

# Performance of GPT-4 in Membership of the Royal College of Paediatrics and Child Health-style examination questions

Richard Armitage 

**To cite:** Armitage R. Performance of GPT-4 in Membership of the Royal College of Paediatrics and Child Health-style examination questions. *BMJ Paediatrics Open* 2024;**8**:e002575. doi:10.1136/bmjpo-2024-002575

Received 7 February 2024  
Accepted 10 February 2024

## ABSTRACT

The large language model (LLM) ChatGPT has been shown to have considerable utility across medicine and healthcare. This paper aims to explore the capabilities of GPT-4 (Generative Pre-trained Transformer 4) in answering Membership of the Royal College of Paediatrics and Child Health (MRCPCH) written paper-style questions. GPT-4 was subjected to four publicly available sample papers designed for those preparing to sit MRCPCH theory components. The model received no specialised training or reinforcement. The average score across all four papers was 78.1%. The model provided reasoning for its answers despite this not being required by the questions. This performance strengthens the case for incorporating LLMs into supporting roles for practising clinicians and medical education in paediatrics.

## INTRODUCTION

ChatGPT is a large language model (LLM) that generates human-like text. Both the potential utility of and threats posed by ChatGPT have been recognised in medical education,<sup>1</sup> in proving antimicrobial prescribing advice,<sup>2</sup> and in writing discharge summaries,<sup>3</sup> clinic letters<sup>4</sup> and simplified radiology reports,<sup>5</sup> while the ethical concerns regarding their use in healthcare have also been highlighted. This paper aims to explore the capabilities of the publicly available GPT-4 (Generative Pre-trained Transformer 4) (as of 11 December 2023) in answering questions in the style of Membership of the Royal College of Paediatrics and Child Health (MRCPCH) written papers. This is so the potential for GPT-4 to augment clinical practice and medical education within contemporary paediatrics can be considered.

## METHODS

The Royal College of Paediatrics and Child Health makes available four sample papers (and their answers) on its website for those preparing to sit the MRCPCH theory components: one Foundation of Practice paper, one Theory and Science paper, and two Applied Knowledge in Practice papers. The author

subjected GPT-4 to all 114 questions (and their multiple-choice answer options) in these papers on 11 December 2023. The model had received no further training or reinforcement and was initially prompted with the instruction 'Pretend you are a paediatric specialty doctor in the UK. Answer the following questions as if you are a paediatric specialty doctor in the UK.' For each question, the model was prompted with the question's textual information and multiple-choice answer options. Any question images were attached to the prompt. Only one prompt was provided for each question, except when the model reported being unable to view or interpret an attached image (only one further prompt was made in each instance).

## RESULTS

The model scored 89.3% on the Foundation of Practice examination, 92.9% on Theory and Science examination, 67.9% on Applied Knowledge in Practice 1, and 63.3% on Applied Knowledge in Practice 2. The average score across all four papers was 78.1%. The model provided reasoning for its answers despite not being required to do so.

## DISCUSSION

The model performed especially well in the first two papers, which were exclusively composed of textual data. In answering these knowledge-based multiple-choice questions, and in its answer explanations, GPT-4 demonstrated impressive capabilities which imitated sound clinical reasoning. It performed less well in the second two papers, particularly in questions requiring image interpretation. In one question, while the model analysed the image (a technetium-99m scan), its answer was not based on its analysis of the image, but on the reasoning that this kind of scan is most commonly used to diagnose Meckel's



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

School of Medicine, University of Nottingham, Nottingham, UK

## Correspondence to

Dr Richard Armitage; richard.armitage@nhs.net



diverticulum, meaning Meckel's diverticulum was the most likely correct answer.

All incorrect answers that were not due to limitations around image interpretation were accompanied by strong hallucinations in which incorrect answers were explained with the same degree of confidence with which explanations of correct answers were provided (the model never disclosed ignorance, uncertainty or hesitancy),<sup>6</sup> thereby revealing the technology's lack of true 'understanding.' Although the questions are publicly available—meaning they might have been included in GPT-4's training data—the model's clinical reasoning ability implies that the LLM is not simply retrieving answers it has been previously exposed to.

The apparent competence of this LLM in this domain does not reflect the genuine aptitude that is required by paediatric specialty doctors, who confront real-world clinical situations in which relevant information is largely unstructured and thus unlike the succinct and uncontaminated packages of MRCPCH-style questions. Furthermore, the MRCPCH-style questions test nothing other than the foundational knowledge requirements of a practising paediatric specialty doctor, and do not assess the professional attitudes and clinical skills that are foundational to safe and effective practice. Despite these limitations, these results strengthen the case for LLM to augment both clinical practice and medical education within contemporary paediatrics.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; internally peer reviewed.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

#### ORCID iD

Richard Armitage <http://orcid.org/0000-0003-1165-6753>

#### REFERENCES

- 1 Armitage RC. Chatgpt: the threats to medical education. *Postgrad Med J* 2023;99:1130–1.
- 2 Howard A, Hope W, Gerada A. Chatgpt and antimicrobial advice: the end of the consulting infection doctor? *Lancet Infect Dis* 2023;23:405–6.
- 3 Patel SB, Lam K. Chatgpt: the future of discharge summaries? *Lancet Digit Health* 2023;5:e107–8.
- 4 Ali SR, Dobbs TD, Hutchings HA, *et al*. Using Chatgpt to write patient clinic letters. *Lancet Digit Health* 2023;5:e179–81.
- 5 Jeblick K, Schachtner B, Dextl J, *et al*. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified Radiology reports. *Eur Radiol* 2023.
- 6 Alkaissi H, McFarlane SI. Artificial hallucinations in chatGPT: implications in scientific writing. *Cureus* 2023;15:e35179.