Impact Surgery

Original Research

e-ISSN: 3033-4470

Performance of OpenAI's GPT-4 in a mock MRCS Part A Examination

Ibrahim Inzarul Haq[1], Siddarth Raj[1], Ali Ridha[1,2], Arun O'Sullivan[1], Imran Ahmed[1,2], Farhan Syed[1], Chetan Khatri[1,2]

Correspondence: Ibrahim Inzarul Haq, Trauma and Orthopaedics Department, University Hospital Coventry and Warwickshire, Clifford Bridge Road, Coventry, United Kingdom, CV2 2DX. ibrahimhaq@hotmail.com

Abstract

Introduction: OpenAI's latest iteration of a Large Language Model (LLM); GPT-4 (Generative Pre-trained Transformer 4) has demonstrated its proficiency against various professional examination standards like the USMLE (United States Medical Licensing Examination), FRCS (Fellowship of the Royal Colleges of Surgeons) and the United States Bar. However, GPT-4's capability with the MRCS (Membership of the Royal College of Surgeons) Part A has not yet been investigated.

Methodology: A representative MRCS Part A examination that was prepared and provided by "TeachMeSurgery" based on the MRCS Intercollegiate Curriculum was used to assess GPT-4's performance. Each question was processed via the web-based interface of ChatGPT (Chat Generative Pre-trained Transformer) Plus.

Results: GPT-4 scored 87.2% on Applied Basic Sciences (157/180) and 86.7% on Principles of Surgery in General (104/120 questions), achieving an overall score of 261/300 (87%), which is above the typical passing threshold. GPT-4 scored 100% in four out of the eleven predefined curriculum areas, which included: Pharmacology, Microbiology, Data Interpretation and Audit, and The Surgical Care of Children. GPT-4's weakest performance was in the Medico-Legal Aspects of Surgical Practice, in which it scored 33.3%.

Conclusion: GPT-4 successfully passed the mock MRCS Part A without any specialised preparatory training. Further research could look at integrating the GPT-4 model to enhance a trainee surgeon's examination preparation and its wider use in surgical training.

1. University Hospitals Coventry & Warwickshire NHS Trust, Clifford Bridge Road, Coventry, United Kingdom, CV2 2DX.

2. Warwick Clinical Trials Unit, Clinical Sciences and Research Laboratories, University Hospitals Coventry & Warwickshire NHS Trust, Clifford Bridge Road, Coventry, United Kingdom. CV2 2DX

Cite as: Haq, I., Raj, S., Ridha, A., O'Sullivan, A., Ahmed, I., Syed, F., & Khatri, C. (2024). Performance of OpenAI's GPT-4 in a mock MRCS Part A Examination. *Impact Surgery*, 1(5), 172–176. https://doi.org/10.62463/surgery.89

Introduction

Generative Pre-trained Transformer 4 (GPT-4) is a Large Language Model (LLM) released in March 2023 by OpenAI. LLMs are artificial intelligence (AI) models capable of higher order reasoning and generating human-like outputs [1]. Compared to traditional deep learning AI models, LLMs produce more coherent and relevant content due to their use of natural language processing[2]. The web-based interface of ChatGPT (a free-to-use AI system) has allowed public access to GPT-4 without the need for technical knowledge and has caused a rapid exploration into potential applications of AI in clinical medicine and medical education. For instance, AI has exhibited a rudimentary capability to analyse chest radiographs[3].

In medical education, AI can be used to assist medical students with history-taking with an AI bot posing as an imaginary patient [4]. This has been shown to be a good alternative to bedside teaching, especially within the early clinical years. However, the use of AI has not yet been adopted into mainstream medical education, despite the increasing number of research papers highlighting its potential[5]. This is partly due to a lack of high-level research demonstrating the accuracy and efficacy of AI in medical education [5].

Perhaps understandably, the adoption of AI into medical practice and education has been met with apprehension by many members of the medical community. Concerns regarding patient privacy and bias within source data have been raised[6]. The management of sensitive patient data causes concerns about potential breaches, especially since this data may be processed remotely. Additionally, AI systems may exhibit diagnostic inaccuracies because they are trained on extensive databases that include unverified or incorrect medical information.

In addition, a nationwide survey of medical students yielded concerns that AI could damage doctor-patient relationships, devalue the medical profession and potentially lead to unemployment amongst doctors [7]. However, the majority of those surveyed agreed that there was potential for AI to improve clinicians' access to information and patients' access to healthcare services, and to reduce errors. 93.8% of those surveyed thought they should be given structured training on AI applications in healthcare settings [7].

GPT-4 has demonstrated its proficiency against established professional examination standards like the United States Medical Licensing Examination (USMLE), the Fellow of the Royal College of Surgeons Examination (FRCS) and the United States Bar [8–10]. These evaluations have yielded mixed success rates, underscoring the dynamic nature of AI's performance. Nevertheless, an unexplored avenue remains: the application of GPT-4 in the context of the Membership of the Royal College of Surgeons (MRCS) Part A examination, which is a mandatory, written exam for postgraduate doctors within the UK healthcare system who wish to enter specialist surgical training. It has a pass rate of 30 to 40% and the pass mark varies from 69 to 75%[11]. In contrast to the FRCS, the MRCS Part A is a broader, less specialised examination designed for doctors at an earlier stage of their surgical training. Given that GPT-4's inability to pass FRCS was in part due to its limitations in critical thinking, surgical principles and decision-making skills, it is possible GPT-4 would be more likely to pass the MRCS Part A.

Exploring GPT-4's ability to pass the MRCS Part A can highlight LLMs potential in examination scenarios and evaluate its limitations. The MRCS Part A is a particularly suitable benchmark compared to other examinations because it assesses a broad range of medical knowledge at an earlier stage of professional development, focusing on fundamental surgical principles and basic clinical skills. This study aimed to determine whether GPT-4 can pass the MRCS Part A and evaluate whether it has potential for further use in medical education and examination settings.

Methods

The MRCS Part A examination is structured into two distinct sections: Applied Basic Sciences (ABS) with 180 questions and Principles of Surgery in General (POSG) with 120 questions, resulting in a comprehensive total of 300 questions. The overall examination adheres to a predefined curriculum breakdown published by the Intercollegiate Committee for Basic Surgical Examinations (ICBSE), focusing heavily on Applied Surgical Anatomy (75 questions) but also includes a range of other topics, such as Applied Surgical Physiology (45 questions) and Common Surgical Conditions (45 questions) [11]. Candidates need to attain a pass mark in both papers to successfully clear the examination.

A representative MRCS Part A examination was provided by TeachMeSurgery in an excel spreadsheet format [12]. It was chosen for this study, for having a comprehensive multiple choice question bank and its focus on surgical topics. The representative examination was based on the MRCS curriculum published by the ICBSE and was reviewed by senior authors at TeachMeSurgery. Each question was written as a Single Best Answer (SBA) with a clinical vignette and four potential answers.

Each question was processed via the web-based interface of ChatGPT Plus, which is the most advanced LLM available at the time of this study and has been used to benchmark the capabilities of AI in the other papers [1]. ChatGPT Plus costs USD $20 per month and utilizes GPT-4 'Advanced Data Analysis', which allows

for documents and spreadsheets to be uploaded and analysed by ChatGPT. We uploaded the mock exam with the four potential answers in an Excel spreadsheet. We then proceeded to prompt ChatGPT to answer each question iteratively.

GPT-4 does not have direct access to the internet and has been trained on a large database until September 2021 at the time of this study.

Results

GPT-4 achieved a score of 87% (261/300) in the representative paper provided. The score was consistent in most subsections of the exam. In the Applied Basic Sciences paper, the score was 87.2% (157/180) and in the Principles of Surgery in General the score was 86.7% (104/120). GPT-4 scored 100% in four out of the eleven predefined curriculum areas, which included: Pharmacology, Microbiology, Data Interpretation and Audit, and The Surgical Care of Children. GPT-4's weakest performance was in the Medico-Legal Aspects of Surgical Practice, in which it scored 33.3%. detailed scores are presented in Table 1 and 2.

Table 1: Breakdown of GPT-4's results on the Membership of the Royal College of Surgeons (MRCS) Part A examination for Applied Basic Sciences

Applied Basic Sciences Total	87.2% (157/180)
Applied Surgical Anatomy	82.5% (99/120)
Applied Surgical Pathology	91.7% (33/36)
Pharmacology	100.0% (9/9)
Microbiology	100.0% (7/7)
Imaging	80.0% (4/5)
Data Interpretation & Audit	100.0% (5/5)

Table 2: Breakdown of GPT-4's results on the Membership of the Royal College of Surgeons (MRCS) Part A examination for Principles of Surgery in General Total

Principles of Surgery in General Total	86.7% (104/120)
Common Congenital and Acquired Surgical Conditions	86.7% (39/45)
Pre-op Management	97.1% (34/35)
Assessment & Management of Trauma	76.7% (23/30)
Surgical Care of Children	100.0% (7/7)
Medico-Legal Aspects of Surgical Practice	33.3% (1/3)

Development of the prompt

A meticulous prompt development process was implemented to ensure its success. The strategy involved crafting the simplest possible prompt without providing any additional information to GPT-4. We conveyed to GPT-4 that we had uploaded a dataset consisting of MRCS Part A examination questions in an Excel spreadsheet and requested it to provide the correct answer out of four potential choices. This initial prompt revealed several challenges that the authors promptly addressed.

Subsequent prompts made it explicit that the questions required a selection of the single best answer, necessitating GPT-4 to pick the most appropriate response. Challenges related to batching questions were identified, where GPT-4 tended to oversimplify, modify data, or randomly guess the correct answer, often selecting 'A' without providing reasoning. Although GPT-4 in our preliminary testing always provided an answer we made it explicit that GPT-4 is to answer each question. This is because the MRCS Part A exam doesn't employ negative marking, meaning there are no penalties for incorrect answers. To counteract these issues, we instructed GPT-4 not to modify any data and conducted a one-by-one assessment of each question for accuracy.

It's important to note that no updates were made to GPT-4 during the testing phase, and distinct chat environments were employed to evaluate GPT-4 independently of prior information. Each section, ABS, and PSOG, was evaluated separately, with recorded correct and incorrect answers, yielding percentage scores out of 100 for both individual sections and the overall score.

The specific prompt, which gained consensus among all authors, is outlined in Appendix 1, providing details of the prompt and GPT-4's subsequent responses. This format enabled the authors to iterate through each question. Four sample outputs by GPT-4 can be found in Appendix 2.

Discussion

This study showed that GPT 4 is able to pass the MRCS A, a UK postgraduate surgical exam. It was found that GPT-4 is less successful with answering questions regarding UK guidelines and with the management of conditions as shown in Appendix 2. GPT-4 is however very proficient with factual questions especially regarding anatomy, pharmacology and data analysis. This is likely because there is only one correct answer in the selection and there is no need for further clinical reasoning to choose the most appropriate answer.

This contrasts with the work by Saad et al.[9], who demonstrated that GPT-4 lacked the clinical expertise to pass the FRCS Orthopaedic Part A examination. However, the FRCS requires a higher standard of knowledge compared to the MRCS examinations aimed

at professionals in the early years of their career. The FRCS (Ortho) is sat by surgeons with at least 10 years of clinical experience prior to becoming a consultant in the United Kingdom. It was noted by Saad et al.[9] that GPT-4 struggled with the SBA format. This could be due to the nature of SBA questions as multiple answers can technically be correct however the single best choice answer may require clinical experience, knowledge and interpretation of a scenario[13]. The "best" answer typically reflects the most appropriate choice given the clinical context, which can vary based on individual experience and judgment. Therefore, the ability to select the most suitable answer is closely tied to one's practical understanding and application of clinical principles.

In contrast to the MRCS, the questions in the FRCS examination exhibit increased complexity, featuring a greater number of distractors. Consequently, tackling FRCS questions necessitates a deeper reliance on both clinical expertise and theoretical knowledge.

It is important to note that GPT-4 has been trained on a large database but as far as we are aware has not been designed specifically for medical examinations or for the MRCS Part A in particular. However, future investigations could explore the potential benefits of training the model using a sample of mock exam questions or by using different prompts for the different subsections of the examination. Similar to human students, refining theoretical knowledge through exam practice is crucial, particularly for formats like SBA questions. With further appropriate training, GPT-4's performance in successfully tackling forthcoming exams could potentially be enhanced. GPT-4 could also be enhanced if it was able to access up-to-date clinical guidelines in the UK. However, at the time of writing, GPT-4 only has access to knowledge up till September 2021.

GPT-4 can pass the MRCS Part A and is also able to give the reasons for its answers. This could have implications in medical education such as helping students review questions with a explanation as to why they have gotten the question wrong. From this study the knowledge base of GPT-4 is accurate overall and can be relied upon mostly in pharmacology and microbiology as GPT-4 achieved 100%.

In the future, if GPT-4 acquires the ability to view or create anatomical images, it could greatly enhance surgical training. This advancement would be especially beneficial for trainees, as GPT-4 could simulate intra-operative anatomy and label complex images that are challenging for the untrained eye. It could also be valuable in laparoscopic training sessions.

Moreover, GPT-4 can support surgical education by simulating clinical scenarios, offering personalised feedback, and generating educational content. Additionally, it could aid in data interpretation and audits by analysing and collecting data, which could improve research outcomes and free up trainees to focus more on hands-on experience in the operating theatre.

One of the key strengths in this paper is that we carefully selected our prompts prior to full testing of the paper. Careful selection of a prompt is imperative for achieving meaningful output from GPT-4. A comprehensive evaluation of the prompt was conducted to ensure that GPT-4 was tested to its maximum potential. A significant limitation is that the official MRCS Part A exam includes five options for each question, however, the questions in this study only included four options, which could have made the exam easier for GPT-4. The representative examination used in this study was primarily text-based: none of the questions included required GPT-4 to interpret either images, such as prosections or radiographs or lab results, such as blood tests, which is unlike the official exam. This study only investigates GPT-4's performance on the MRCS Part A examination, however, candidates must also pass the Part B examination to be fully certified. The latter extensively tests communication and clinical skills, which cannot currently be assessed in the context of GPT-4 or other LLMs.

Future studies can also investigate the performance of different LLMs such as PALM2[14] in the context of the MRCS Part A examination. Thereafter, the ability of GPT-4 to generate questions to be used for revision can also be investigated, however the validity of these questions would need robust testing. GPT-4 can pass a representative MRCS Part A paper and can be used as a tool for medical education to help students understand the rationale behind questions. This paper highlights the strengths and weaknesses of LLMs in sitting clinical examinations and how it can be improved in future iterations.

Acknowledgements

The authors would like to express thanks to TeachMeSurgery for providing mock MRCS Part A questions free of charge. TeachMeSurgery was not involved in the design or conduct of the study.

Competing Interests

None of the authors have any competing interests to declare.

Declarations

Data is provided within the manuscript and supplementary information files

Author Contributions

Concept and design: Ibrahim Inzarul Haq, Siddarth Raj, Ali Ridha, Arun O'Sullivan, Imran Ahmed, Farhan Syed, Chetan Khatri; Acquisition, analysis, or interpretation of data: Ibrahim Inzarul Haq, Siddarth Raj, Ali Ridha; Drafting of the manuscript: Ibrahim Inzarul Haq, Siddarth Raj; Critical review of the

manuscript for important intellectual content: Ibrahim Inzarul Haq, Siddarth Raj, Ali Ridha, Arun O'Sullivan, Imran Ahmed, Chetan Khatri; Supervision: Farhan Syed, Chetan Khatri

References

1. Introducing ChatGPT Plus. https://openai.com/blog/chatgpt-plus (accessed Oct 5, 2023).

2. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023; published online April 13. DOI:10.48550/arXiv.2303.12712.

3. Akhter Y, Singh R, Vatsa M. AI-based radiodiagnosis using chest X-rays: A review. Front Big Data 2023; 6: 1120989.

4. Co M, John Yuen TH, Cheung HH. Using clinical history taking chatbot mobile app for clinical bedside teachings – A prospective case control study. Heliyon 2022; 8: e09751.

5. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. Am J Transl Res 2023; 15: 4820–8.

6. Cooper A, Rodman A. AI and Medical Education - A 21st-Century Pandora's Box. N Engl J Med 2023; 389: 385–7.

7. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ 2022; 22: 772.

8. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2: e0000198.

9. Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: A critical analysis. Surg J R Coll Surg Edinb Irel 2023; 21: 263–6.

10. 1Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 Passes the Bar Exam. 2023; published online March 15. DOI:10.2139/ssrn.4389233.

11. Intercollegiate Committee for Basic Surgical Examinations 2018/19 Annual Report [Internet]. Intercollegiate Committee for Basic Surgical Examinations. https://www.intercollegiatemrcsexams.org.uk/-/media/files/imrcs/policies-and-reports/icbse-annual-report-201819-final.docx.

12. TeachMeSurgery - Making Surgery Simple. TeachMeSurgery. https://teachmesurgery.com/ (accessed Oct 5, 2023).

13. Mirbahai L, W Adie J. Applying the utility index to review single best answer questions in medical education assessment. Arch Epidemiol Public Health 2020; 2. DOI:10.15761/AEPH.1000113.

14. Google AI PaLM 2. Google AI. https://ai.google/discover/palm2/ (accessed Oct 5, 2023).