Original Paper

Trialling a Large Language Model (ChatGPT) in General Practice With the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care

Arun James Thirunavukarasu[1], BA; Refaat Hassan[1], BA; Shathar Mahmood[1], BA; Rohan Sanghera[1], BA; Kara Barzangi[1], BA; Mohanned El Mukashfi[1], BA; Sachin Shah[2], MBBS

[1]University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom

[2]Attenborough Surgery, Bushey Medical Centre, Bushey, United Kingdom

Corresponding Author:
Arun James Thirunavukarasu, BA
University of Cambridge School of Clinical Medicine
Box 111 Cambridge Biomedical Campus
Cambridge, CB2 0SP
United Kingdom
Phone: 44 0 1223 336732 ext 3
Email: ajt205@cantab.ac.uk

Abstract

Background: Large language models exhibiting human-level performance in specialized tasks are emerging; examples include Generative Pretrained Transformer 3.5, which underlies the processing of ChatGPT. Rigorous trials are required to understand the capabilities of emerging technology, so that innovation can be directed to benefit patients and practitioners.

Objective: Here, we evaluated the strengths and weaknesses of ChatGPT in primary care using the Membership of the Royal College of General Practitioners Applied Knowledge Test (AKT) as a medium.

Methods: AKT questions were sourced from a web-based question bank and 2 AKT practice papers. In total, 674 unique AKT questions were inputted to ChatGPT, with the model's answers recorded and compared to correct answers provided by the Royal College of General Practitioners. Each question was inputted twice in separate ChatGPT sessions, with answers on repeated trials compared to gauge consistency. Subject difficulty was gauged by referring to examiners' reports from 2018 to 2022. Novel explanations from ChatGPT—defined as information provided that was not inputted within the question or multiple answer choices—were recorded. Performance was analyzed with respect to subject, difficulty, question source, and novel model outputs to explore ChatGPT's strengths and weaknesses.

Results: Average overall performance of ChatGPT was 60.17%, which is below the mean passing mark in the last 2 years (70.42%). Accuracy differed between sources (P=.04 and .06). ChatGPT's performance varied with subject category (P=.02 and .02), but variation did not correlate with difficulty (Spearman ρ=–0.241 and –0.238; P=.19 and .20). The proclivity of ChatGPT to provide novel explanations did not affect accuracy (P>.99 and .23).

Conclusions: Large language models are approaching human expert–level performance, although further development is required to match the performance of qualified primary care physicians in the AKT. Validated high-performance models may serve as assistants or autonomous clinical tools to ameliorate the general practice workforce crisis.

(JMIR Med Educ 2023;9:e46599) doi: 10.2196/46599

KEYWORDS

ChatGPT; large language model; natural language processing; decision support techniques; artificial intelligence; AI; deep learning; primary care; general practice; family medicine; chatbot

Introduction

Deep learning is a form of artificial intelligence (AI), which facilitates the development of exquisitely organized processing within an artificial neural network architecture, composed of multiple layers of interlinked perceptron nodes [1]. During supervised training of these models, the nature and weighting of communicating links between perceptrons is tuned to

XSL•FO
RenderX

optimize performance in a predefined task. While also applied to structured (tabulated) data, as with longer-established computational techniques, deep learning has enabled AI to work with unstructured inputs and outputs, such as images, videos, and sounds [1]. In recent years, natural language processing (NLP) has leveraged deep learning to extend the analytical and productive capability of computational models to unstructured language.

Generative Pretrained Transformer 3.5 (GPT-3.5) is a large language model (LLM), trained on a data set of over 400 billion words from articles, books, and other forms of media on the internet [2]. ChatGPT is a web-based chatbot that uses GPT-3.5 to directly answer users' queries. Unlike most chatbots previously trialed in clinical settings, ChatGPT facilitates free-text input and spontaneous output, as opposed to manually designed finite-state inputs and outputs [3]. ChatGPT has already begun to be trialed in medical contexts and has garnered attention for attaining sufficient accuracy in medical licensing examinations to graduate as a doctor, with even better performance recorded since the release of GPT-4 as the application's backend LLM [4-6]. As primary care struggles with poor recruitment, increasing workload, and early retirement [7-9], the introduction of autonomous decision aids and advisors may complement existing initiatives to improve the provision of general practitioners (GPs) [7,10]. Innovation in this sector would enable maximizing of the value provided by practicing GPs, likely benefiting deprived and rural areas—where fewer doctors serve the population—the most [11].

The Applied Knowledge Test (AKT) of the Membership of the Royal College of General Practitioners (RCGP) must be passed for GPs to complete their training in the United Kingdom. A total of 200 questions—mostly multiple choice but with occasional requirement to input numbers or select from a longer list of potential answers—must be answered in 190 minutes by candidates at a computer workstation. Questions test mostly clinical knowledge (80%), as well as evidence-based practice (10%) and primary care organizational and management skills (10%). All questions are designed to test higher-order reasoning rather than simple factual recall.

Before trials of clinical applications of NLP chatbots can be designed, the proposed purpose of applications such as ChatGPT must be established, requiring thorough investigation of their strengths and weaknesses. To evaluate the utility of ChatGPT in primary care settings, we used the AKT as an existing standard met by all UK GPs. The distinct sections of the AKT enabled the investigation of the opportunities afforded by ChatGPT (and LLMs more broadly), as well as the limitations of currently available technology. Through this work, we aimed to provide suggestions as to how clinical and computational research should proceed with the design and implementation of NLP chatbots, supported by empirical data.

Methods

Overview

AKT questions were sourced from the RCGP's GP SelfTest platform [12], as well as 2 publicly available practice papers [13,14]. Twenty questions were extracted from each subject category on the GP SelfTest platform, and all questions were extracted from the practice papers. Two researchers matched the subject categories of the practice papers' questions to those defined in GP SelfTest and in AKT examiners' reports from 2018 to 2022, with disagreements resolved through discussion and arbitration by a third researcher. Questions and multiple answer choices were copied from these three sources for entry into ChatGPT. Questions with multiple parts were prepared as distinct entries. Questions requiring appraisal of non–plain text elements that could not be copied into ChatGPT were excluded from the study. Duplicate questions were identified by a single researcher and excluded from the study.

Every eligible question was inputted into ChatGPT (January 30, 2023, version; OpenAI) on 2 separate occasions between January 30 and February 9, 2023, in separate sessions to avoid the second trial from being influenced by previous dialogue. ChatGPT's answer was recorded, and its whole reply to each question was recorded for further analysis. If ChatGPT failed to provide a definitive answer, the question was retrialed up to 3 times, after which ChatGPT's answer was recorded as "null" if no answer was provided. Correct answers (ie, the "ground truth") was defined as the answers provided by GP SelfTest and the practice papers—these were recorded for every eligible question. ChatGPT's responses were screened for "novel explanations"—defined as any information provided that was not included in the question or multiple choice answers—by a single researcher.

The scores required to pass the AKT in every examination undertaken in the last 2 years were collected from RCGP examiners' reports for the AKT between 2018 and 2022 [15]. Additionally, the number of recommendations of "room for improvement" for each subject category in the last 5 years were collected to use as a measure of "difficulty" in subsequent analysis.

ChatGPT's answers in both trials were compared to the correct answers to gauge performance and were compared to recent pass marks to assess ChatGPT's prospects of passing the AKT. ChatGPT's answers were compared between the 2 trials to measure the consistency of its responses. Performance was analyzed with respect to difficulty, explanation novelty, source, and subject to explore the strengths and weaknesses of ChatGPT. Nonparametric statistical analysis was undertaken due to the nonrandom nature of question design and small number of questions in some subjects. Effect sizes were reported with 95% CI and *P* values, with statistical significance concluded where $P<.05$. Statistical analysis was conducted in R (version 4.1.2; R Foundation for Statistical Computing), and figures were produced using Affinity Designer (version 1.10.6; Serif Ltd).

Ethics Approval

Ethics approval was not required for this study as human participants were not involved.

Results

In total, 720 questions were identified, which increased to 733 questions after multipart questions were separated into distinct

XSL•FO
RenderX

entries. In total, 674 unique questions were ultimately inputted into ChatGPT after duplicate and incompatible questions were excluded (Figure 1). Incompatibility was due to the question including an image in 35 cases and the inclusion of a table in 11 cases.

Exemplar questions and answers are depicted in Figure S1 in Multimedia Appendix 1. Overall performance was consistent: 59.94% (404/674) on the first run and 60.39% (407/674) on the second run. ChatGPT expressed uncertainty or did not provide an answer to repeated inquiry on 4 occasions in the first trial and on 6 occasions in the second trial, corresponding to 1.48% and 2.25% of incorrect answers, respectively. ChatGPT gave the same answer on both runs in response to 83.23% (561/674) of the questions, indicating variability in a significant proportion of cases. For reference, the average pass mark for the AKT in the last 2 years has been 70.42%, ranging from 69.00% to 71.00% [15]. Performance differed by question source (Table 1): variation was significant in the second (Fisher exact test, P=.04) but not the first (Fisher exact test, P=.06) trial. This indicates that question difficulty (for ChatGPT) differed between sources, although differences in performance were not large (Figure S2 in Multimedia Appendix 1).

Figure 1. Flowchart illustrating how questions were sourced and processed before inputting into ChatGPT and extracting answers for further analysis. GP: general practitioner.

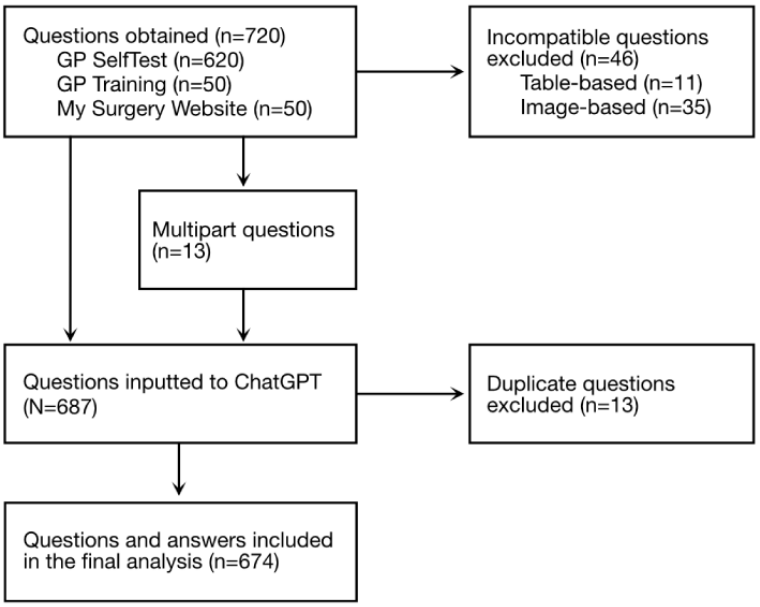


Table 1. Overall performance of ChatGPT in both trials, stratified by question source.

Source	GP[a] SelfTest [12]	My Surgery Website [13]	GP Training Schemes [14]
Questions, n	599	44	31
Trial 1, n (%)			
Correct answers	368 (61.60)	23 (52.27)	13 (41.94)
Incorrect answers	231 (38.56)	21 (47.73)	18 (58.06)
Trial 2, n (%)			
Correct answers	372 (62.10)	21 (47.73)	14 (45.16)
Incorrect answers	227 (37.90)	23 (52.27)	17 (54.85)

[a]GP: general practitioner.

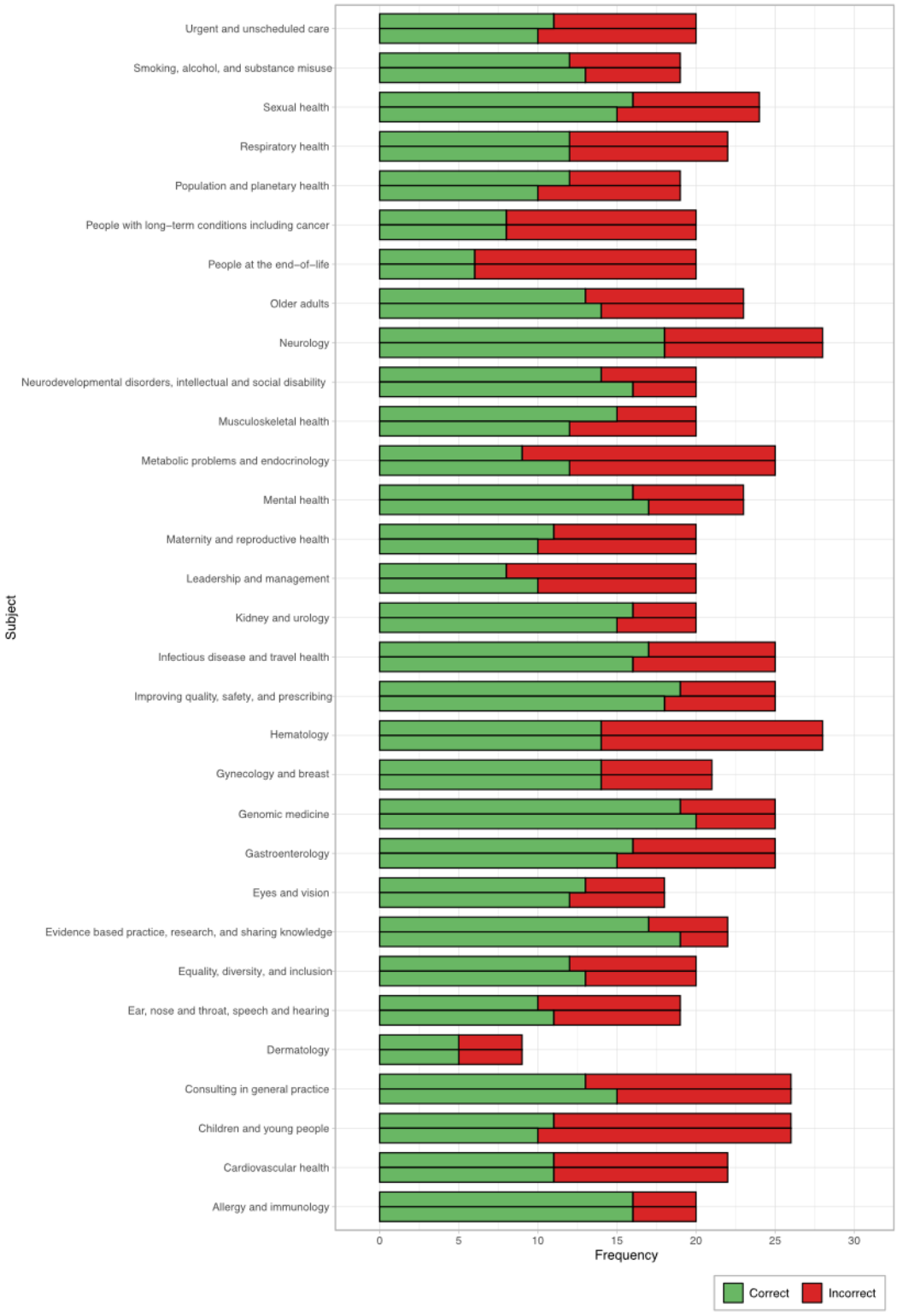
Performance was highly variable between subjects (Figure 2), with significant variation observed in the first (Fisher exact test estimated over 10^6 iterations, P=.02) and second (Fisher exact test estimated over 10^6 iterations, P=.02) trials. Subject variation did not correlate with the difficulty indicated by the frequency of recommendations of "room for improvement" by the RCGP (Spearman correlation coefficient for the first run [ρ]=–0.241, P=.19; Spearman ρ for the second run=–0.238, P=.20; Figure 3). Average accuracy over 75% was exhibited in 4 subjects: intellectual and social disability, kidney and urology, genomic medicine, and allergy and immunology (Table S1 in Multimedia Appendix 1). Accuracy under 50% on average was exhibited in 5 subjects: leadership and management, metabolic problems and endocrinology, children and young people, people with long-term conditions including cancer, and people at the end-of-life (Table S1 in Multimedia Appendix 1).

ChatGPT provided novel explanations in response to 58 (8.61%) questions in the first run and 66 (9.79%) questions in the second run. A novel explanation was provided in response to just 18

XSL•FO
RenderX

(2.67%) questions in both runs, illustrating significant stochasticity in the relationship between prompt and output. The proclivity of ChatGPT to provide a novel explanation had no bearing on accuracy in the first (Fisher exact test odds ratio 1.02, 95% CI 0.57-1.85, P>.99) or second (Fisher exact test odds ratio 0.72, 95% CI 0.42-1.24, P=.23) iterations (Figure 4).

Figure 2. ChatGPT's performance in 674 questions on the Membership of the Royal College of General Practitioners Applied Knowledge Test, stratified by subject category. The higher bar within each subject corresponds to the first trial; the lower bar corresponds to the second trial.



XSL•FO
RenderX

Figure 3. Correlation between ChatGPT performance and subject difficulty, expressed in terms of the Spearman rank correlation coefficient (ρ).

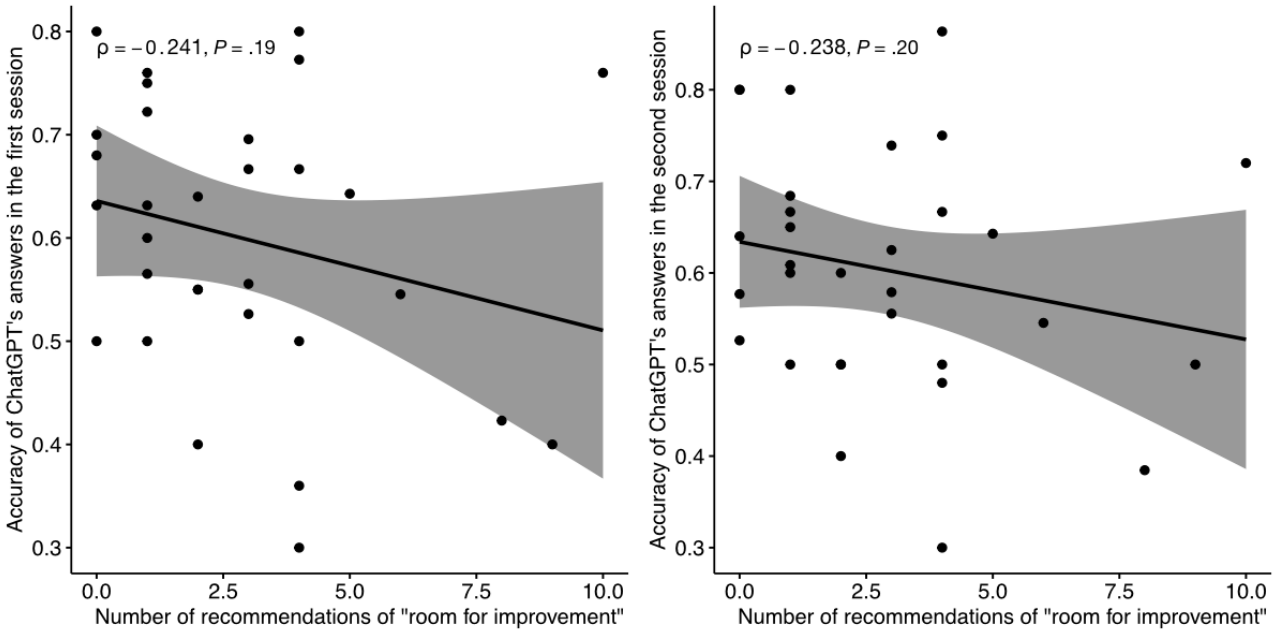
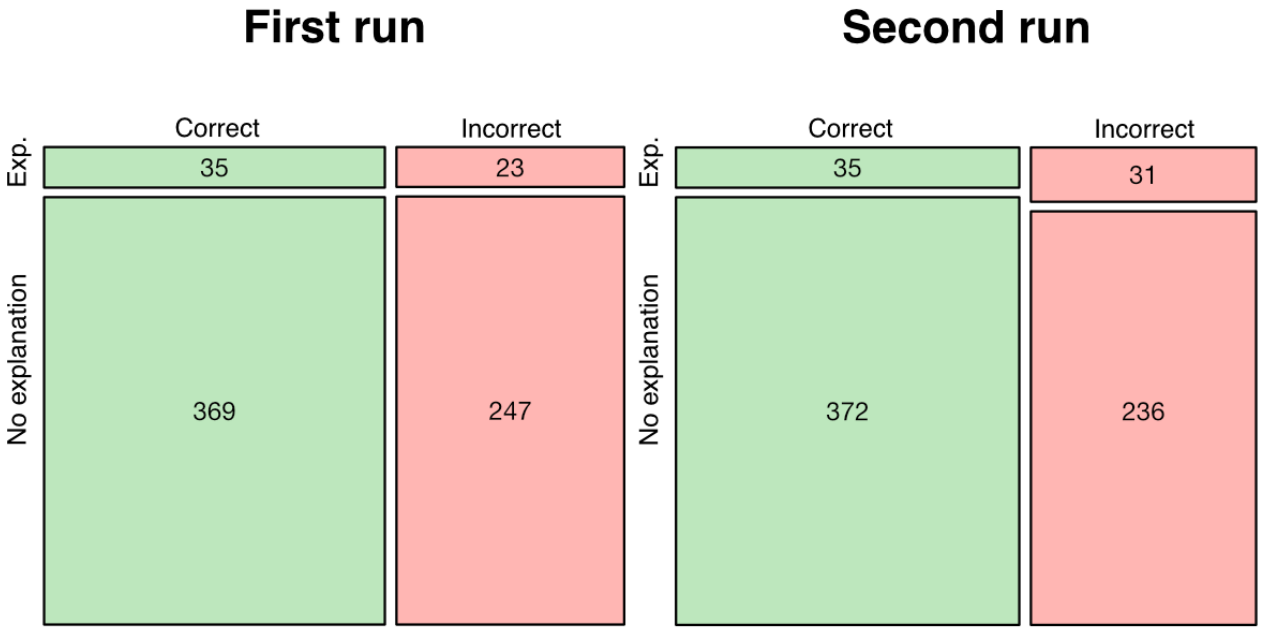


Figure 4. Mosaic plot depicting the relationship between ChatGPT's proclivity to provide a novel explanation and answer accuracy. Exp.: explanation provided.



Discussion

This study makes 5 significant observations. First, performance in a national primary care examination cannot be passed by ChatGPT, although the platform came close in terms of accuracy to AKT pass marks in recent years. Contrary to some academic and media reports, AI cannot replace human doctors who remain indispensable within general practice. As ChatGPT attained sufficient performance to pass medical school examinations, its semantic knowledge base appears to lie between the minimum standards to graduate as a doctor and to qualify as a GP [5,16]. Second, ChatGPT's performance is highly variable between subjects, suggesting that NLP applications must be deployed within highly specified roles to avoid compromising efficacy. Given the impressive performance of ChatGPT in certain subjects of the AKT, chatbots may be capable of providing useful input within narrowly defined portions of primary care.

Third, ChatGPT expresses uncertainty or technical limitation in a small minority of the cases in which it provides an incorrect answer. This limits the confidence patients and practitioners may place in chatbots' answers, as there is no obvious way to determine the model's uncertainty. This increases the risk of decisions based on inaccurate answers that occur too frequently to allow these applications to be deployed without supervision; this limits the current potential of this technology to automate health care processes. Additionally, use of ChatGPT as an educational tool in primary care is compromised by its frequent

XSL•FO
RenderX

errors, which may not be noticed by learners. Fourth, the proclivity or ability of ChatGPT to provide novel explanations has no bearing on the accuracy of its responses, which remains inconsistent—the application frequently "hallucinates," describing inaccurate information as lucidly as with correct facts. This compounds the issues regarding application of chatbots as decision support tools or educational assistants as discussed above. Lastly, the difficulty of subject categories based on GP trainee performance does not correlate with ChatGPT's performance at the subject level—human perceptions or manifestations of complexity or difficulty cannot be translated to NLP models without validation.

This study comprehensively assesses the performance of ChatGPT across the domains of primary care assessed in the AKT, with a large sample size providing a realistic estimate of the application's prospects were it to sit an official AKT paper. This provides valuable insight into NLP chatbots' strengths and weaknesses as applied to general practice and facilitates research into model development and implementation based on data-driven conclusions. However, there were 2 limitations to this study. First, passing the AKT does not equate to demonstrating ability to perform as a GP; subsequent models with improved performance may or may not be appropriate for autonomous deployment. GPs' knowledge and skills are tested in a variety of ways from medical school onward, with the AKT representing just one of many official assessments. Second, questions containing images or tables could not be inputted to ChatGPT, which may have affected our results. Emerging multimodal LLMs such as GPT-4 are compatible with all questions in the AKT, and our protocol provides a benchmark and methodology for trials of future models.

ChatGPT has garnered particular attention in recent months due to its performance in tasks previously considered completable by humans alone, such as passing medical school examinations such as the United States Medical Licensing Examination [5,16]. Other LLMs have exhibited similar achievements, such as FlanPaLM [17]. The ability of ChatGPT to accurately answer questions, provide useful advice, and triage based on clinical vignettes consistently exceeds that of a layperson [5,18]. However, the accuracy of computational models' answers to medical questions is yet to exceed that of fully trained physicians, with findings in the present context of primary care being no exception [16,17]. When ChatGPT is used as a medical advice chatbot, advice seekers are only able to identify that the source of provided advice is computational 65% of the time [19]. It follows that health care providers must protect their patients from inaccurate information provided by this technology, as they are unable to differentiate between computational and human advice [19]. This requirement for oversight limits the potential of LLMs to meaningfully change practice, as performance equivalent to that of experts is the minimum standard to justify autonomous deployment: there must be confidence in the accuracy and trustworthiness of answers from these applications [20,21].

The excellent performance of ChatGPT in certain sections of the AKT indicates that deployment may be feasible within strictly bounded tasks. NLP chatbots may provide useful assistance to clinicians, but application as an autonomous decision maker is not currently justified by exhibited performance. Examples of potential uses include interpretation of objective data such as laboratory reports, triage (a fully automated conveyor model or with human management of edge cases), and semiautonomous completion of administrative tasks such as clinic notes, discharge summaries, and referral letters [21,22]. Further work is required to engineer models with supraexpert performance in any domain of primary care, which could justify deployment as an autonomous component of care provision [21]. Additionally, uncertainty indicators or contingency messages where the model is unable to answer with accuracy could improve confidence in the information provided and, therefore, safety [19,20]. Specific study is required to ensure that new tools reduce rather than increase workload for GPs [23-25]. As this technology continues to advance, individualistic care must not be sacrificed: general practice consulting involves long-term development of a therapeutic relationship between patients and physicians, and chatbots should not be allowed to change this dynamic into an impersonal, transactional arrangement [21,24]. Optimal management of patients' issues is governed by patients' wishes and circumstances in addition to the empirical evidence base.

Chatbots leveraging advanced NLP models are an exciting innovation with the potential to ameliorate staffing pressures that disproportionately affect deprived areas [11]. However, improvement in domain-specific tasks is required to enable this technology to make a meaningful contribution. Improvement is not a simple matter of increasing the size of the data set used to train these large language models. Larger models do not always exhibit superior performance in highly specialized tasks such as answering medical questions [26]. This is likely due to most available training material being irrelevant to medical tasks, as text is sourced from across the internet. While training may be improved by sourcing greater volumes of domain-specific text, development is complicated by restricted-access sensitive patients' data, which likely comprises the largest unused source of information for large language models. Concerns regarding privacy and transparency of use currently limit the access of the largest NLP engineering companies to these data [27]. Alternative means of improving performance include fine-tuning by inputting a set of prompts or instructions to the model before it is deployed on a medical task. Fine-tuning has been shown to improve the performance of models beyond that of larger (but untuned) models, and fine-tuned LLMs are still state-of-the-art in terms of performance in medical questions, despite competition from ChatGPT, GPT-3.5, and GPT-4 [6,17,26,28]. It follows that similar tuning protocols may be applied to GPT-3.5 or ChatGPT to further optimize performance—this may be explored in backend development or by chatbot users experimenting with initial prompts before initiating a trial.

Effective applications must be rigorously trialed in the same context as the one they are intended to be deployed in the future [24,29]. As evidence supporting the integration of previously developed chatbots into primary care has suffered from poor reporting quality and high risk of bias, improved research practices are necessary to ensure that contemporary innovation fulfils its potential in terms of translated into impactful changes

XSL•FO
RenderX

in clinical practice [30]. Validated NLP models may be more broadly applicable, such as within different language mediums, but revalidation and proper clinical governance are essential mechanisms to protect patients from harm [31]. As LLM-based chatbots have only recently begun to exhibit human or near-human ability to complete complicated tasks [3], a new set of evidence is about to be generated: this represents an opportunity to improve research practices to maximize the chance of innovative applications translating into impactful changes in clinical practice [22]. NLP technology may prove to be an integral part of a solution to the issues of staffing shortages, population growth, and health care inequities.

Acknowledgments

AJT and SS extend their thanks to Dr Sandip Pramanik for his advice and tutelage.

Authors' Contributions

AJT and SS conceived and designed the study. AJT, RH, SM, RS, KB, MEM, and SS undertook data collection. AJT conducted data analysis and visualization. AJT, RH, and SS drafted the manuscript. SM, RS, KB, and MEM provided feedback on the manuscript and assisted with redrafting. All authors approved the submitted version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Exemplar questions and answers on the ChatGPT interface; mosaic plots stratifying performance by question source; and table stratifying performance by subject alongside the number of recommendations for improvement given by examiners based on human examination performance.
[PDF File (Adobe PDF File), 684 KB-Multimedia Appendix 1]

References

1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan 7;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. 2020 Presented at: 34th Conference on Neural Information Processing Systems (NeurIPS 2020); December 6-12, 2020; Vancouver, BC URL: https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
3. Parmar P, Ryu J, Pandya S, Sedoc J, Agarwal S. Health-focused conversational agents in person-centered care: a review of apps. NPJ Digit Med 2022 Feb 17;5(1):21 [FREE Full text] [doi: 10.1038/s41746-022-00560-6] [Medline: 35177772]
4. James CA, Wheelock K, Woolliscroft J. Machine learning: the next paradigm shift in medical education. Acad Med 2021 Jul 01;96(7):954-957. [doi: 10.1097/ACM.0000000000003943] [Medline: 33496428]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb 9;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online March 20, 2023. [FREE Full text]
7. Majeed A. Shortage of general practitioners in the NHS. BMJ 2017 Jul 10;358:j3191 [FREE Full text] [doi: 10.1136/bmj.j3191] [Medline: 28694250]
8. Sturmberg JP, O'Halloran DM, McDonnell G, Martin CM. General practice work and workforce: interdependencies between demand, supply and quality. Aust J Gen Pract 2018 Aug 01;47(8):507-513. [doi: 10.31128/ajgp-03-18-4515]
9. Razai MS, Majeed A. General practice in England: the current crisis, opportunities, and challenges. J Ambul Care Manage 2022;45(2):135-139. [doi: 10.1097/JAC.0000000000000410] [Medline: 35202030]
10. Marchand C, Peckham S. Addressing the crisis of GP recruitment and retention: a systematic review. Br J Gen Pract 2017 Mar 13;67(657):e227-e237. [doi: 10.3399/bjgp17x689929]
11. Nussbaum C, Massou E, Fisher R, Morciano M, Harmer R, Ford J. Inequalities in the distribution of the general practice workforce in England: a practice-level longitudinal analysis. BJGP Open 2021 Aug 17;5(5):BJGPO.2021.0066. [doi: 10.3399/bjgpo.2021.0066]
12. GP SelfTest. Royal College of General Practitioners. URL: https://elearning.rcgp.org.uk/course/index.php?categoryid=56 [accessed 2023-02-15]
13. MRCGP Applied Knowledge Test. Royal College of General Practitioners. URL: https://www.mysurgerywebsite.co.uk/website/IGP604/files/MRCGP%20AKT%20questions%20with%20answers.pdf [accessed 2023-02-15]
14. AKT Example Questions. Royal College of General Practitioners. 2019. URL: https://gp-training.hee.nhs.uk/cornwall/wp-content/uploads/sites/86/2021/04/RCGP-Sample-questions-2019-with-answers.pdf [accessed 2023-02-15]

XSL•FO
RenderX

15. MRCGP: Applied Knowledge Test (AKT). Royal College of General Practitioners. URL: https://www.rcgp.org.uk/mrcgp-exams/applied-knowledge-test [accessed 2023-02-15]
16. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. 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 2023 Feb 08;9:e45312 [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
17. Singhal K, Azizi S, Tu T, Madhavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. arXiv. Preprint posted online December 26, 2022. [doi: 10.48550/arXiv.2212.13138]
18. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. medRxiv. :5067 Preprint posted online February 1, 2023. [FREE Full text] [doi: 10.1101/2023.01.30.23285067] [Medline: 36778449]
19. Nov O, Singh N, Mann DM. Putting ChatGPT's medical advice to the (Turing) test. medRxiv. Preprint posted online January 24, 2023. [doi: 10.1101/2023.01.23.23284735]
20. Koman J, Fauvelle K, Schuck S, Texier N, Mebarki A. Physicians' perceptions of the use of a chatbot for information seeking: qualitative study. J Med Internet Res 2020 Nov 10;22(11):e15185 [FREE Full text] [doi: 10.2196/15185] [Medline: 33170134]
21. Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General practitioners' attitudes toward artificial intelligence-enabled systems: interview study. J Med Internet Res 2022 Jan 27;24(1):e28916 [FREE Full text] [doi: 10.2196/28916] [Medline: 35084342]
22. Gunasekeran DV, Tham Y, Ting DSW, Tan GSW, Wong TY. Digital health during COVID-19: lessons from operationalising new models of care in ophthalmology. Lancet Digit Health 2021 Feb;3(2):e124-e134. [doi: 10.1016/s2589-7500(20)30287-9]
23. Fletcher E, Burns A, Wiering B, Lavu D, Shephard E, Hamilton W, et al. Workload and workflow implications associated with the use of electronic clinical decision support tools used by health professionals in general practice: a scoping review. BMC Prim Care 2023 Jan 20;24(1):23 [FREE Full text] [doi: 10.1186/s12875-023-01973-2] [Medline: 36670354]
24. Tossaint-Schoenmakers R, Versluis A, Chavannes N, Talboom-Kamp E, Kasteleyn M. The challenge of integrating eHealth into health care: systematic literature review of the Donabedian model of structure, process, and outcome. J Med Internet Res 2021 May 10;23(5):e27180 [FREE Full text] [doi: 10.2196/27180] [Medline: 33970123]
25. Kremer L, Lipprandt M, Röhrig R, Breil B. Examining mental workload relating to digital health technologies in health care: systematic review. J Med Internet Res 2022 Oct 28;24(10):e40946 [FREE Full text] [doi: 10.2196/40946] [Medline: 36306159]
26. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. arXiv. Preprint posted online October 20, 2022. [doi: 10.48550/arXiv.2210.11416]
27. Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. J Med Ethics 2020 Jun 26;46(6):367-377 [FREE Full text] [doi: 10.1136/medethics-2019-105472] [Medline: 32457202]
28. Matias Y, Corrado G. Our latest health AI research updates. The Keyword. Google. 2023. URL: https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/ [accessed 2023-03-16]
29. Thirunavukarasu AJ, Hassan R, Limonard A, Savant SV. Accuracy and reliability of self-administered visual acuity tests: systematic review of pragmatic trials. medRxiv. Preprint posted online February 3, 2023. [doi: 10.1101/2023.02.03.23285417]
30. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
31. Malamas N, Papangelou K, Symeonidis AL. Upon improving the performance of localized healthcare virtual assistants. Healthcare (Basel) 2022 Jan 04;10(1):99 [FREE Full text] [doi: 10.3390/healthcare10010099] [Medline: 35052263]

Abbreviations

AI: artificial intelligence
AKT: Applied Knowledge Test
GP: general practitioner
GPT: Generative Pretrained Transformer
LLM: large language model
NLP: natural language processing
RCGP: Royal College of General Practitioners

XSL•FO
RenderX

Edited by T Leung, T de Azevedo Cardoso, G Eysenbach; submitted 20.02.23; peer-reviewed by D Gunasekeran, S Pesälä, D Patel; comments to author 30.03.23; revised version received 31.03.23; accepted 11.04.23; published 21.04.23

Please cite as:
Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, Shah S
Trialling a Large Language Model (ChatGPT) in General Practice With the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care
JMIR Med Educ 2023;9:e46599
URL: https://mededu.jmir.org/2023/1/e46599
doi: 10.2196/46599
PMID:

©Arun James Thirunavukarasu, Refaat Hassan, Shathar Mahmood, Rohan Sanghera, Kara Barzangi, Mohanned El Mukashfi, Sachin Shah. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.04.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX