association of Neurosurgical Societies. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

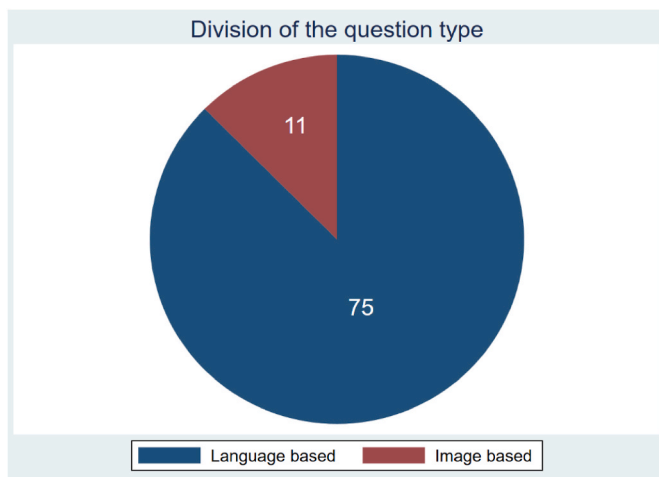


Fig. 1. Questions division based on content. The breakdown of the analyzed EANS questions by question type into language based (n = 75) and image based (n = 11) questions is shown in a pie chart.

is 60 % but it is usually defined and depends on the average performance of all residents. It is open for all residents in accredited neurosurgical programs in Europe, and to all neurosurgeons with a license to practice neurosurgery to register for the exam. Part II of the exam is the oral clinical problem-solving and patient management test, which requires submission of the logbook and confirmation of specialty training in neurological surgery (Stienen et al., 2020; Whitfield et al., 2023). It is not a theoretical examination, unlike the Part I examination, and was not assessed in this study.

Since the launch of ChatGPT 3 in 2020, several large language models (LLMs) based artificial intelligence (AI) algorithms have been developed and have become popular for commercial and scientific use. ChatGPT (OpenAI; San Francisco, CA), is a LLM chatbot that uses self-attention mechanisms and a large amount of training data to generate human-like text responses to prompt content created by users. It is extremely capable of handling long-range dependencies and forming coherent and contextually appropriate responses (Kung et al., 2023). Apart from ChatGPT, there are other commercially useable LLMs such as Bing (Microsoft Corporation, Redmond, WA) and Bard (Alphabet Inc., Mountain View, CA) and even more scientifically specialized models such as BioMedLM (Stanford) and BioGPT, which derive their information from research publications.

Recent literature demonstrated that AI is able not only to outperform participants in medical exams (Kung et al., 2023; Guerra et al., 2023) but also to design contemporary, cohesive, and valid SBA questions (E et al., 2023). Given the above, the authors, who are an expression of the past and current EANS Young Neurosurgeons Committee (M.L.G., S.M., F.S., M.S., G.R.), the ETIN Task Force (M.I.), the Exam Committee (M.I., M.G., P.W.) and the Ethico-Legal Committee (M.G.), joined forces to explore the accuracy of ChatGPT 3.5, Bing and Bard in the written part of the EANS board examination and its performance compared to the results of human examinees.

2. Methods

According to the St. Gallen Ethics Committee, no ethics approval was required for this study. We included 86 representative SBA questions with five possible answers from general clinical and scientific fields in neurosurgery, which have been chosen by the current EANS board exam committee between 07/2023 and 08/2023. An important criterion for the selection of questions was that all questions had appeared at least ten times in previous EANS Board exams.

Questions were subdivided according to the content in two categories – text-based (n = 75; 87%) and image-based (IB; n = 11; 13%)

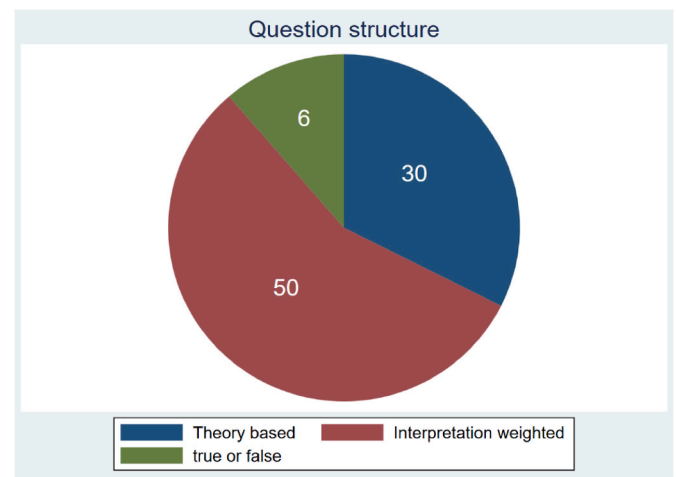


Fig. 2. Questions division based on structure. The breakdown of the analyzed EANS questions by question structure into theory based (n = 30), interpretation based (n = 50), and true or false (n = 6) questions is shown in a pie chart.

(Fig. 1). A further classification was made based on the formal structure of the questions into three categories – theory-based meaning questions aiming for an exact data-based e.g. anatomical structures, neurophysiological parameters, etc. (n = 30; 35%), interpretation-weighted questions, which are based on the analysis of information and e.g. deriving a diagnosis or therapy solution (n = 50; 58 %) and questions with a true/false or correct/incorrect answer structure (n = 6; 7%) (Fig. 2).

All questions were tested with the three freely available LLM algorithms - ChatGPT 3.5, Bing and Bard. The AI and the average participants' results were statistically analyzed with descriptive and comparative analysis through one-way analysis of variance (ANOVA) tests based on StataSE 15.1 (StataCorp, College Station, TX). Furthermore, we analyzed the accuracy of LLMs in different categories based on this differentiation. The analysis was carried out including and excluding IB questions. P-values of <0.05 were considered as statistically significant.

3. Results

In general, the LLMs scored similarly to the participants regarding all questions (participants mean correct 59%; between the groups: F (3/340) = 0.51, p = 0.68) with Bard being the most accurate LLM, even surpassing the participants' results (mean correct 62%; Fig. 3), followed by Chat GPT (mean correct 58%) and Bing (mean correct 53%). When IB questions were excluded, all LLMs scored higher (mean correct Bard 69%, ChatGPT 65%, Bing 60 %) than the participants (mean correct 59 %), however without reaching statistical significance (between the groups F (3/296) = 0.95, p = 0.42; Fig. 4).

In theory-based questions, the mean correct LLM scoring (mean correct Bard 87%, Bing 83%, ChatGPT 80%) exceeded the human participants' scoring (mean correct 60%, F (3/116) = 3.52, p = 0.0173, with significant difference of the mean correct between Bard and the participants: p = 0.023; Fig. 5). Further exclusion of IB questions in this category provided a statistically meaningful difference with LLMs outscoring the residents (F (3/112) = 3.28, p = 0.0237, with a significant difference of the mean correct between Bard and the participants: p = 0.03; Fig. 6).

Based on interpretation-weighted questions, the LLMs showed a strong tendency to achieve lower results (mean correct Bard 46%, ChatGPT 44%, Bing 36%) than the human participants not reaching statistical difference (mean correct 59%; F (3/196) = 2.28, p = 0.0809; Fig. 7). After exclusion of IB questions, the LLMs again scored close (mean correct Bard 58%, ChatGPT 55%, Bing 45%) to the human participants' results (mean correct 59%; F (3/156) = 0.82, p = 0.4864;

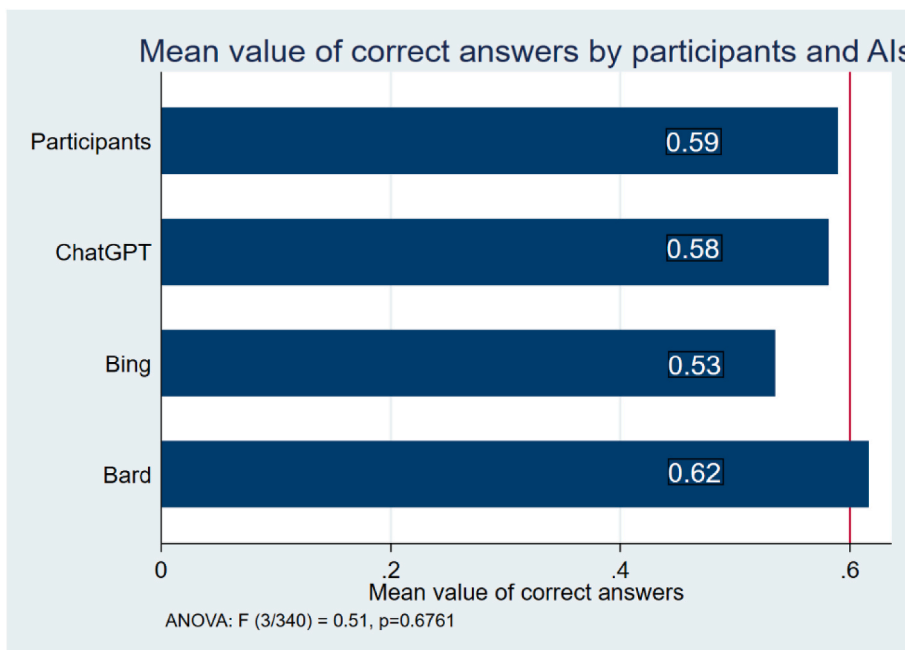


Fig. 3. Mean value of correct answers by participants and artificial intelligence (AI) in all categories. The bar chart shows the average correct answer rate for all questions (participants 59%, ChatGPT 58%, Bing 53%, Bard 62%). The ANOVA test showed no significant difference ($F(3/340) = 0.51, p = 0.676$). The red vertical line shows the approximate pass mark of 60% correct answers on the exam.

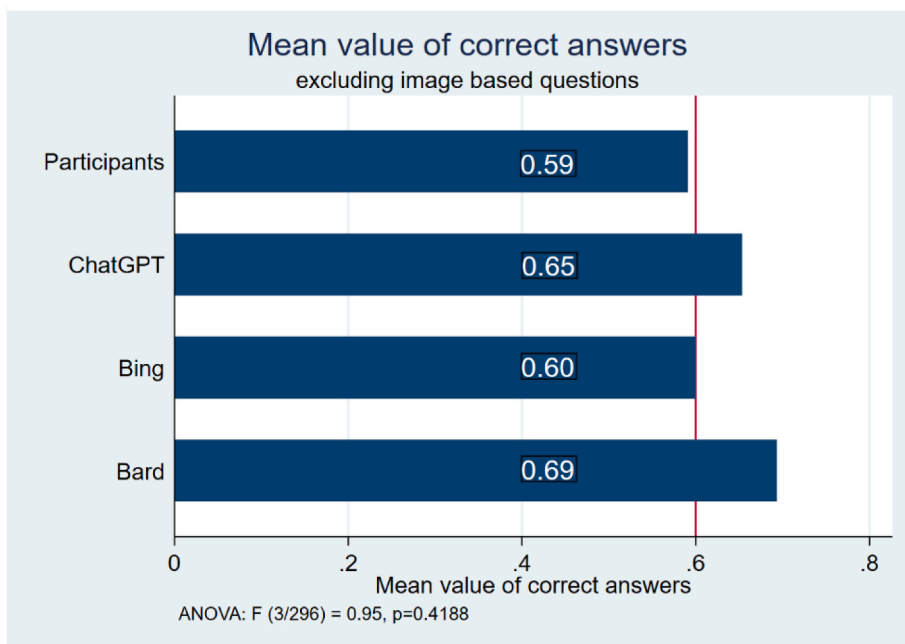


Fig. 4. Mean value of correct answers by participants and artificial intelligence (AI) in all categories excluding image-based questions. The bar chart shows the average correct answer rate for all questions (participants 59%, ChatGPT 65%, Bing 60%, Bard 69%). The ANOVA test showed no significant difference ($F(3/296) = 0.95, p = 0.4188$). The red vertical line shows the approximate pass mark of 60% correct answers on the exam.

Fig. 8).

The questions were also categorized according to their structure into true or false questions. Questions in which either a correct or incorrect answer option was sought. The analysis of a difference in the correct response rate concerning this task structure excluding IB questions revealed no significant difference between the groups ($F(3/20) = 0.20, p = 0.90$).

4. Discussion

For decades, the EANS board exam has been an educational standard for European board-certified neurosurgeons. However, AI encompasses the power to revolutionize the current concept of neurosurgical training and examination. We demonstrated corresponding to previous publications (Kung et al., 2023; Guerra et al., 2023; E et al., 2023; Johnson et al., 2023; Mannam et al., 2023), that all three commercial LLMs can achieve similar results as human participants, or even surpass their

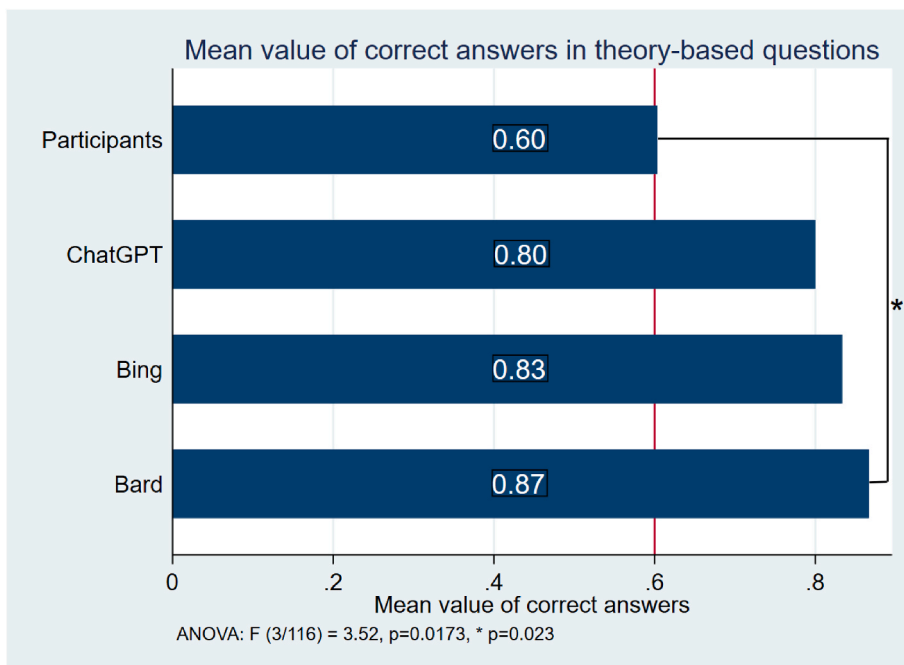


Fig. 5. Mean value of correct answers in all theory-based questions by participants and artificial intelligence (AI). The bar chart shows the average correct answer rate for all questions (participants 60%, ChatGPT 80%, Bing 83%, Bard 87%). The ANOVA test showed a significant difference within the groups between the participants and Bard (marked with an asterisk: $p = 0.023$; $F (3/116) = 3.52$, $p = 0.0173$). The red vertical line indicates the approximate pass mark of 60% correct answers on the exam.

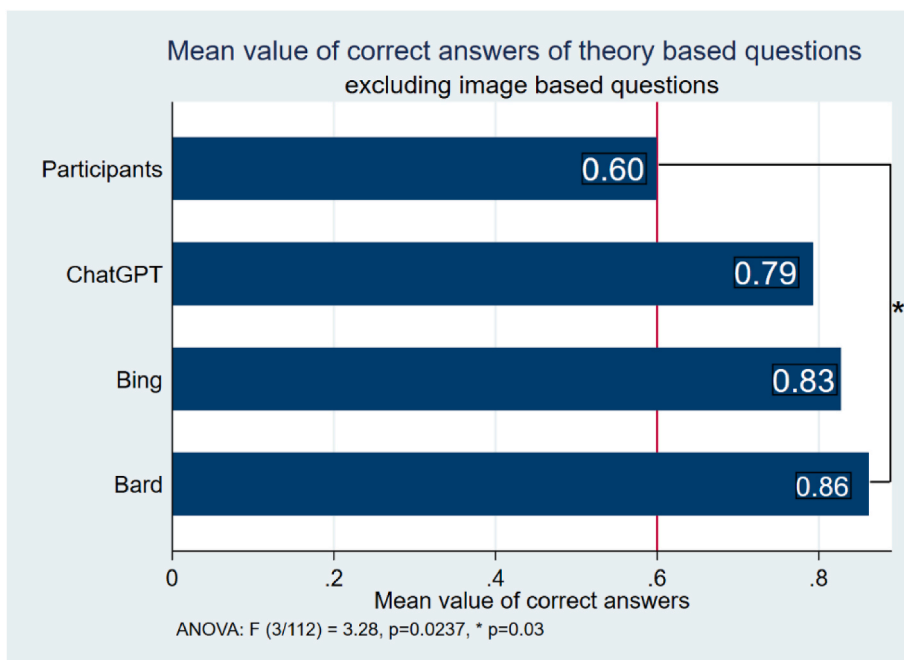


Fig. 6. Mean value of correct answers of theory-based questions excluding image-based by participants and artificial intelligence (AI). The bar chart shows the average correct response rate for all questions (participants 60%, ChatGPT 79%, Bing 83%, Bard 86%). The ANOVA test showed a significant difference within the groups between the participants and Bard (marked with an asterisk: $p = 0.03$; $F (3/112) = 3.28$, $p = 0.0237$). The red vertical line indicates the approximate pass mark of 60% correct answers on the exam.

performance in certain domains. Even without access to PubMed-indexed journals and recent medical publications, the AI software was able to analyze and correctly answer most questions. The highest scores and accuracy were as expected achieved in theory-based questions, as LLMs are assumed to better recall exact details than human beings. The lowest scores and accuracy were unsurprisingly obtained in

image-based and interpretation-weighted questions. There is an easy explanation for this result, since the AI algorithms, that we assessed, are LLMs and still unable to interpret image data. However, there are also limits to text interpretation for LLMs. The accuracy of the LLMs' answers increases with the amount of information given in the questions. The accuracy of the answers also depends on the database used to train the

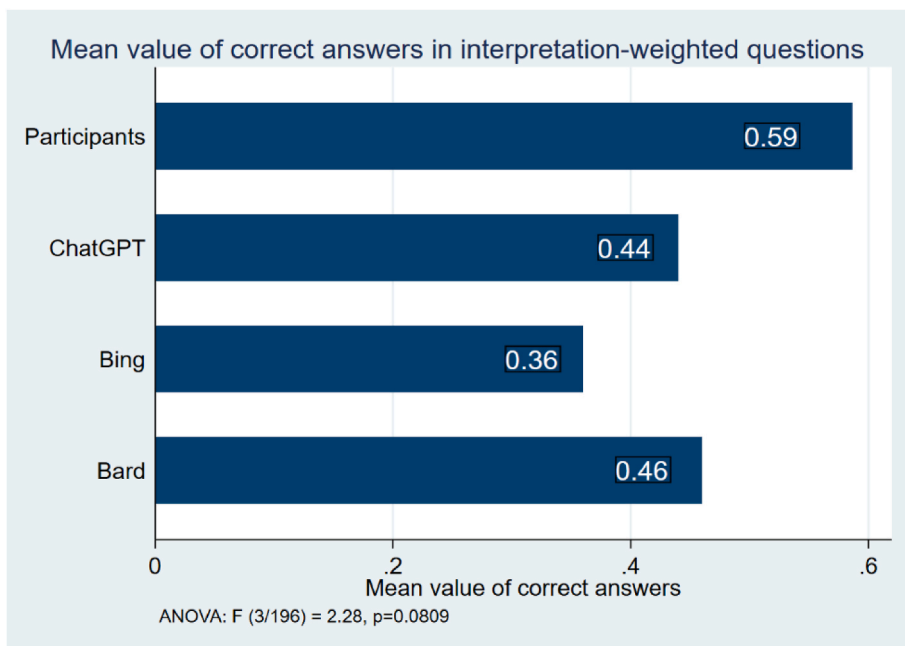


Fig. 7. Mean value of correct answers in all interpretation-weighted questions by participants and artificial intelligence (AI). The bar chart shows the average correct answer rate for all questions (participants 59%, ChatGPT 44%, Bing 36%, Bard 46%). There was no significant difference in the ANOVA test ($F(3/196) = 2.28, p = 0.0809$).

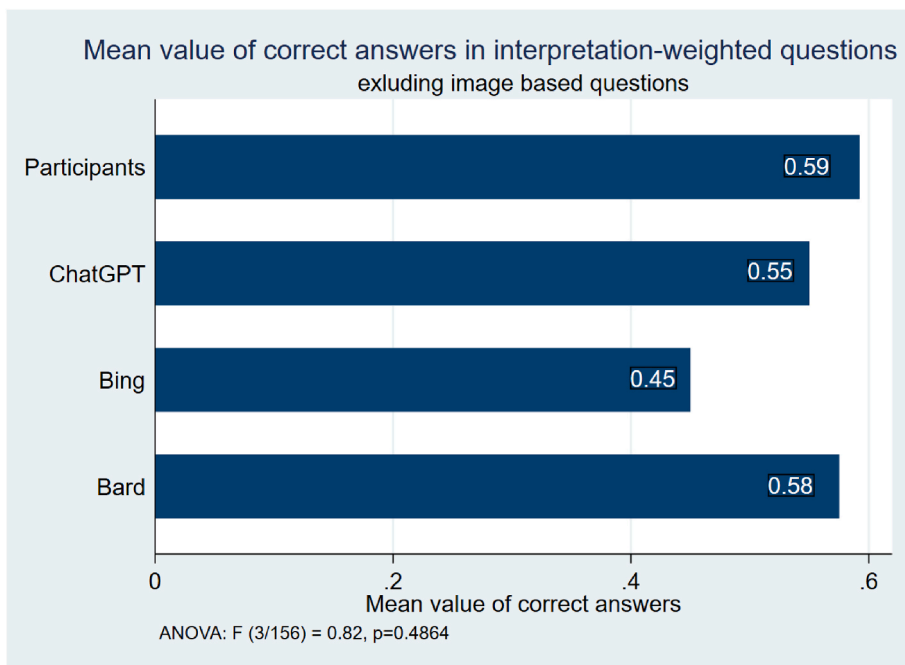


Fig. 8. Mean value of correct answers in all interpretation-weighted questions excluding image-based questions by participants and artificial intelligence (AI). The bar chart shows the average correct answer rate for all questions (participants 59%, ChatGPT 55%, Bing 45%, Bard 58%). There was no significant difference in the ANOVA test ($F(3/156) = 0.82, p = 0.4864$).

model. For questions based on an interpretation structure, this means a certain dependence on the information provided. The complexity of questions and the SBA type might even present a certain limitation for current AI models as they are designed with conversational interfaces more capable of providing context and content to open questions (Johnson et al., 2023; Gilson et al., 2023). A recent publication (Saad et al., 2023) reinforces our theory that general AI might still underperform compared to human examinees when asked to answer questions

requiring higher-order critical thinking. Further development and combination of language-based and image-based models is a matter of time, and the continuous updates of AI will surely enable the LLM to become even more accurate in the areas where they currently do not perform optimally.

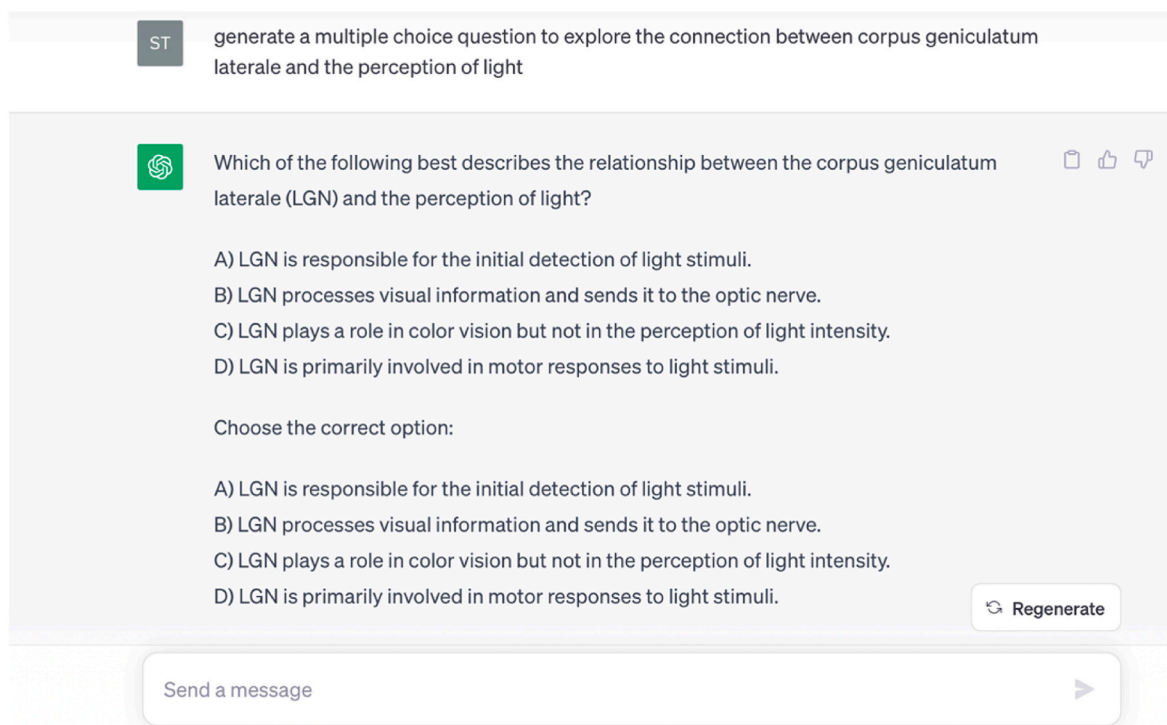


Fig. 9. Example for a SBA question generated with AI (Chat GPT 3.5, OpenAI) - The generated question does not yet meet the threshold for inclusion in the board exam due to poor quality and lack of a clinical vignette as well as unclear sort of relationship (e.g. anatomical, clinical, physiological), which makes it in this form unsuitable for use. All new SBA questions need to describe a scenario – usually clinical and to assess the applied knowledge of the candidates. However, it demonstrates the creativity and reliability of AI to generate SBA questions based on scientific content.

4.1. Ethical and practical implications

The results of this study imply two main questions for further debate – must the current concept of written board exam be altered, and what shall be the role of AI in the further development of the exam?

We demonstrated that in its current form, the part I written exam does not necessarily require human or natural intelligence, and might be accomplished with the help of AI, which reveals the need for certain changes. Although cheating might be a possible scenario, which would provide more residents with the possibility to apply for part II of the exam, the oral exam is performed in person and the examination supervisors would prohibit the use of LLMs during its conduction. Also, trust in medical specialists should be maintained not only throughout educational events and training courses but also during exams despite the rise of modern technological advances.

Another ethical dilemma that will have to be considered is the potential impact of AI on clinical practice. Having a tool that is as good or even better than “end-of-training neurosurgeon” on the one hand may lead the medical professionals to exceedingly rely on such technology in their clinical decisions and on the other hand may demotivate young neurosurgeons in their educational growth which naturally implies a lot of effort to memorize and analyze a huge quantity of information during their life-long process of continuous education.

However, LLMs also might provide false or outdated information and mislead even the specialists, if blindly relied on (Liu et al., 2023). A thorough and regular review of the sources of information and sequential validation of the provided output by healthcare professionals is essential at this stage of development (Ali et al., 2023; Sorin et al., 2023).

Access to AI technology is currently available to anyone – both medical professionals and the general public (patients, lawyers, etc). Potential differences of opinion between humans and artificial intelligence (which is shown to be at least as good at passing the theoretical exam) will have to be considered. There is a certain need for regulations

and security measures especially concerning ethical and legal responsibility. For example, in case of false information delivered by the LLM during the teaching process even in the form of an incorrect exam question, or potentially during a clinical decision, it must be defined who might be held responsible. Also, LLM might become targets for adversarial cyber-attacks (Finlayson et al., 2019), thus compromising confidential educational and residents’ information is not without risk. Robust security measures and continuous updates are mandatory for further integration in the teaching and exam process but the utilization of AI LLMs in the EANS board exam seems to be inevitable. Research in the near future might include the use of ChatGPT for generating and proofing new questions, including assistance during the various steps of syntax, semantics, and scientific accuracy verification.

4.2. Future perspectives on using AI for neurosurgical education

There is an increased need for neurosurgical specialists worldwide and AI might help to reduce the costs and effort for training and examination (Stienen et al., 2017), (Stengel et al., 2022). Since LLMs are powerful algorithms, that have the potential to quickly access specific information, to learn based on provided sources, and to develop decision-making abilities, they might be exceptional learning tools for residents preparing to take the written exam. LLMs not only answer SBA questions accurately but also deliver a very proficient explanation. Irrespective of this, the basic specialist knowledge required to interpret and categorize whether the answer given by the LLM is conclusive is indispensable.

Further development and combination with, e.g., the EANS educational platforms such as the EANS Academy might be interesting soon since LLMs might categorize, explain, and deliver the content in a highly structured and efficient way, saving time and energy, which one would invest to search for particular content. Recent data provided insights that AI is not only able to reproduce and reorganize data accurately but to support clinical decision-making (Liu et al., 2023; Ben-Shabat et al.,



ChatGPT

Case Vignette:

A 35-year-old male was involved in a motor vehicle accident. He was not wearing a seatbelt and was ejected from the vehicle. Emergency medical services found him unconscious at the scene. On arrival at the hospital, the patient is unresponsive with a Glasgow Coma Scale (GCS) of 8 (E2V2M4). Initial imaging reveals a left-hemispheric epidural hematoma with midline shift.

Multiple Choice Question:

What is the most appropriate surgical approach for managing the left-hemispheric epidural hematoma in this patient with a Glasgow Coma Scale (GCS) of 8?

- A) Burr hole evacuation
- B) Craniotomy with evacuation of hematoma and control of bleeding
- C) Conservative management with close observation

Correct Answer: B) Craniotomy with evacuation of hematoma and control of bleeding

Explanation:

In the case of a left-hemispheric epidural hematoma with midline shift and neurological deterioration (as evidenced by a low GCS), the most appropriate surgical approach is a craniotomy. This procedure involves the surgical opening of the skull to evacuate the hematoma, relieve pressure on the brain, and control any bleeding. Burr hole evacuation and stereotactic aspiration are less invasive procedures and may be considered in certain cases; however, a craniotomy allows for better access to the hematoma and thorough control of bleeding. Conservative management or administration of anticoagulants would not be suitable in a situation where urgent surgical intervention is necessary to prevent further neurological deterioration and potential mortality.

Fig. 10. Example of a case vignette and SBA question generated with AI (Chat GPT 3.5, OpenAI) – although not a comprehensive question and clinical case vignette, this example represents very well the capacity of AI and gives us the possibility for further developments.

2021) and even to design new SBA questions and case vignettes (E et al., 2023) (Fig. 9 and 10). However, we have significant concerns that AI-generated questions may focus on factual recall rather than applied knowledge. Designing clinical case vignettes and SBA questions might be challenging and occasionally affords a panel of specialists who certainly need a lot more time to create unique exam questions. AI might not only be utilized for automated scoring of student papers but also as a provider for new exercises and exams (Guo et al., 2023). Additionally, it can individually facilitate the learning process, based on the residents' needs and enable a more tailored educational approach (Zoia et al., 2022) (see Fig. 10).

4.3. Strengths

We explored in this study the capability of three commercial LLMs to pass the EANS board exam. We not only demonstrated that all three LLMs performed highly and were able to pass the current written part of the EANS board exam but also that they were even better in certain categories than the neurosurgical residents. This study showcases the capacity of already existing AI models, which might be successfully integrated into the further development of the EANS board exam.

4.4. Limitations

A major limitation of this study is its design as we explored only the capability of AI LLMs on pre-selected questions from prior EANS board exams retrospectively. A prospective study might be of importance regarding the assessment of exam questions based on AI before letting the residents take the exam. We also have not explored the possibility of taking the exam questions several times, which theoretically might improve the general scores of the AI models. AI models receive a regular update through the training on inputs of the users and it is reasonable to hypothesize that each new iteration might cause an increase in performance (Johnson et al., 2023). Moreover, we performed an analysis based on only 86 representative questions, which did not cover each aspect or topic in the broad field of neurosurgery. Further subanalysis based on a higher number of questions from various clinical and scientific backgrounds might be interesting to define subcategories, where AI out or still underperforms compared to the human participants. A further investigation that was not carried out concerns the formulation and semantics of the question, which could influence the precision of the answers given by LLMs. Also more recent LLMs updates e.g. Chat GPT 4 might even achieve higher scores in exams than the prior versions.

5. Conclusions

AI passed part I of the EANS board exam in representative, selected SBA questions, and achieved results close to or even better than the human exam participants with higher accuracy in non-IB questions. It appears that AI is most accurate in theory-based questions with less demand for interpretation. These results raise several ethical and practical implications, which may impact the current concept for the written EANS board exam.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Ali, K., et al., 2023. ChatGPT-A Double-Edged Sword for Healthcare Education? Implications for Assessments of Dental Students. *Eur J Dent Educ*.
- Ben-Shabat, N., et al., 2021. Assessing the performance of a new artificial intelligence-driven diagnostic support tool using medical board exam simulations: clinical vignette study. *JMIR Med Inform* 9 (11), e32507.
- E, K., et al., 2023. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Med. Educ.* 23 (1), 772.
- EANS, 2023. *EANS board examination webpage*. Available from: <https://www.eans.org/page/Exams>.
- Finlayson, S.G., et al., 2019. Adversarial attacks on medical machine learning. *Science* 363 (6433), 1287–1289.
- Gilson, A., et al., 2023. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 9, e45312.
- Guerra, G.A., et al., 2023. GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and Neurosurgery Residents on Neurosurgery Written Board-like Questions. *World Neurosurg.*
- Guo, A.A., Li, J., 2023. Harnessing the power of ChatGPT in medical education. *Med. Teach.* 45 (9), 1063.
- Johnson, D., et al., 2023. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: an Evaluation of the Chat-GPT Model. *Res Sq*.
- Kung, T.H., et al., 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2 (2), e0000198.
- Liu, S., et al., 2023. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J. Am. Med. Inf. Assoc.* 30 (7), 1237–1245.
- Mannam, S.S., et al., 2023. Large Language Model-Based Neurosurgical Evaluation Matrix: A Novel Scoring Criteria to Assess the Efficacy of ChatGPT as an Educational Tool for Neurosurgery Board Preparation. *World Neurosurg.*
- Saad, A., et al., 2023. Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. *Surgeon* 21 (5), 263–266.
- Sorin, V., et al., 2023. Large language models for oncological applications. *J. Cancer Res. Clin. Oncol.* 149 (11), 9505–9508.
- Stengel, F.C., et al., 2022. Transformation of neurosurgical training from "see one, do one, teach one" to AR/VR & simulation - a survey by the EANS Young Neurosurgeons. *Brain Spine* 2, 100929.
- Stienen, M.N., et al., 2016. Residency program trainee-satisfaction correlate with results of the European board examination in neurosurgery. *Acta Neurochir.* 158 (10), 1823–1830.
- Stienen, M.N., et al., 2017. eLearning resources to supplement postgraduate neurosurgery training. *Acta Neurochir.* 159 (2), 325–337.
- Stienen, M.N., et al., 2020. Procedures performed during neurosurgery residency in Europe. *Acta Neurochir.* 162 (10), 2303–2311.
- Whitfield, P.C., et al., 2023. European training requirements in neurological surgery: a new outcomes-based 3 stage UEMS curriculum. *Brain Spine* 3, 101744.
- Zoia, C., et al., 2022. The EANS young neurosurgeons committee's vision of the future of European neurosurgery. *J. Neurosurg. Sci.* 66 (6), 473–475.