BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email info.bmjopen@bmj.com

BMJ Open: first published as 10.1136/bmjopen-2024-086148 on 26 December 2024. Downloaded from http://bmjopen.bmj.com/ on June 7, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

BMJ Open

ChatGPT (GPT-4) versus Doctors on Complex Cases in Family Medicine: An Observational Comparative Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-086148
Article Type:	Original research
Date Submitted by the Author:	06-Mar-2024
Complete List of Authors:	Arvidsson, Rasmus; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Praktikertjänst AB, Hälsocentralen Sankt Hans Entezarjou, Artin; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Capio Healthcare Centre Angered Sundemo, David; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Lerum Närhälsan Primary Healthcare Center Wikberg, Carl; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Region Vastra Gotaland, Research, Education, Development & Innovation, Primary Health Care Gunnarsson, Ronny; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Region Vastra Gotaland, Närhälsan, Vårdcentralen Hemlösa
Keywords:	Artificial Intelligence, Primary Health Care, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

ChatGPT (GPT-4) versus Doctors on Complex Cases in Family Medicine: An Observational Comparative Study

Rasmus Arvidsson[1,2], Ronny Gunnarsson[1,3,4], Artin Entezarjou[1,5], David Sundemo[1,6], Carl Wikberg[1,4]

[1]General Practice, School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[2]Hälsocentralen Sankt Hans, Lund, Sweden

[3]Region Vastra Gotaland, Närhälsan, Vårdcentralen Hemlösa
Göteborg, Sweden

[4]Research, Education, Development & Innovation, Primary Health Care, Region Västra Götaland, Sweden

[5]Capio Healthcare Centre Angered, Angered, Sweden

[6]Lerum Närhälsan Primary Healthcare Center, Lerum, Sweden

Corresponding author

Rasmus Arvidsson (rasmusalinks@gmail.com), ORCID 0009-0006-0387-3108

ORCIDs of other authors

Ronny Gunnarsson (0000-0001-9183-3072), Artin Entezarjou (0000-0002-7418-8750), David Sundemo (0000-0002-5871-1636), Carl Wikberg (0000-0002-6494-5922)

ABSTRACT

Background: Recent breakthroughs in AI research include the development of generative pretrained transformers (GPT). ChatGPT has been shown to perform well when answering several sets of medical multiple choice questions. However, it has not been tested on writing free text assessments of complex cases in primary care.

Objectives: Compare ChatGPT (GPT-4) to real doctors.

Design and Setting: A blinded observational comparative study conducted in the Swedish primary care setting. ChatGPT's and real doctors' responses for cases from the Swedish family medicine specialist exam were scored by blinded reviewers, and the scores compared.

Participants: Anonymous answers from the Swedish family medicine specialist exam 2017–2022 were used.

Outcome measures: Primary: Mean difference in scores between ChatGPT and randomly selected responses by human doctors, as well as between ChatGPT and top-tier responses. Secondary: Correlation between differences in response length and response score. Intraclass correlation coefficient between reviewers.

Results: The mean scores were 6.0, 7.2 and 4.5 for randomly selected doctor responses, top-tier doctor responses and ChatGPT responses, respectively, on a 10-point scale. The scores for the random doctor responses were, on average, 1.6 points higher than those of ChatGPT ($p < 0.001$, 95 % confidence interval (CI) 0.9–2.2) and the top-tier doctor scores were, on average, 2.7 points higher than those of ChatGPT ($p < 0.001$, 95 % CI 2.2–3.3).

Conclusion: On complex primary care cases, ChatGPT-4 performs worse than human doctors taking the Family Medicine Specialist Exam. Future GPT based chatbots may perform better, but comprehensive evaluations are needed before implementing chatbots for medical decision support in primary care.

STRENGTHS AND LIMITATIONS OF THIS STUDY

Strengths

- Each response was scored by two independent, blinded reviewers.
- Detailed scoring keys provided excellent interrater correlation.
- Evaluation of long-form free text responses to complex cases, relevant for primary care.

Limitations

- The result may not be fully generalisable to other countries and languages.
- There was no penalty for the presence of extra, unnecessary, information in the responses when scoring.

KEYWORDS

Artificial Intelligence

Primary Health Care

Medical Informatics

Technology Assessment, Biomedical

BACKGROUND

Artificial intelligence (AI) in medicine has been the subject of an increasing amount of research, even though real world applications are relatively few [1–3]. Over the last few years, large AI models called generative pretrained transformers (GPT) have demonstrated remarkable abilities beyond simple text generation, such as answering questions and participating in chat conversations. ChatGPT from OpenAI is arguably one of the most well-known models.

Countless clinical applications could be envisioned for an AI system which can accurately answer any questions from healthcare staff and patients. The impact could be enormous in primary health care, where healthcare staff need to keep themselves up-to-date on a broad spectrum of medical conditions.

ChatGPT, GPT versions 3.5 and 4, have demonstrated human-level performance on several professional benchmarks [4] and achieved moderate to excellent results in various medical exams [5–10] but did not pass the general practice licensing exams of Taiwan and the UK [11,12]. However, the medical questions in these assessments have typically been multiple-choice questions, which is different from a clinician asking the chatbot for advice on the management of real patient cases. Also, the studies focusing on general practice have both tested the 3.5 version of ChatGPT, which may perform significantly worse than the 4.0 version [6,10]. At the time of writing, research has not explored ChatGPT's (GPT-4) ability to provide comprehensive free text assessments of complex medical cases, prevalent in primary care. For example, primary care cases often involve intricate patient symptoms combined with social or behavioural factors, adding to their complexity. This study aims to address this gap.

METHODS

Study Design

This study compared the performance of ChatGPT (GPT-4) with responses from human doctors on cases from the Swedish family medicine specialist exam. The responses from three distinct groups were scored and compared: A) randomly selected doctor responses, B) top-tier doctor responses, and C) responses generated by ChatGPT.

Objective and outcome measures

The objective was to compare ChatGPT (GPT-4) to real doctors, regarding their ability to write comprehensive assessments of complex cases from primary care.

Primary outcome measure: Mean difference in scores between ChatGPT and randomly selected responses by human doctors, as well as between ChatGPT and top-tier responses.

Secondary outcome measures: Correlation between differences in response length and response score. Intraclass correlation coefficient between reviewers.

Data collection

Sourcing of medical cases

Forty-eight cases from previous versions of the Swedish family medicine specialist exam were utilised for this study. These exams are publicly available on the Swedish Association of General Practice (SFAM)'s, website [13]. The cases require comprehensive responses, typically several paragraphs of free text. They are often complex, with symptoms pointing towards different diseases and complicating factors such as social problems, addiction, poor compliance, legal aspects, and patients near the end of life.

Sourcing of doctor responses, group A and B

Group A: A digital random choice function was used to draw a single anonymous response for each case, from the human responses submitted to the exam when it took place. Group B: SFAM has published an example of a top-tier response for each case, chosen by the exam reviewers as an example of a very good response, and these were used for group B [13].

Obtaining ChatGPT responses, group C

Medical cases were sent to ChatGPT in an automated manner through OpenAI's API [14], using the version of GPT-4 released on August 3, 2023. Apart from the case itself, additional instructions were sent along with each case to provide some context, based on the written instructions included in the 2022 exam [See Supplemental file 1]. A separate chat session was created for each case.

Scoring the responses

For each case, SFAM has published an evaluation guide, containing a couple of main points which should be included in a good answer, but the precise scoring guide used for the exam is not public. To quantify the performance of each exam response, the published evaluation guide for each case was adapted into a criteria-based scoring guide. Each scoring guide could award a score ranging from 0 to 10 points. This adaptation involved rephrasing each evaluation guide into a set of true-or-false criteria. For each criterion met, a specific number of points would be awarded [See Supplemental file 2]. A group of three medical doctors, blinded to the origins of the responses, rated the responses using the scoring guide. Each response was scored by two of the three raters, and the average of their scores was used for the statistical analysis. The same pair of raters assessed all responses pertaining to the same case.

Statistical analysis

Sample size calculation

The level of significance was set to 0.025, power to 0.8 and minimal difference between groups to be detected was set to one point, which resulted in a required sample size of 48 cases.

Data Analysis

After scoring the responses to all 48 cases, the difference between each doctor group and ChatGPT was calculated for each case. A paired t-test was used to compare each doctor group with ChatGPT, pairing the scores by question.

To assess the reliability of the averaged scores derived from the raters' use of the scoring guide, we conducted an Intraclass Correlation Coefficient (ICC) analysis, specifically employing the Two-Way Mixed-Effects Model for the mean of k raters, utilising the psych package in R [15,16].

In addition, we examined the differences in response length (number of words) between the top-tier and ChatGPT responses. A paired t-test was used with the lengths paired by question. As a measure of the information density, we divided the score with the number of words for each response. Finally, a linear regression analysis was performed to explore the relationship between the difference in lengths and the difference in scores. The latter was set as the dependent variable and the former as the independent variable. The OLS function from the statsmodels library was employed for this analysis [17].

RESULTS

ChatGPT (GPT-4) scored lower than any doctor group (Table 1). The differences between groups were statistically significant (Table 2).

Table 1: Mean score, length and points per 100 words of each group

Group	Mean score	Mean length (words)	Points per 100 words
A: Random doctor	6.0 (5.5–6.6)	*320 (280–350)*	*2.2 (1.9–2.5)*
B: Top-tier doctor	7.2 (6.8–7.6)	350 (320–380)	2.3 (2.0–2.6)
C: ChatGPT (GPT-4)	4.5 (3.9–5.0)	290 (280–300)	1.6 (1.4–1.7)

95 % confidence intervals (CI) in parenthesis. The points per 100 words is the average of the points per 100 words for each case.

Table 2: Differences in scores between ChatGPT (GPT-4) and the doctor groups

	Mean difference (95 % CI)	P-value
A vs C: Random doctor versus ChatGPT	1.6 (0.9–2.2)	< 0.001
B vs C: Top-tier doctor versus ChatGPT	2.7 (2.2–3.3)	< 0.001

The ICC for the scores from the three raters was 0.92 (95% CI 0.90–0.94, p < 0.001), demonstrating excellent reliability of the scoring guide.

The top-tier responses were on average 60 words longer than ChatGPT's (p < 0.001, 95 % CI 30–97). The correlation between differences in length and differences in scores of responses between ChatGPT and the top-tier answers was not statistically significant (p = 0.11).

DISCUSSION

The main finding was that graduating specialists in general practice scored significantly better than ChatGPT (GPT-4) on the Swedish family medicine specialist exam, with top-tier responses scoring almost three points better. What such a difference corresponds to in practice differs a lot from case to case. For example, in one case, ChatGPT scored 2.75 points lower than the top-tier response due to mentioning one less important differential diagnosis and two fewer aspects of treatment and follow-up. Notably, the information density was also higher for the two doctor groups than for ChatGPT, indicating that human doctors are better at conveying relevant information concisely.

Comparison with the existing literature

In one study, ChatGPT (GPT-4) passed every test in a series of dermatology licensing exams, achieving over 80% for the English version (pass level: 60 %) [6]. No data were presented on the performance of real dermatologists for comparison. On the other hand, the average score of ChatGPT (GPT 3.5) was only 60.17% on the general practice licensing exam of the UK (pass level ≈ 70%) [12], and it scored 41.6% on the corresponding Taiwanese licensing exam (pass level = 60%) [11]. This aligns well with our results, even though we used a newer version of ChatGPT. Both these studies, and several similar studies in other medical disciplines [7–10], used multiple choice questions, which is a task very different from providing free text responses to complex clinical cases. Providing free text answers more closely resembles the requirements of a chatbot used for decision support in clinical practice. Many used GPT-3.5, which may perform significantly worse than GPT-4.

One study examined questions posted by patients online, on a forum where volunteering doctors responded [18]. In the study, three licensed health care professionals evaluated the free text responses. In 79% of the cases, they favoured ChatGPT's (GPT-3.5) responses over the doctors and the quality score was 21 % lower for doctors on average, as scored on a 5 category ordinal scale. These findings are opposite to the findings of our study, where the randomly selected doctors' responses scored higher in 71% of the cases, even though a newer version of ChatGPT was used. The questions and responses in the patient forum were typically shorter and simpler than the primary care cases used in our study, and the responses were not assessed on specific medical criteria. In a recent preprint, a novel chatbot AI, named AMIE, has been fine tuned to perform a diagnostic interview with a patient through chat [19]. It was compared with general practitioners on objective structured clinical examination (OSCE) cases and outperformed general practitioners on most metrics, including suggesting relevant differential diagnoses. This suggests that higher performance is already possible from AI models, but evaluating ChatGPT is still highly relevant, since it is widely accessible and may hypothetically already be used by patients and clinicians.

Strengths and limitations

This is the first study of ChatGPT's (GPT-4) performance on complex primary care cases with long-form free text responses, rather than multiple choice. As such, it mimics the scenario where a clinician posts a case summary of a real patient in order to get input on the management. The scoring system was a relatively clear way to quantify the amount of useful content in each answer, and demonstrated excellent reliability. No penalty was given to superfluous content, however, which could favour respondents writing longer, but less relevant, responses. The cases used in our study are representative for Swedish primary care, which may differ somewhat from other countries [20]. This should be taken into account when generalising our results to other countries. Using a different set of instructions to be sent with each case would have influenced the responses, for better or for worse [21]. The cases utilised in the study are publicly available online and could have been part of ChatGPT's training data, but the correct answers are not available in direct association with the questions, so we find it unlikely that this would have affected the result.

Implications for current practice and future research

ChatGPT (GPT-4) falls short in medical accuracy when writing comprehensive assessments of complex primary care cases, compared to human doctors. The difference in performance is both statistically significant and clinically relevant. Hence, case assessments by ChatGPT, version 4, should not be used directly by primary care doctors. Nor should ChatGPT be implemented as a doctor or nurse substitute for patients. However, new versions and models can be expected to arise.

Future research on medical chatbots should focus on evaluating emerging models on representative questions asked by clinicians and patients in a clinical setting. At the same time, in line with the previously mentioned AMIE medical chatbot [19], researchers and developers should aim to optimise the performance of such chatbots, for example by training them specifically on reliable medical information, optimising prompt engineering techniques [21,22], using algorithms for processing a single question in multiple steps or allowing the chatbots access to external sources of information and tools, including other categories of AI-models [23,24].

DECLARATIONS

Ethics approval

Responses from human participants of the Swedish family medicine specialist exam were obtained anonymously without an identifier linking responses to individuals. According to Swedish regulations collecting this type of data anonymously does not require an ethics approval.

Patient involvement statement

There was no patient or public involvement in the development of this study.

Availability of data and materials

At the time of writing, the scores are in the review process for being published in the Swedish National Data Service's Data Organisation and Information System repository [25]. Three examples of cases and their corresponding scoring guides have been translated to English and included as supplemental file 2. The original cases, evaluation guides and top-tier responses are publicly available in Swedish at SFAM's website [13], from where they were used in this study with permission.

Funding

This work was supported by the Västra Götaland region Sweden, a publicly run health care provider.

Competing Interests

None.

Authors' contributions

The study was planned by authors RA, RG, AE, DS and CW. Author RA collected the cases and responses, and compiled the scoring guides. Authors DS and AE participated in the group of raters who scored the responses. Author RA performed the statistical analysis and created the draft of the manuscript. Authors RA, RG, AE, DS and CW participated in discussing results and refining the manuscript.

Acknowledgements

Special thanks to The Swedish Association of General Practice, SFAM, for granting permission to use the Swedish family medicine specialist exam. Specifically, Karin Lindhagen was very helpful in compiling the randomly selected responses. Special thanks also to Dr Abed Alsabbagh, who participated in the group reviewing and scoring the responses.

References

1. Niu J, Tang W, Xu F, et al. Global Research on Artificial Intelligence from 1990–2014: Spatially-Explicit Bibliometric Analysis. ISPRS Int J Geo-Inf. 2016 May;5(5):66. DOI:

10.3390/ijgi5050066

2. Yin J, Ngiam KY, Teo HH. Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. J Med Internet Res. 2021 Apr 22;23(4):e25759. DOI: 10.2196/25759
3. Triantafyllidis AK, Tsanas A. Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature. J Med Internet Res. 2019 Apr 5;21(4):e12286. DOI: 10.2196/12286
4. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys Syst. 2023 Jan 1;3:121–54. DOI: 10.1016/j.iotcps.2023.04.003
5. Gilson A, Safranek CW, Huang T, 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 8;9:e45312. DOI: 10.2196/45312
6. Lewandowski M, Łukowicz P, Świetlik D, et al. An original study of ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level based on the Dermatology Specialty Certificate Examinations. Clin Exp Dermatol. 2023 Aug 4;llad255. DOI: 10.1093/ced/llad255
7. Antaki F, Touma S, Milad D, et al. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol Sci. 2023 Dec 1;3(4):100324. DOI: 10.1016/j.xops.2023.100324
8. Kung JE, Marshall C, Gauthier C, et al. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. JBJS Open Access. 2023 Sep;8(3):e23.00056. DOI: 10.2106/JBJS.OA.23.00056
9. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023 Mar 22;29(3):721–32. DOI: 10.3350/cmh.2023.0089
10. Taloni, Andrea, Borselli, Massimiliano, Scarsi, Valentina, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology | Scientific Reports. Sci Rep. 2023 Oct 29;13(18562). DOI: 10.1038/s41598-023-45837-2
11. Weng TL, Wang YM, Chang S, et al. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc. 2023 Aug;86(8):762. DOI: 10.1097/JCMA.0000000000000946
12. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. 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 Apr 21;9:e46599. DOI: 10.2196/46599
13. Specialistexamen – SFAM [Internet]. [cited 2023 Sep 30]. Available from: https://sfam.se/specialistexamen/
14. OpenAI API [Internet]. [cited 2023 Sep 30]. Available from: https://openai.com/blog/openai-api
15. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research [Internet]. 2024 [cited 2024 Mar 2]. Available from: https://CRAN.R-project.org/package=psych
16. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155–63. DOI: 10.1016/j.jcm.2016.02.012
17. Perktold J, Skipper Seabold, Sheppard K, et al. statsmodels/statsmodels: Release 0.10.0 [Internet]. Zenodo; 2023 [cited 2023 Oct 4]. Available from: https://www.statsmodels.org/dev/release/version0.10.htmlDOI: 10.5281/ZENODO.593847
18. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023 Jun 1;183(6):589–96. DOI: 10.1001/jamainternmed.2023.1838
19. Tu T, Palepu A, Schaekermann M, et al. Towards Conversational Diagnostic AI

[Internet]. arXiv; 2024 [cited 2024 Mar 4]. Available from: http://arxiv.org/abs/2401.05654DOI: 10.48550/arXiv.2401.05654

20. Kringos D, Boerma W, Bourgueil Y, et al. The strength of primary care in Europe: an international comparative study. Br J Gen Pract J R Coll Gen Pract. 2013 Nov;63(616):e742-750. DOI: 10.3399/bjgp13X674422
21. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023 Oct 4;25(1):e50638. DOI: 10.2196/50638
22. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Adv Neural Inf Process Syst. 2022 Dec 6;35:24824–37.
23. Wu Q, Bansal G, Zhang J, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation [Internet]. arXiv.org. 2023 [cited 2023 Oct 8]. Available from: https://arxiv.org/abs/2308.08155v2
24. Xi Z, Chen W, Guo X, et al. The Rise and Potential of Large Language Model Based Agents: A Survey [Internet]. arXiv.org. 2023 [cited 2023 Oct 8]. Available from: https://arxiv.org/abs/2309.07864v3
25. Scores of responses by doctors and ChatGPT on the Swedish family medicine specialist exam. Available from [To be published]: https://doris.snd.gu.se/ DOI: https://doi.org/10.5878/j8jh-5128

Supplemental files

File name	**Title**	**Description**
Supplemental file 1	ChatGPT Settings and instructions	Describes the settings used when prompting ChatGPT through OpenAI's API, as well as the instructions (prompt) used when sending the cases to ChatGPT, and a short rationale behind the choice of instructions.
Supplemental file 2	Examples of cases and scoring guides	Three examples of cases and their corresponding scoring guides, translated to English.

ChatGPT Settings and instructions

Supplemental file 1

CHATGPT SETTINGS

These are the settings used when prompting ChatGPT through OpenAI's API. See the API reference for more information on what each of the settings means [1].

Model	gpt-4
Version	August 3 Version, 2023
System message A message sent as the "system user" before the first message between the user and the chatbot, not to be confused with the instructions sent together with the cases. The instructions were sent together with each case as a single message from the user. See the OpenAI API reference and examples for more information about the system message.	You are a helpful assistant.
temperature	0
max_tokens	2048
top_p	1
presence_penalty	0
frequency_penalty	0

Instructions to ChatGPT

```
Below are questions from a test designed to test ChatGPT's ability
to answer medical questions. These are fictitious situations in a
primary health care centre. Please, describe how you would handle
the described situations as a doctor in the context of a primary
health care centre with the resources and conditions that apply
there.

General instructions:
Think about and identify the following
• What is the most significant problem of the case - and why?
• Which other aspects should also be taken into account - justify!
You only act on the facts that are described in the assignment.
Justify your statements, decisions, prescriptions.
Write down your reflections! In case of different options (eg
diagnoses): which is most likely?
If there is any information you are missing: please, explain why!

Here is the first scenario:
"<the case is inserted here>"

Since this is a fictitious test to evaluate the AI:s own medical
ability, I ask you not to refer to talking to a doctor. We
understand that we should not take medical advice from you and
apply in real cases. Please, write the answer just as if you were
a doctor answering an exam!
```

The prompt used when sending the cases to ChatGPT's. This version is translated from Swedish to English, while the Swedish version was used in the study

Rationale behind the choice of prompt

The instructions are mostly a copy of the instructions included in the Swedish Family Medicine Specialist Exam of 2022. They were adapted to be directed towards ChatGPT rather than towards a human doctor. The parts stressing that the cases are fictitious and for evaluation purposes only, were added to discourage ChatGPT from writing disclaimers and referring to seeing an actual doctor. In this study, a 'zero-shot' strategy was employed, meaning that no examples of good answers were provided to ChatGPT. An alternative approach could have involved using a few other cases with 'perfect' answer examples as a guide, which might have steered ChatGPT towards the desired answer style. However, this could potentially limit the model's inherent flexibility. Other research in the domain of testing ChatGPT's performance on medical examinations has often adopted a 'zero-shot' approach [2,3].

References

1. OpenAI API [Internet]. [cited 2023 Sep 30]. Available from: https://openai.com/blog/openai-api

2. 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 8;9:e45312.

3. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. An original study of ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level based on the Dermatology Specialty Certificate Examinations. Clin Exp Dermatol. 2023 Aug 4;llad255.

Example Cases and Scoring Guides

Supplemental file 2

Below are examples of three cases and their corresponding scoring guides. Here, they are translated to English, whereas the original Swedish versions were used in the study.

CASE 1

Bengt, 76 years old, comes for an annual check-up for diabetes and hypertension. You don't know him from before. His wife is accompanying him, and they tell you that he has generally deteriorated in the last year. He is tired and feels weak, has a headache, has a poor appetite, is dizzy and falls easily, coughs a bit, feels sick and has become more forgetful.

Before the doctor's appointment: blood pressure 130/70, height 177 cm, weight 71 kg, pulse 95/minute. Drugs: Metformin 500 mg 2x2, Felodipine 10 mg 1x1, Enalapril 5 mg 1x1. The medicine list also contains Betolvidon 1 mg 1x1 and Folacin 1 mg 1x1, but he is out of these medications.

Lab: HbA1C 34, fasting blood glucose 7.6. ALT 0.36. Creatinine 67. Potassium 4.5. Cholesterol 3.8. Cobalamin 291. Folate 8.

Please note that the units of the blood tests were not stated in the case, but they are expressed in the standard units used in Sweden: HbA1C: mmol/mol, glucose: mmol/L, ALT: µkat/L, Creatinine: µmol/L, Potassium: mmol/L, Cholesterol: mmol/L, Cobalamin: pmol/L, Folate: nmol/L.

SCORING GUIDE FOR CASE 1

- Does the answer mention the need to let them talk and provide more information? (0.5 p)
- Does it prioritise reasonably between the various problems? (0.5p)
- Does it suggest an ECG? (0.75p)
- Does it suggest haemoglobin (Hb)? (0.75p)
- Does it suggest a neurological examination? (0.5p)
- Does it express the need to assess his alcohol consumption? (0.5p)
- Does it mention heart disease/failure as a possible diagnosis? (0.5p)
- Does it mention low blood pressure as a possible diagnosis? (0.5p)
- Does it mention low blood sugar (hypoglycaemia) as a possible diagnosis? (0.5p)
- Does it mention anaemia as a possible diagnosis? (0.5p)
- Does it mention memory loss/dementia as a possible diagnosis? (0.5p)
- Does it mention inactivity as a possible diagnosis? (0.5p)
- Does it suggest a review of his medications? (2p)
- Does it suggest reducing the metformin dose? (0.5p)
- Does it suggest reducing the felodipine dose? (0.5p)
- Does it suggest a trial of reduced enalapril dose? (0.5p)

CASE 2

Filip's parents have scheduled an appointment with you at the clinic because Filip has a poor appetite. He suffers from abdominal pain and intermittent diarrhoea. The symptoms have been present for about a year. Occasionally, he also experiences joint pain. Filip is 11 years old, perceived by his parents as small and thin, and his younger sister is almost as tall as him. His mother is worried about Filip being short for his age and whether his symptoms need further investigation. His father is less concerned, recalling his own late puberty and being the shortest in his class for a long time.

When Filip is asked to describe his symptoms, he points to his parents, but he politely answers a few supplementary questions himself.

Physical examination:

General Appearance: Not noticeably thin, seems energetic and happy. Mouth and throat: normal. Lymph nodes in the neck, axillae, and groin: normal. Thyroid: not enlarged. Heart/lungs: Normal findings. Abdomen: Non-tender without abnormal findings. Joints: No signs of joint disease.

SCORING GUIDE FOR CASE 2

- Does the answer mention gluten intolerance (celiac disease) as a possible diagnosis? (1p)
- Does it mention lactose intolerance as a possible diagnosis? (1p)
- Does it mention inflammatory bowel disease as a possible diagnosis? (1p)
- Does it mention gastrointestinal symptoms related to anxiety and social circumstances as a possible cause? (1p)
- Does it mention constipation as a possible diagnosis? (1p)
- Does it mention the need to ask about diet and bowel habits as part of the consultation? (0.5p)
- Does it mention the need to ask about signs of puberty as part of the consultation? (0.25p)
- Does it mention the need to ask about school and friends as part of the consultation? (0.25p)
- Does the answer suggest testing for glucose levels? (0.5p)
- Does the answer suggest conducting inflammatory marker tests, such as CRP (C-reactive protein)? (0.5p)
- Does the answer suggest measuring weight and height? (0.5p)
- Does the answer suggest conducting a test for gluten intolerance (anti-tissue transglutaminase (TTG) antibody)? (0.5p)
- Does the answer suggest requesting and evaluating the school nurse's records of growth curves as part of the assessment? (1p)
- In the case of normal test results and growth measurements, does the answer suggest reassurance and no further evaluation? (1p)

CASE 3

Astrid, 87 years old, has been living in a special care facility for 2 months, where you are the responsible physician. She had breast cancer two years ago, which was treated with breast resection. Following her recovery, she was able to live independently, without requiring assistance. Follow-up checks were unremarkable, and she was prescribed antiestrogen therapy.

About 3.5 months ago, Astrid experienced numbness and some weakness in her right hand, and an outpatient CT scan was ordered. However, before the scan, her symptoms worsened, including motor aphasia and epileptic seizures which were difficult to control with treatment. The CT scan revealed brain metastases likely originating from the breast tumour. She was started on betamethasone to reduce intracranial swelling, with a plan to gradually reduce the dose. She declined palliative radiation therapy and was offered a place in the care facility. At the first meeting with you there, she was ambulatory, could manage her activities of daily living, understood everything said but could not speak coherently. She exhibited a heightened emotional sensitivity, as is sometimes observed after a brain injury, which prompted a trial of mood-stabilising medication (Citalopram 10 mg x 1).

A week later, Astrid's condition deteriorated, with instability while walking, frequent falls, weakness in her right leg, and confusion.

Upon examination, the only new findings were difficulty rising from bed and a flaccid paralysis in her right arm. She still had motor aphasia.

Astrid's daughter contacts you after your visit to the facility, inquiring about the ongoing care and treatment of her mother.

SCORING GUIDE FOR CASE 3

- Does the response recognize and characterise this as a palliative care situation? (1.5p)
- Does the response suggest a palliative care discussion with Astrid and her daughter? (1.5p)
- Does the response suggest a medication review with an emphasis on discontinuing non-essential medications? (1.5p)
- Does the response analyse and compare the suitability of continued treatment at the care facility versus hospitalisation in this case? (1.5p)
- Does the response emphasise attentive and compassionate nursing care? (1.5p)
- Does the response propose the use of palliative medications, such as those for symptom relief and pain management? (1.5p)
- Does the response mention the possibility of applying for a "caregiver's allowance" (to support individuals caring for a terminally ill family member)? (1p)

BMJ Open

ChatGPT (GPT-4) versus Doctors on Complex Cases of the Swedish Family Medicine Specialist Exam: An Observational Comparative Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-086148.R1
Article Type:	Original research
Date Submitted by the Author:	25-Aug-2024
Complete List of Authors:	Arvidsson, Rasmus; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Praktikertjänst AB, Hälsocentralen Sankt Hans Gunnarsson, Ronny; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Region Vastra Gotaland, Närhälsan, Vårdcentralen Hemlösa Entezarjou, Artin; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Capio Healthcare Centre Angered Sundemo, David; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Lerum Närhälsan Primary Healthcare Center Wikberg, Carl; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Region Vastra Gotaland, Research, Education, Development & Innovation, Primary Health Care
Primary Subject Heading:	General practice / Family practice
Secondary Subject Heading:	Health informatics
Keywords:	Artificial Intelligence, Primary Health Care, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

For peer review only

Enseignement Superieur (ABES) .
Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

ChatGPT (GPT-4) versus Doctors on Complex Cases of the Swedish Family Medicine Specialist Exam: An Observational Comparative Study

Rasmus Arvidsson[1,2], Ronny Gunnarsson[1,3,4], Artin Entezarjou[1,5], David Sundemo[1,6], Carl Wikberg[1,4]

[1]General Practice, School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[2]Hälsocentralen Sankt Hans, Lund, Sweden

[3]Region Vastra Gotaland, Närhälsan, Vårdcentralen Hemlösa

Göteborg, Sweden

[4]Research, Education, Development & Innovation, Primary Health Care, Region Västra Götaland, Sweden

[5]Capio Healthcare Centre Angered, Angered, Sweden

[6]Lerum Närhälsan Primary Healthcare Center, Lerum, Sweden

Corresponding author

Rasmus Arvidsson (rasmusalinks@gmail.com), ORCID 0009-0006-0387-3108

ORCIDs of other authors

Ronny Gunnarsson (0000-0001-9183-3072), Artin Entezarjou (0000-0002-7418-8750), David Sundemo (0000-0002-5871-1636), Carl Wikberg (0000-0002-6494-5922)

ABSTRACT

Background: Recent breakthroughs in AI research include the development of generative pretrained transformers (GPT). ChatGPT has been shown to perform well when answering several sets of medical multiple choice questions. However, it has not been tested on writing free text assessments of complex cases in primary care.

Objectives: To compare the performance of ChatGPT, version GPT-4, with that of real doctors.

Design and Setting: A blinded observational comparative study conducted in the Swedish primary care setting. Responses from GPT-4 and real doctors to cases from the Swedish family medicine specialist exam were scored by blinded reviewers, and the scores compared.

Participants: Anonymous responses from the Swedish family medicine specialist exam 2017–2022 were used.

Outcome measures: Primary: Mean difference in scores between GPT-4's responses and randomly selected responses by human doctors, as well as between GPT-4's responses and top-tier responses by human doctors. Secondary: Correlation between differences in response length and response score. Intraclass correlation coefficient between reviewers.

Results: The mean scores were 6.0, 7.2, and 4.5 for randomly selected doctor responses, top-tier doctor responses, and GPT-4 responses, respectively, on a 10-point scale. The scores for the random doctor responses were, on average, 1.6 points higher than those of GPT-4 ($p < 0.001$, 95 % confidence interval (CI) 0.9–2.2) and the top-tier doctor scores were, on average, 2.7 points higher than those of GPT-4 ($p < 0.001$, 95 % CI 2.2–3.3).

Conclusion: On complex primary care cases, GPT-4 performs worse than human doctors taking the Family Medicine Specialist Exam. Future GPT-based chatbots may perform better, but comprehensive evaluations are needed before implementing chatbots for medical decision support in primary care.

STRENGTHS AND LIMITATIONS OF THIS STUDY

Strengths

- Each response was scored by two independent, blinded reviewers.
- Detailed scoring keys provided excellent interrater correlation.
- Evaluation of long-form free text responses to complex cases, relevant for primary care.

Limitations

- The result may not be fully generalisable to other countries and languages.
- There was no penalty for the presence of extra, unnecessary, information in the responses when scoring.

KEYWORDS

Artificial Intelligence

Primary Health Care

Medical Informatics

Technology Assessment, Biomedical

BACKGROUND

Artificial intelligence (AI) in medicine has been the subject of increasing research, even though real-world applications are relatively few [1–3]. Over the last few years, large AI models called generative pretrained transformers (GPT) have demonstrated remarkable abilities beyond simple text generation, such as answering questions and participating in chat conversations. ChatGPT from OpenAI is arguably one of the most well-known models. At the time of this study, their two latest AI models are GPT-3.5 and GPT-4, with GPT-4 being the most advanced.

Countless clinical applications could be envisioned for an AI system that can accurately answer questions from healthcare staff and patients. The impact could be enormous in primary health care, where healthcare staff need to keep themselves up-to-date on a broad spectrum of medical conditions.

GPT-3.5 and GPT-4 have demonstrated human-level performance on several professional benchmarks [4] and achieved moderate to excellent results in various medical exams [5–10] but did not pass the general practice licensing exams of Taiwan and the United Kingdom (UK) [11,12]. However, the medical questions in these assessments have typically been multiple-choice questions, which differ from a clinician asking the chatbot for advice on managing real patient cases. Additionally, the studies focusing on general practice have tested GPT-3.5, which may perform significantly worse than GPT-4 [6,9]. At the time of writing, research has not explored GPT-4's ability to provide comprehensive free text assessments of primary care cases.

The Swedish family medicine specialist exam is not mandatory, but it is a valuable credential taken by resident doctors in general medicine as they become certified specialists. One part of the exam is a written test with eight complex cases that often involve intricate symptoms combined with social or behavioural factors, requiring comprehensive long-form responses. Our research question investigates how GPT-4 performs in comparison to real doctors taking the exam.

METHODS

Study Design

This study compared the performance of GPT-4 with responses from human doctors on cases from the Swedish family medicine specialist exam. The responses from three distinct groups were scored and compared: A) randomly selected doctor responses, B) top-tier doctor responses, and C) responses generated by GPT-4.

Objective and outcome measures

The objective was to compare GPT-4 to real doctors, regarding their ability to write comprehensive assessments of complex cases from primary care.

Primary outcome measure: Mean difference in scores between GPT-4 and randomly selected responses by human doctors, as well as between GPT-4 and top-tier responses.

Secondary outcome measures: Correlation between differences in response length and response score. Intraclass correlation coefficient between reviewers.

Data collection

Sourcing of medical cases

All cases from the Swedish family medicine specialist exam from 2017 to 2022 were utilised for this study, totalling 48 cases [See supplemental file 1 for examples]. These exams are publicly available on the Swedish Association of General Practice (SFAM)'s website [13]. The cases require comprehensive responses, typically consisting of several paragraphs of free text. They are often complex, involving symptoms indicative of various diseases and complicating factors such as social problems, addiction, poor compliance, legal aspects, and patients near the end of life. Table 1 provides a summary of the number of cases addressing different topics.

Table 1: Number of Cases Addressing Different Topics

Topic	Case count	Topic	Case count
Diagnostics	43	Multi-morbidity/ Polypharmacy	6
Treatment	37	Well-Child Care	6
Communication	17	Prioritisation/ Level of Care	6
Lifestyle/Social Factors	15	Dissatisfied Patients	5
Red Flags for Serious Diseases	8	Sickness Benefit	4

Overmedication/Side Effects	7	Legal Matters, Addiction, Miscellaneous	15

Sourcing of doctor responses, group A and B

Anonymous responses from past exams were used. Group A: A digital random choice function was used to draw a single anonymous response for each case, from all the human responses submitted to the exam when it took place. Group B: The Swedish Association of General Practice, SFAM, has published an example of a top-tier response for each case. These responses were chosen arbitrarily by the exam reviewers as the best response for each question, in their opinion, and were used for Group B [13].

Obtaining GPT-4 responses, group C

Medical cases were sent to GPT-4 in an automated manner through OpenAI's application programming interface (API) [14], using the version of GPT-4 released on August 3, 2023. Apart from the case itself, additional instructions were sent along with each case to provide some context, based on the written instructions included in the 2022 exam [See supplemental file 2]. A single response was collected for each case, without any follow-up questions [See supplemental file 1]. A separate chat session was created for each case.

Scoring the responses

For each case, SFAM has published an evaluation guide that includes a few main points which should be included in a good answer, although the precise scoring guide used for the exam is not public. To quantify the performance of each exam response, the published evaluation guide for each case was adapted into a criteria-based scoring guide. Each scoring guide could award a score ranging from 0 to 10 points. This adaptation involved rephrasing each evaluation guide into a set of true-or-false criteria. The original evaluation guide was followed as closely as possible, but in cases where it was vaguely phrased, official Swedish medical guidelines were consulted to help formulate the criteria. For each criterion met, a specific number of points was awarded [See supplemental file 1]. A group of three medical doctors, blinded to the origins of the responses, rated the responses using the scoring guide. Each response was scored by two of the three raters, and the average of their scores was used for the statistical analysis. The same pair of raters assessed all responses pertaining to the same case. The doctor creating the scoring guide is a specialist in general practice, whereas two of the reviewers are residents nearing the end of their residency, and one is a licensed doctor working in general practice. The evaluators were selected based on their expertise and availability.

Statistical analysis

Sample size calculation

In the primary research question, we aimed to make two group comparisons, each producing a p-value. Using the Bonferroni approach to adjust for multiple testing, the level of significance was set to 0.025. The power was set to 0.8 and minimal difference between groups to be detected was set to one point, which resulted in a required sample size of 48 cases.

Data Analysis

After scoring the responses to all 48 cases, the difference between each doctor group and GPT-4 was calculated for each case. A paired t-test was used to compare each doctor group with GPT-4, pairing the scores by question.

To assess the reliability of the averaged scores derived from the raters' use of the scoring guide, we conducted an Intraclass Correlation Coefficient (ICC) analysis, specifically employing the Two-Way Mixed-Effects Model for the mean of k raters, utilising the psych package in R [15,16].

In addition, we examined the differences in response length (number of words) between the top-tier and GPT-4 responses. A paired t-test was used paired by question. As a measure of the information density, we divided the score with the number of words for each response. Finally, a linear regression analysis was performed to explore the relationship between the difference in lengths and the difference in scores. The latter was set as the dependent variable and the former as the independent variable. The OLS function from the statsmodels library was employed for this analysis [17].

RESULTS

GPT-4 scored lower than any doctor group (Table 2). The differences between groups were statistically significant (Table 3). For examples of responses, see supplemental file 1. The complete scores are available in a public repository [18].

Table 2: Mean score, length and points per 100 words of each group

Group	Mean score	Mean length (words)	Points per 100 words
A: Random doctor	6.0 (5.5–6.6)	*320 (280–350)*	*2.2 (1.9–2.5)*
B: Top-tier doctor	7.2 (6.8–7.6)	350 (320–380)	2.3 (2.0–2.6)
C: GPT-4	4.5 (3.9–5.0)	290 (280–300)	1.6 (1.4–1.7)

95 % confidence intervals (CI) in parenthesis. The points per 100 words is the average of the points per 100 words for each case.

Table 3: Differences in scores between GPT-4 and the doctor groups

	Mean difference (95 % CI)	P-value
A vs C: Random doctor versus GPT-4	1.6 (0.9–2.2)	< 0.001
B vs C: Top-tier doctor versus GPT-4	2.7 (2.2–3.3)	< 0.001
A vs B: Top-tier vs random doctor	1.2 (0.7–1.7)	< 0.001

The ICC for the scores from the three raters was 0.92 (95% CI 0.90–0.94, p < 0.001), demonstrating excellent reliability of the scoring guide.

The top-tier responses were on average 60 words longer than GPT-4's (p < 0.001, 95 % CI 30–97). The correlation between differences in length and differences in scores of responses between GPT-4 and the top-tier answers was not statistically significant (p = 0.11).

DISCUSSION

The main finding was that GPT-4 scored significantly lower than any group of doctors on the Swedish family medicine specialist exam, with top-tier responses scoring almost three points higher (Table 3). This statistically significant difference indicates that graduating specialists in general practice perform better than GPT-4 in writing comprehensive assessments of complex primary care cases.

What such a difference corresponds to in practice differs a lot from case to case. For example, in one case, GPT-4 scored 2.75 points lower than the top-tier response due to mentioning one fewer important differential diagnosis and two fewer aspects of treatment and follow-up. Notably, the information density was also higher for the two doctor groups than for GPT-4, indicating that human doctors are better at conveying relevant information concisely. Despite these limitations, ChatGPT's performance is impressive, considering it is not a registered medical device and has not been specifically trained for medical use.

There was also a significant difference between the top-tier and randomly selected doctor responses, raising the question of what requirements should be met by a medical chatbot. Is it enough for it to perform better than the average doctor, or should it aim to match or exceed the best responses from a group of doctors?

Comparison with the existing literature

In one study, GPT-4 passed every test in a series of dermatology licensing exams, achieving over 80% for the English version (pass level: 60 %) [6]. No data were presented on the performance of real dermatologists for comparison. On the other hand, the average score of GPT-3.5 was only 60.17% on the general practice licensing exam of the UK (pass level ≈ 70%) [12], and it scored 41.6% on the corresponding Taiwanese licensing exam (pass level = 60%) [11]. This aligns well with our results, even though we used GPT-4. Both these studies, and several similar studies in other medical disciplines [7–10], used multiple choice questions, which is a task very different from providing free text responses to complex clinical cases. Providing free text answers more closely resembles the requirements of a chatbot used for decision support in clinical practice. Many used GPT-3.5, which may perform significantly worse than GPT-4.

One study examined questions posted by patients online, on a forum where volunteering doctors responded [19]. In the study, three licensed health care professionals evaluated the free text responses. In 79% of the cases, they favoured GPT-3.5 responses over the doctors and the quality score was 21 % lower for doctors on average, as scored on a 5-category ordinal scale. These findings are opposite to the findings of our study, where the randomly selected doctors' responses scored higher in 71% of the cases, even though GPT-4 was used. The questions and responses in the patient forum were typically shorter and simpler than the primary care cases used in our study, and the responses were not assessed on specific medical criteria. In a recent preprint, a novel chatbot AI, named AMIE, has been fine tuned to perform a diagnostic interview with a patient through chat [20]. It was compared with general practitioners on objective structured clinical examination (OSCE) cases and outperformed general practitioners on most metrics, including suggesting relevant differential diagnoses. This suggests that higher performance is already possible from AI models, but evaluating GPT-4 is still highly relevant, since it is widely accessible and may hypothetically already be used by patients and clinicians.

Strengths and limitations

This is the first study of GPT-4 performance on complex primary care cases with long-form free text responses, rather than multiple choice. As such, it mimics the scenario where a clinician posts a case summary of a real patient in order to get input on the management. The scoring system was a relatively clear way to quantify the amount of useful content in each answer and demonstrated excellent reliability. No penalty was given to superfluous content, however, which could favour respondents writing longer, but less relevant, responses. The cases used in our study are representative for Swedish primary care, which may differ somewhat from other countries [21]. This should be taken into account when generalising our results to other countries.

The set of instructions sent to GPT-4 with each case, sometimes called the "prompt," may influence the quality of responses [22]. This is its own area of research, and optimising the prompt was beyond the scope of this study (see supplemental file 2 for the rationale behind our choice of prompt). The cases utilised in the study are publicly available online and could have been part of GPT-4's training data, but the correct answers are not available in direct association with the questions, so we find it unlikely that this would have affected the result. In some cases, the reviewers could guess which answer was written by GPT-4, which may have introduced some bias. However, the impact of this bias was likely reduced by the use of the scoring guide, which focused on the presence and absence of specific criteria rather than an overall subjective assessment of the answer quality.

Implications for current practice and future research

GPT-4 falls short in medical accuracy when writing comprehensive assessments of complex primary care cases, compared to human doctors. The difference in performance is both statistically significant and clinically relevant. Hence, case assessments by GPT-4, should not be used directly by primary care doctors. Nor should GPT-4 be implemented as a doctor or nurse substitute for patients. However, new versions and models can be expected to arise.

Future research on medical chatbots should focus on evaluating emerging models on representative questions asked by clinicians and patients in a clinical setting. At the same time, in line with the previously mentioned AMIE medical chatbot [20], researchers and developers should aim to optimise the performance of such chatbots, for example by training them specifically on reliable medical information, optimising prompt engineering techniques [22,23], using algorithms for processing a single question in multiple steps or allowing the chatbots access to external sources of information and tools, including other categories of AI-models [24,25]. If reliable medical chatbots are developed, they could profoundly impact general practice. Initial contact, triage, and management of simple cases could conceivably be handled directly by a medical chatbot. Additionally, these chatbots could serve as constantly available expert advisors for medical staff.

DECLARATIONS

Ethics approval

Responses from human participants of the Swedish family medicine specialist exam were obtained anonymously without an identifier linking responses to individuals. According to Swedish regulations, collecting this type of data anonymously does not require ethics approval.

Patient involvement statement

There was no patient or public involvement in the development of this study.

Availability of data and materials

The scores are published in the Swedish National Data Service's Data Organisation and Information System repository. Three examples of cases and their corresponding scoring guides and GPT-4 responses have been translated to English and included as supplemental file 1. The original cases, evaluation guides and top-tier responses are publicly available in Swedish at SFAM's website, from where they were used in this study with permission.

Funding

Västra Götaland region, Sweden. Grant no: NA

Competing Interests

None.

Authors' contributions

The study was planned by authors RA, RG, AE, DS and CW. Author RA collected the cases and responses, and compiled the scoring guides. Authors DS and AE participated in the group of raters who scored the responses. Author RA performed the statistical analysis and created the draft of the manuscript. Authors RA, RG, AE, DS and CW participated in discussing results and refining the manuscript. RA is responsible for the overall content as guarantor.

Acknowledgements

Special thanks to The Swedish Association of General Practice, SFAM, for granting permission to use the Swedish family medicine specialist exam. Specifically, Karin Lindhagen was very helpful in compiling the randomly selected responses. Special thanks also to Dr Abed Alsabbagh, who participated in the group reviewing and scoring the responses.

References

1. Niu J, Tang W, Xu F, et al. Global Research on Artificial Intelligence from 1990–2014: Spatially-Explicit Bibliometric Analysis. ISPRS Int J Geo-Inf. 2016 May;5(5):66. DOI: 10.3390/ijgi5050066
2. Yin J, Ngiam KY, Teo HH. Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. J Med Internet Res. 2021 Apr 22;23(4):e25759. DOI: 10.2196/25759
3. Triantafyllidis AK, Tsanas A. Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature. J Med Internet Res. 2019 Apr 5;21(4):e12286. DOI: 10.2196/12286
4. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys Syst. 2023 Jan 1;3:121–54. DOI: 10.1016/j.iotcps.2023.04.003
5. Gilson A, Safranek CW, Huang T, 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 8;9:e45312. DOI: 10.2196/45312

6. Lewandowski M, Łukowicz P, Świetlik D, et al. An original study of ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level based on the Dermatology Specialty Certificate Examinations. Clin Exp Dermatol. 2023 Aug 4;llad255. DOI: 10.1093/ced/llad255
7. Antaki F, Touma S, Milad D, et al. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol Sci. 2023 Dec 1;3(4):100324. DOI: 10.1016/j.xops.2023.100324
8. Kung JE, Marshall C, Gauthier C, et al. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. JBJS Open Access. 2023 Sep;8(3):e23.00056. DOI: 10.2106/JBJS.OA.23.00056
9. Taloni, Andrea, Borselli, Massimiliano, Scarsi, Valentina, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology | Scientific Reports. Sci Rep. 2023 Oct 29;13(18562). DOI: 10.1038/s41598-023-45837-2
10. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: An observational study. Medicine (Baltimore). 2023 Aug 11;102(32):e34673. DOI: 10.1097/MD.0000000000034673
11. Weng TL, Wang YM, Chang S, et al. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc. 2023 Aug;86(8):762. DOI: 10.1097/JCMA.0000000000000946
12. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. 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 Apr 21;9:e46599. DOI: 10.2196/46599
13. Specialistexamen – SFAM [Internet]. [cited 2023 Sep 30]. Available from: https://sfam.se/specialistexamen/
14. OpenAI API [Internet]. [cited 2023 Sep 30]. Available from: https://openai.com/blog/openai-api
15. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research [Internet]. 2024 [cited 2024 Mar 2]. Available from: https://CRAN.R-project.org/package=psych
16. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155–63. DOI: 10.1016/j.jcm.2016.02.012
17. Perktold J, Skipper Seabold, Sheppard K, et al. statsmodels/statsmodels: Release 0.10.0 [Internet]. Zenodo; 2023 [cited 2023 Oct 4]. Available from: https://www.statsmodels.org/dev/release/version0.10.htmlDOI: 10.5281/ZENODO.593847
18. [dataset] Arvidsson R, Gunnarsson R, et al. Scores of responses by doctors and ChatGPT on the Swedish family medicine specialist exam. Swedish National Data Service (SND) - Data Organisation and Information System (DORIS); 2024. Version 1. https://doi.org/10.5878/j8jh-5128
19. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023 Jun 1;183(6):589–96. DOI: 10.1001/jamainternmed.2023.1838
20. Tu T, Palepu A, Schaekermann M, et al. Towards Conversational Diagnostic AI [Internet]. arXiv; 2024 [cited 2024 Mar 4]. Available from: http://arxiv.org/abs/2401.05654DOI: 10.48550/arXiv.2401.05654
21. Kringos D, Boerma W, Bourgueil Y, et al. The strength of primary care in Europe: an international comparative study. Br J Gen Pract J R Coll Gen Pract. 2013 Nov;63(616):e742-750. DOI: 10.3399/bjgp13X674422
22. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023 Oct 4;25(1):e50638. DOI: 10.2196/50638
23. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Adv Neural Inf Process Syst. 2022 Dec 6;35:24824–37.

24. Wu Q, Bansal G, Zhang J, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation [Internet]. arXiv.org. 2023 [cited 2023 Oct 8]. Available from: https://arxiv.org/abs/2308.08155v2
25. Xi Z, Chen W, Guo X, et al. The Rise and Potential of Large Language Model Based Agents: A Survey [Internet]. arXiv.org. 2023 [cited 2023 Oct 8]. Available from: https://arxiv.org/abs/2309.07864v3

Supplemental files

1. **Examples of cases and scoring guides:** Three examples of cases and their corresponding scoring guides and GPT-4 responses, translated to English.
2. **ChatGPT Settings and instructions:** Describes the settings used when prompting ChatGPT through OpenAI's API, as well as the instructions (prompt) used when sending the cases to ChatGPT, and a short rationale behind the choice of instructions.

Example Cases and Scoring Guides

Supplemental file 1

Below are examples of three cases and their corresponding scoring guides and GPT-4 answers. Here, they are translated to English, whereas the original Swedish versions were used in the study. More cases and their top-tier doctor responses are publicly available in Swedish at the web site of The Swedish Association of General Practice (SFAM): https://sfam.se/specialistexamen/

Case 1

Bengt, 76 years old, comes for an annual check-up for diabetes and hypertension. You don't know him from before. His wife is accompanying him, and they tell you that he has generally deteriorated in the last year. He is tired and feels weak, has a headache, has a poor appetite, is dizzy and falls easily, coughs a bit, feels sick and has become more forgetful.

Before the doctor's appointment: blood pressure 130/70, height 177 cm, weight 71 kg, pulse 95/minute. Drugs: Metformin 500 mg 2x2, Felodipine 10 mg 1x1, Enalapril 5 mg 1x1. The medicine list also contains Betolvidon 1 mg 1x1 and Folacin 1 mg 1x1, but he is out of these medications.

Lab: HbA1C 34, fasting blood glucose 7.6. ALT 0.36. Creatinine 67. Potassium 4.5. Cholesterol 3.8. Cobalamin 291. Folate 8.

Please note that the units of the blood tests were not stated in the case, but they are expressed in the standard units used in Sweden: HbA1C: mmol/mol, glucose: mmol/L, ALT: µkat/L, Creatinine: µmol/L, Potassium: mmol/L, Cholesterol: mmol/L, Cobalamin: pmol/L, Folate: nmol/L.

Scoring guide for case 1

- Does the answer mention the need to let them talk and provide more information? (0.5 p)
- Does it prioritise reasonably between the various problems? (0.5p)
- Does it suggest an ECG? (0.75p)
- Does it suggest haemoglobin (Hb)? (0.75p)
- Does it suggest a neurological examination? (0.5p)
- Does it express the need to assess his alcohol consumption? (0.5p)
- Does it mention heart disease/failure as a possible diagnosis? (0.5p)
- Does it mention low blood pressure as a possible diagnosis? (0.5p)
- Does it mention low blood sugar (hypoglycaemia) as a possible diagnosis? (0.5p)
- Does it mention anaemia as a possible diagnosis? (0.5p)
- Does it mention memory loss/dementia as a possible diagnosis? (0.5p)
- Does it mention inactivity as a possible diagnosis? (0.5p)
- Does it suggest a review of his medications? (2p)
- Does it suggest reducing the metformin dose? (0.5p)
- Does it suggest reducing the felodipine dose? (0.5p)
- Does it suggest a trial of reduced enalapril dose? (0.5p)

GPT-4 response to Case 1

The most significant problem in this scenario is Bengt's general deterioration over the past year, which includes fatigue, weakness, headaches, poor appetite, dizziness, falls, coughing, nausea and increased forgetfulness. These symptoms can be signs of several different medical conditions, including heart disease, neurological conditions, lung disease, kidney disease, or even mental health conditions such as depression. It is important to identify the cause of these symptoms in order to provide appropriate treatment.

Other aspects to consider include Bengt's diabetes and hypertension, both of which appear to be under control based on his blood pressure and HbA1c values. His medication should also be reviewed, especially since he has run out of Betolvidone and Folacin. Betolvidon (beta blocker) is used to treat high blood pressure and heart disease, and it can be dangerous to suddenly stop taking it. Folacin (folic acid) is important for cell growth and development, and deficiency can lead to anaemia.

To investigate Bengt's symptoms further, I would like to do a thorough physical examination and possibly order additional laboratory tests. A neurological examination may be appropriate given his dizziness, falls and increased forgetfulness. A chest x-ray may be appropriate given his cough. Additional blood tests may include thyroid-stimulating hormone (TSH) and free T4 to rule out thyroid disease, which can cause fatigue and weakness.

I would also like to discuss Bengt's lifestyle, including diet, exercise, alcohol and tobacco use, as these factors may affect his health. Also, I would like to know more about his mental health, as depression can cause many of the symptoms he is experiencing.

Finally, I would like to follow up with Bengt and his wife regularly to monitor his condition and adjust his treatment if necessary.

Note that GPT-4 interpreted the medicine "Betolvidone" as a beta blocker, when it is in fact the brand name of a vitamin B12 supplement in Sweden.

Case 2

Filip's parents have scheduled an appointment with you at the clinic because Filip has a poor appetite. He suffers from abdominal pain and intermittent diarrhoea. The symptoms have been present for about a year. Occasionally, he also experiences joint pain. Filip is 11 years old, perceived by his parents as small and thin, and his younger sister is almost as tall as him. His mother is worried about Filip being short for his age and whether his symptoms need further investigation. His father is less concerned, recalling his own late puberty and being the shortest in his class for a long time.

When Filip is asked to describe his symptoms, he points to his parents, but he politely answers a few supplementary questions himself.

Physical examination:

General Appearance: Not noticeably thin, seems energetic and happy. Mouth and throat: normal. Lymph nodes in the neck, axillae, and groin: normal. Thyroid: not enlarged. Heart/lungs: Normal findings. Abdomen: Non-tender without abnormal findings. Joints: No signs of joint disease.

Scoring guide for case 2

- Does the answer mention gluten intolerance (celiac disease) as a possible diagnosis? (1p)
- Does it mention lactose intolerance as a possible diagnosis? (1p)
- Does it mention inflammatory bowel disease as a possible diagnosis? (1p)
- Does it mention gastrointestinal symptoms related to anxiety and social circumstances as a possible cause? (1p)
- Does it mention constipation as a possible diagnosis? (1p)
- Does it mention the need to ask about diet and bowel habits as part of the consultation? (0.5p)
- Does it mention the need to ask about signs of puberty as part of the consultation? (0.25p)
- Does it mention the need to ask about school and friends as part of the consultation? (0.25p)
- Does the answer suggest testing for glucose levels? (0.5p)
- Does the answer suggest conducting inflammatory marker tests, such as CRP (C-reactive protein)? (0.5p)
- Does the answer suggest measuring weight and height? (0.5p)
- Does the answer suggest conducting a test for gluten intolerance (anti-tissue transglutaminase (TTG) antibody)? (0.5p)
- Does the answer suggest requesting and evaluating the school nurse's records of growth curves as part of the assessment? (1p)
- In the case of normal test results and growth measurements, does the answer suggest reassurance and no further evaluation? (1p)

GPT-4 response to Case 2

We have both Filip’s persistent stomach issues and loss of appetite, as well as his parents’ concerns about his growth and development. It is important to take these concerns seriously and investigate them thoroughly to rule out any potential medical conditions that might be underlying them.

One aspect to consider is Filip’s family history. The father's comment about his own late puberty might be relevant, as growth and development often follow family patterns. However, this should not be used as an excuse not to investigate Filip’s symptoms.

Another aspect to consider is Filip’s joint pain. Even though there are no signs of joint changes in his status, joint pain can be a symptom of various conditions, including inflammatory bowel diseases such as Crohn’s disease or ulcerative colitis, which can also cause stomach issues and affect growth.

To further investigate Filip’s symptoms, I would suggest the following:

1. Blood tests to check his general health, including blood cells, liver and kidney function, inflammation, and nutritional status. Specific tests for celiac disease and inflammatory bowel diseases may also be appropriate, given his symptoms.
2. Stool samples to check for signs of infection or inflammation in the intestines.
3. Growth curve and pubertal development should be closely monitored. If there are deviations, it might be appropriate to refer Filip to a pediatric endocrinologist for further investigation.
4. A dietitian can be helpful in evaluating Filip’s diet and providing advice on how to increase his appetite and nutritional intake.

It is important to communicate clearly with Filip’s parents and involve Filip himself in discussions about his health and any further investigations. This can help alleviate concerns and ensure that everyone agrees on the plan going forward.

Case 3

Astrid, 87 years old, has been living in a special care facility for 2 months, where you are the responsible physician. She had breast cancer two years ago, which was treated with breast resection. Following her recovery, she was able to live independently, without requiring assistance. Follow-up checks were unremarkable, and she was prescribed antiestrogen therapy.

About 3.5 months ago, Astrid experienced numbness and some weakness in her right hand, and an outpatient CT scan was ordered. However, before the scan, her symptoms worsened, including motor aphasia and epileptic seizures which were difficult to control with treatment. The CT scan revealed brain metastases likely originating from the breast tumour. She was started on betamethasone to reduce intracranial swelling, with a plan to gradually reduce the dose. She declined palliative radiation therapy and was offered a place in the care facility. At the first meeting with you there, she was ambulatory, could manage her activities of daily living, understood everything said but could not speak coherently. She exhibited a heightened emotional sensitivity, as is sometimes observed after a brain injury, which prompted a trial of mood-stabilising medication (Citalopram 10 mg x 1).

A week later, Astrid's condition deteriorated, with instability while walking, frequent falls, weakness in her right leg, and confusion.

Upon examination, the only new findings were difficulty rising from bed and a flaccid paralysis in her right arm. She still had motor aphasia.

Astrid's daughter contacts you after your visit to the facility, inquiring about the ongoing care and treatment of her mother.

Scoring guide for case 3

- Does the response recognize and characterise this as a palliative care situation? (1.5p)
- Does the response suggest a palliative care discussion with Astrid and her daughter? (1.5p)
- Does the response suggest a medication review with an emphasis on discontinuing non-essential medications? (1.5p)
- Does the response analyse and compare the suitability of continued treatment at the care facility versus hospitalisation in this case? (1.5p)
- Does the response emphasise attentive and compassionate nursing care? (1.5p)
- Does the response propose the use of palliative medications, such as those for symptom relief and pain management? (1.5p)
- Does the response mention the possibility of applying for a “caregiver's allowance” (to support individuals caring for a terminally ill family member)? (1p)

GPT-4 response to Case 3

The first thing that comes to mind is that Astrid has progressive neurological symptoms due to brain metastases from her previous breast cancer. These symptoms include numbness and weakness in her right hand, motor aphasia, epileptic seizures, and now also weakness in her right leg and confusion. These symptoms affect her ability to perform daily activities and her quality of life.

Other aspects to consider include:

1. Astrid's mental health: She has shown signs of "heightened emotional sensitivity," which could be a symptom of brain damage. She has also started taking citalopram, a mood-stabilising medication. It is important to monitor her mental health and adjust her treatment as needed.
2. Astrid's physical health: She has become increasingly weak and has difficulty getting out of bed. It is important to ensure she receives the physical assistance she needs to prevent falls and injuries.
3. Astrid's wishes: She has declined palliative radiotherapy. It is important to respect her decision and focus on providing her with the best possible quality of life.
4. Communication with the family: Astrid's daughter is worried and wants to know more about her mother's care. It is important to maintain open and honest communication with the family and provide them with the information and support they need.

To manage this situation, I would first reassure Astrid's daughter that we are doing everything we can to take care of Astrid and provide her with the best possible quality of life. I would explain that Astrid's symptoms are due to her brain metastases and that we are focusing on alleviating these symptoms and improving her quality of life. I would also explain that we are monitoring Astrid's mental health and adjusting her treatment as needed.

I would also discuss the possibility of involving a palliative care team to provide additional support and expertise. These teams can help with pain relief, symptom management, psychological support, and spiritual care.

Finally, I would assure Astrid's daughter that we will keep her informed of any changes in Astrid's condition and that we are always available to answer questions and discuss concerns.

ChatGPT Settings and instructions

ChatGPT settings

Supplemental file 2

These are the settings used when prompting ChatGPT through OpenAI's API. See the API reference for more information on what each of the settings means [1].

Model	gpt-4
Version	August 3 Version, 2023
System message A message sent as the "system user" before the first message between the user and the chatbot, not to be confused with the instructions sent together with the cases. The instructions were sent together with each case as a single message from the user. See the OpenAI API reference and examples for more information about the system message.	You are a helpful assistant.
temperature	0
max_tokens	2048
top_p	1
presence_penalty	0
frequency_penalty	0

Instructions to ChatGPT

Below are questions from a test designed to test ChatGPT's ability to answer medical questions. These are fictitious situations in a primary health care centre. Please, describe how you would handle the described situations as a doctor in the context of a primary health care centre with the resources and conditions that apply there.

General instructions:

Think about and identify the following

- What is the most significant problem of the case - and why?
- Which other aspects should also be taken into account - justify!

You only act on the facts that are described in the assignment.

Justify your statements, decisions, prescriptions.

Write down your reflections! In case of different options (eg diagnoses): which is most likely?

If there is any information you are missing: please, explain why!

Here is the first scenario:

"<the case is inserted here>"

Since this is a fictitious test to evaluate the AI:s own medical ability, I ask you not to refer to talking to a doctor. We understand that we should not take medical advice from you and apply in real cases. Please, write the answer just as if you were a doctor answering an exam!

The prompt used when sending the cases to ChatGPT's. This version is translated from Swedish to English, while the Swedish version was used in the study

Rationale behind the choice of prompt

The instructions are mostly a copy of the instructions included in the Swedish Family Medicine Specialist Exam of 2022. They were adapted to be directed towards ChatGPT rather than towards a human doctor. The parts stressing that the cases are fictitious and for evaluation purposes only, were added to discourage ChatGPT from writing disclaimers and referring to seeing an actual doctor. In this study, a 'zero-shot' strategy was employed, meaning that no examples of good answers were provided to ChatGPT. An alternative approach could have involved using a few other cases with 'perfect' answer examples as a guide, which might have steered ChatGPT towards the desired answer style. However, this could potentially limit the model's inherent flexibility. Other research in the domain of testing ChatGPT's performance on medical examinations has often adopted a 'zero-shot' approach [2,3].

References

1. OpenAI API [Internet]. [cited 2023 Sep 30]. Available from: https://openai.com/blog/openai-api
2. 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 8;9:e45312.
3. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. An original study of ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level based on the Dermatology Specialty Certificate Examinations. Clin Exp Dermatol. 2023 Aug 4;llad255.

BMJ Open

ChatGPT (GPT-4) versus Doctors on Complex Cases of the Swedish Family Medicine Specialist Exam: An Observational Comparative Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2024-086148.R2
Article Type:	Original research
Date Submitted by the Author:	13-Nov-2024
Complete List of Authors:	Arvidsson, Rasmus; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Praktikertjänst AB, Hälsocentralen Sankt Hans Gunnarsson, Ronny; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Region Vastra Gotaland, Närhälsan, Vårdcentralen Hemlösa Entezarjou, Artin; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Capio Healthcare Centre Angered Sundemo, David; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Lerum Närhälsan Primary Healthcare Center Wikberg, Carl; University of Gothenburg Institute of Medicine, General Practice / Family Medicine, School of Public Health and Community Medicine, Sahlgrenska Academy; Region Vastra Gotaland, Research, Education, Development & Innovation, Primary Health Care
Primary Subject Heading:	General practice / Family practice
Secondary Subject Heading:	Health informatics
Keywords:	Artificial Intelligence, Primary Health Care, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

ChatGPT (GPT-4) versus Doctors on Complex Cases of the Swedish Family Medicine Specialist Exam: An Observational Comparative Study

Rasmus Arvidsson[1,2], Ronny Gunnarsson[1,3,4], Artin Entezarjou[1,5], David Sundemo[1,6], Carl Wikberg[1,4]

[1]General Practice, School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[2]Hälsocentralen Sankt Hans, Lund, Sweden

[3]Region Vastra Gotaland, Närhälsan, Vårdcentralen Hemlösa

Göteborg, Sweden

[4]Research, Education, Development & Innovation, Primary Health Care, Region Västra Götaland, Sweden

[5]Capio Healthcare Centre Angered, Angered, Sweden

[6]Lerum Närhälsan Primary Healthcare Center, Lerum, Sweden

Corresponding author

Rasmus Arvidsson (rasmusalinks@gmail.com), ORCID 0009-0006-0387-3108

ORCIDs of other authors

Ronny Gunnarsson (0000-0001-9183-3072), Artin Entezarjou (0000-0002-7418-8750), David Sundemo (0000-0002-5871-1636), Carl Wikberg (0000-0002-6494-5922)

ABSTRACT

Background: Recent breakthroughs in AI research include the development of generative pretrained transformers (GPT). ChatGPT has been shown to perform well when answering several sets of medical multiple choice questions. However, it has not been tested on writing free text assessments of complex cases in primary care.

Objectives: To compare the performance of ChatGPT, version GPT-4, with that of real doctors.

Design and Setting: A blinded observational comparative study conducted in the Swedish primary care setting. Responses from GPT-4 and real doctors to cases from the Swedish family medicine specialist exam were scored by blinded reviewers, and the scores compared.

Participants: Anonymous responses from the Swedish family medicine specialist exam 2017–2022 were used.

Outcome measures: Primary: the mean difference in scores between GPT-4's responses and randomly selected responses by human doctors, as well as between GPT-4's responses and top-tier responses by human doctors. Secondary: the correlation between differences in response length and response score; the intraclass correlation coefficient between reviewers; and the percentage of maximum score achieved by each group in different subject categories.

Results: The mean scores were 6.0, 7.2, and 4.5 for randomly selected doctor responses, top-tier doctor responses, and GPT-4 responses, respectively, on a 10-point scale. The scores for the random doctor responses were, on average, 1.6 points higher than those of GPT-4 ($p < 0.001$, 95 % confidence interval (CI) 0.9–2.2) and the top-tier doctor scores were, on average, 2.7 points higher than those of GPT-4 ($p < 0.001$, 95 % CI 2.2–3.3). Following the release of GPT-4o, the experiment was repeated, albeit this time with only a single reviewer scoring the answers. In this follow-up, random doctor responses were scored 0.7 points higher than those of GPT-4o ($p = 0.044$).

Conclusion: On complex primary care cases, GPT-4 performs worse than human doctors taking the Family Medicine Specialist Exam. Future GPT-based chatbots may perform

better, but comprehensive evaluations are needed before implementing chatbots for medical decision support in primary care.

STRENGTHS AND LIMITATIONS OF THIS STUDY

Strengths

- Each response was scored by two independent, blinded reviewers.
- Detailed scoring keys provided excellent interrater correlation.
- Evaluation of long-form free text responses to complex cases, relevant for primary care.

Limitations

- The result may not be fully generalisable to other countries and languages.
- There was no penalty for the presence of extra, unnecessary, information in the responses when scoring.

KEYWORDS

Artificial Intelligence

Primary Health Care

Medical Informatics

Technology Assessment, Biomedical

BACKGROUND

Artificial intelligence (AI) in medicine has been the subject of increasing research, even though real-world applications are relatively few [1–3]. Over the last few years, large AI models called generative pretrained transformers (GPT) have demonstrated remarkable abilities beyond simple text generation, such as answering questions and participating in chat conversations. ChatGPT from OpenAI is arguably one of the most well-known models. At the time of this study, their two latest AI models are GPT-3.5 and GPT-4, with GPT-4 being the most advanced.

Countless clinical applications could be envisioned for an AI system that can accurately answer questions from healthcare staff and patients. The impact could be enormous in primary health care, where healthcare staff need to keep themselves up-to-date on a broad spectrum of medical conditions.

GPT-3.5 and GPT-4 have demonstrated human-level performance on several professional benchmarks [4] and achieved moderate to excellent results in various medical exams [5–10] but did not pass the general practice licensing exams of Taiwan and the United Kingdom (UK) [11,12]. However, the medical questions in these assessments have typically been multiple-choice questions, which differ from a clinician asking the chatbot for advice on managing real patient cases. Additionally, the studies focusing on general practice have tested GPT-3.5, which may perform significantly worse than GPT-4 [6,9]. At the time of writing, research has not explored GPT-4's ability to provide comprehensive free text assessments of primary care cases.

The Swedish family medicine specialist exam is not mandatory, but it is a valuable credential taken by resident doctors in general medicine as they become certified specialists. One part of the exam is a written test with eight complex cases that often involve intricate symptoms combined with social or behavioural factors, requiring comprehensive long-form responses. Our research question investigates how GPT-4 performs in comparison to real doctors taking the exam.

METHODS

Study Design

This study compared the performance of GPT-4 with responses from human doctors on cases from the Swedish family medicine specialist exam. The responses from three distinct groups were scored and compared: A) randomly selected doctor responses, B) top-tier doctor responses, and C) responses generated by GPT-4.

Objective and outcome measures

The objective was to compare GPT-4 to real doctors, regarding their ability to write comprehensive assessments of complex cases from primary care.

Primary outcome measure: the mean difference in scores between GPT-4 and randomly selected responses by human doctors, as well as between GPT-4 and top-tier responses.

Secondary outcome measures: the correlation between differences in response length and response score; the intraclass correlation coefficient between reviewers; and the percentage of maximum score achieved by each group in different subject categories.

Data collection

Sourcing of medical cases

All cases from the Swedish family medicine specialist exam from 2017 to 2022 were utilised for this study, totalling 48 cases [See supplemental file 1 for examples]. These exams are publicly available on the Swedish Association of General Practice (SFAM)'s website [13]. The cases require comprehensive responses, typically consisting of several paragraphs of free text. They are often complex, involving symptoms indicative of various diseases and complicating factors such as social problems, addiction, poor compliance, legal aspects, and patients near the end of life. Table 1 provides a summary of the number of cases addressing different topics.

Table 1: Number of Cases Addressing Different Topics

Topic	Case count	Topic	Case count
Diagnostics	43	Multi-morbidity/ Polypharmacy	6
Treatment	37	Well-Child Care	6
Communication	17	Prioritisation/ Level of Care	6
Lifestyle/Social Factors	15	Dissatisfied Patients	5
Red Flags for Serious Diseases	8	Sickness Benefit	4

Overmedication/Side Effects	7	Legal Matters, Addiction, Miscellaneous	15

Sourcing of doctor responses, group A and B

Anonymous responses from past exams were used. Group A: A digital random choice function was used to draw a single anonymous response for each case, from all the human responses submitted to the exam when it took place. Group B: The Swedish Association of General Practice, SFAM, has published an example of a top-tier response for each case. These responses were chosen arbitrarily by the exam reviewers as the best response for each question, in their opinion, and were used for Group B [13].

Obtaining GPT-4 responses, group C

Medical cases were sent to GPT-4 in an automated manner through OpenAI's application programming interface (API) [14], using the version of GPT-4 released on August 3, 2023. Apart from the case itself, additional instructions were sent along with each case to provide some context, based on the written instructions included in the 2022 exam [See supplemental file 2]. A single response was collected for each case, without any follow-up questions [See supplemental file 1]. A separate chat session was created for each case.

Scoring the responses

For each case, SFAM has published an evaluation guide that includes a few main points which should be included in a good answer, although the precise scoring guide used for the exam is not public. To quantify the performance of each exam response, the published evaluation guide for each case was adapted into a criteria-based scoring guide. Each scoring guide could award a score ranging from 0 to 10 points. This adaptation involved rephrasing each evaluation guide into a set of true-or-false criteria. The original evaluation guide was followed as closely as possible, but in cases where it was vaguely phrased, official Swedish medical guidelines were consulted to help formulate the criteria. For each criterion met, a specific number of points was awarded [See supplemental file 1]. A group of three medical doctors, blinded to the origins of the responses, rated the responses using the scoring guide. Each response was scored by two of the three raters, and the average of their scores was used for the statistical analysis. The same pair of raters assessed all responses pertaining to the same case. The doctor creating the scoring guide is a specialist in general practice, whereas two of the reviewers are residents nearing the end of their residency, and one is a licensed doctor working in general practice. The evaluators were selected based on their expertise and availability.

During the review process for this paper, OpenAI released GPT-4o, its latest flagship model. The experiment was subsequently repeated to include responses from GPT-4o. Due to limited availability, it was not possible to reassemble the original panel of evaluators; instead, a single evaluator scored the responses across all groups, including the new GPT-4o group.

Statistical analysis

Sample size calculation

In the primary research question, we aimed to make two group comparisons, each producing a p-value. Using the Bonferroni approach to adjust for multiple testing, the level of significance was set to 0.025. The power was set to 0.8 and minimal difference between groups to be detected was set to one point, which resulted in a required sample size of 48 cases.

Data Analysis

After scoring the responses to all 48 cases, the difference between each doctor group and GPT-4 was calculated for each case. A paired t-test was used to compare each doctor group with GPT-4, pairing the scores by question.

To assess the reliability of the averaged scores derived from the raters' use of the scoring guide, we conducted an Intraclass Correlation Coefficient (ICC) analysis, specifically employing the Two-Way Mixed-Effects Model for the mean of k raters, utilising the psych package in R [15,16].

In addition, we examined the differences in response length (number of words) between the top-tier and GPT-4 responses. A paired t-test was used paired by question. As a measure of the information density, we divided the score with the number of words for each response. Finally, a linear regression analysis was performed to explore the relationship between the difference in lengths and the difference in scores. The latter was set as the dependent variable and the former as the independent variable. The OLS function from the statsmodels library was employed for this analysis [17].

Each individual true-or-false scoring criterion was assigned to a category by the author RA, such as ‘suggest diagnosis’ for points awarded for mentioning a possible diagnosis, and ‘patient history inquiry’ for points awarded for mentioning questions that should be asked of the patient. For more details and definitions of the categories, see supplemental file 3. The top nine most common categories were utilised, and the rest were grouped under 'other'. These categories were then used to compare performance across different subject areas. For each category, we calculated the maximum score and the percentage of that score achieved by each group. The Wilcoxon signed-rank test was used to assess significance in the difference between top-tier and random doctor responses, as well as between GPT-4 and random doctor responses, using the differences in scores paired by scoring criteria.

RESULTS

GPT-4 scored lower than any doctor group (Table 2). The differences between groups were statistically significant (Table 3). For examples of responses, see supplemental file 1. The complete scores are available in a public repository [18].

Table 2: Mean score, length and points per 100 words of each group

Group	Mean score	Mean length (words)	Points per 100 words
A: Random doctor	6.0 (5.5–6.6)	*320 (280–350)*	*2.2 (1.9–2.5)*
B: Top-tier doctor	7.2 (6.8–7.6)	350 (320–380)	2.3 (2.0–2.6)
C: GPT-4	4.5 (3.9–5.0)	290 (280–300)	1.6 (1.4–1.7)

95 % confidence intervals (CI) in parenthesis. The points per 100 words is the average of the points per 100 words for each case.

Table 3: Differences in scores between GPT-4 and the doctor groups

	Mean difference (95 % CI)	P-value
A vs C: Random doctor versus GPT-4	1.6 (0.9–2.2)	< 0.001
B vs C: Top-tier doctor versus GPT-4	2.7 (2.2–3.3)	< 0.001
A vs B: Top-tier vs random doctor	1.2 (0.7–1.7)	< 0.001

The ICC for the scores from the three raters was 0.92 (95% CI 0.90–0.94, p < 0.001), demonstrating excellent reliability of the scoring guide.

The results of the repeated experiment with GPT-4o are not included in the above tables, as a single evaluator scored all groups, making these scores not directly comparable with the original results. However, the original findings were confirmed. Additionally, GPT-4o scored an average of 0.7 points higher than GPT-4 (p = 0.024), though random doctor responses continued to outperform GPT-4o, with an average of 0.7 points higher (p = 0.044).

The top-tier responses were on average 60 words longer than GPT-4's (p < 0.001, 95 % CI 30–97). The correlation between differences in length and differences in scores of responses between GPT-4 and the top-tier answers was not statistically significant (p = 0.11).

The percentage of the total maximum score for each subject category achieved by each group is illustrated in Figure 1. More details about the definition of each category, as well as illustrative examples, are available in supplemental file 3.

DISCUSSION

The main finding was that GPT-4 scored significantly lower than any group of doctors on the Swedish family medicine specialist exam, with top-tier responses scoring almost three points higher (Table 3). This statistically significant difference indicates that graduating specialists in general practice perform better than GPT-4 in writing comprehensive assessments of complex primary care cases.

What such a difference corresponds to in practice differs a lot from case to case. For example, in one case, GPT-4 scored 2.75 points lower than the top-tier response due to mentioning one fewer important differential diagnosis and two fewer aspects of treatment and follow-up. Generally, it appears that GPT-4 significantly lags behind the random doctor responses in critical areas such as suggesting relevant diagnoses, lab tests, physical examinations, referrals, and addressing legal matters. For any general practitioners currently using GPT-4, this finding is concerning, as these are precisely the areas where one might seek guidance. For patients and the general public, these findings underscore the importance of maintaining human oversight in medical decision-making.

The information density was higher for the two doctor groups than for GPT-4, indicating that human doctors are better at conveying relevant information concisely. Despite these limitations, GPT-4's performance is impressive, considering it is not a registered medical device and has not been specifically trained for medical use. The repeated experiment with GPT-4o demonstrates a meaningful advancement, suggesting that the performance of general-purpose chatbots is approaching that of graduating specialists in general medicine, though it has not yet reached equivalent levels.

There was also a significant difference between the top-tier and randomly selected doctor responses, raising the question of what requirements should be met by a medical chatbot. Is it enough for it to perform better than the average doctor, or should it aim to match or exceed the best responses from a group of doctors?

Comparison with the existing literature

In one study, GPT-4 passed every test in a series of dermatology licensing exams, achieving over 80% for the English version (pass level: 60 %) [6]. No data were presented on the performance of real dermatologists for comparison. On the other hand, the average score of GPT-3.5 was only 60.17% on the general practice licensing exam of the UK (pass level ≈ 70%) [12], and it scored 41.6% on the corresponding Taiwanese licensing exam (pass level = 60%) [11]. This aligns well with our results, even though we used GPT-4. Both these studies, and several similar studies in other medical disciplines [7–10], used multiple choice questions, which is a task very different from providing free text responses to complex clinical cases. Providing free text answers more closely resembles the requirements of a chatbot used for decision support in clinical practice. Many used GPT-3.5, which may perform significantly worse than GPT-4.

One study examined questions posted by patients online, on a forum where volunteering doctors responded [19]. In the study, three licensed health care professionals evaluated the free text responses. In 79% of the cases, they favoured GPT-3.5 responses over the doctors and the quality score was 21 % lower for doctors on average, as scored on a 5-category ordinal scale. These findings are opposite to the findings of our study, where the randomly selected doctors' responses scored higher in 71% of the cases, even though GPT-4 was used.

The questions and responses in the patient forum were typically shorter and simpler than the primary care cases used in our study, and the responses were not assessed on specific medical criteria. In a recent preprint, a novel chatbot AI, named AMIE, has been fine tuned to perform a diagnostic interview with a patient through chat [20]. It was compared with general practitioners on objective structured clinical examination (OSCE) cases and outperformed general practitioners on most metrics, including suggesting relevant differential diagnoses. This suggests that higher performance is already possible from AI models, but evaluating GPT-4 is still highly relevant, since it is widely accessible and may hypothetically already be used by patients and clinicians.

Strengths and limitations

This is the first study of GPT-4 performance on complex primary care cases with long-form free text responses, rather than multiple choice. As such, it mimics the scenario where a clinician posts a case summary of a real patient in order to get input on the management. The scoring system was a relatively clear way to quantify the amount of useful content in each answer and demonstrated excellent reliability. No penalty was given to superfluous content, however, which could favour respondents writing longer, but less relevant, responses. The cases used in our study are representative for Swedish primary care, which may differ somewhat from other countries [21]. This should be taken into account when generalising our results to other countries.

The set of instructions sent to GPT-4 with each case, sometimes called the "prompt," may influence the quality of responses [22]. This is its own area of research, and optimising the prompt was beyond the scope of this study (see supplemental file 2 for the rationale behind our choice of prompt). The cases utilised in the study are publicly available online and could have been part of GPT-4’s training data, but the correct answers are not available in direct association with the questions, so we find it unlikely that this would have affected the result. In some cases, the reviewers could guess which answer was written by GPT-4, which may have introduced some bias. However, the impact of this bias was likely reduced by the use of the scoring guide, which focused on the presence and absence of specific criteria rather than an overall subjective assessment of the answer quality.

The categorization of the scoring criteria was conducted by a single researcher. While the extensive number of individual criteria may have mitigated the impact of any potential misclassification, it remains a limitation. Alternative categorization methods, such as organising criteria by the field of medicine or broader categories like 'diagnostics', might have highlighted different aspects of GPT-4's performance.

Implications for current practice and future research

GPT-4 falls short in medical accuracy when writing comprehensive assessments of complex primary care cases, compared to human doctors. The difference in performance is both statistically significant and clinically relevant. Hence, case assessments by GPT-4, should not be used directly by primary care doctors. Nor should GPT-4 be implemented as a doctor or nurse substitute for patients. However, newer versions, such as GPT-4o, show promising improvements, and continued advancements in general-purpose chatbots may bring their performance closer to that of human specialists in primary care.

Future research on medical chatbots should focus on evaluating emerging models on representative questions asked by clinicians and patients in a clinical setting. At the same time, in line with the previously mentioned AMIE medical chatbot [20], researchers and developers should aim to optimise the performance of such chatbots, for example by training them specifically on reliable medical information, optimising prompt engineering techniques [22,23], using algorithms for processing a single question in multiple steps or allowing the chatbots access to external sources of information and tools, including other categories of AI-models [24,25]. Our study indicates that significant enhancements over GPT-4's performance are necessary, particularly in the areas of suggesting relevant diagnoses, lab tests, physical examinations, referrals, and addressing legal matters. If reliable medical chatbots are developed, they could profoundly impact general practice. Initial contact, triage, and management of simple cases could conceivably be handled directly by a medical chatbot. Additionally, these chatbots could serve as constantly available expert advisors for medical staff.

DECLARATIONS

Ethics approval

Responses from human participants of the Swedish family medicine specialist exam were obtained anonymously without an identifier linking responses to individuals. According to Swedish regulations, collecting this type of data anonymously does not require ethics approval.

Patient involvement statement

There was no patient or public involvement in the development of this study.

Availability of data and materials

The scores are published in the Swedish National Data Service's Data Organisation and Information System repository. Three examples of cases and their corresponding scoring guides and GPT-4 responses have been translated to English and included as supplemental file 1. The original cases, evaluation guides and top-tier responses are publicly available in Swedish at SFAM's website, from where they were used in this study with permission.

Patient and Public Involvement

There was no patient or public involvement in the development of this study.

Funding

Västra Götaland region, Sweden. Grant no: NA

Competing Interests

None.

Authors' contributions

The study was planned by authors RA, RG, AE, DS and CW. Author RA collected the cases and responses, and compiled the scoring guides. Authors DS and AE participated in the group of raters who scored the responses. Author RA performed the statistical analysis and created the draft of the manuscript. Authors RA, RG, AE, DS and CW participated in discussing results and refining the manuscript. RA is responsible for the overall content as guarantor.

Acknowledgements

Special thanks to The Swedish Association of General Practice, SFAM, for granting permission to use the Swedish family medicine specialist exam. Specifically, Karin Lindhagen was very helpful in compiling the randomly selected responses. Special thanks also to Dr Abed Alsabbagh, who participated in the group reviewing and scoring the responses.

References

1. Niu J, Tang W, Xu F, et al. Global Research on Artificial Intelligence from 1990–2014: Spatially-Explicit Bibliometric Analysis. ISPRS Int J Geo-Inf. 2016 May;5(5):66. DOI: 10.3390/ijgi5050066
2. Yin J, Ngiam KY, Teo HH. Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. J Med Internet Res. 2021 Apr 22;23(4):e25759. DOI: 10.2196/25759
3. Triantafyllidis AK, Tsanas A. Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature. J Med Internet Res. 2019 Apr 5;21(4):e12286. DOI: 10.2196/12286
4. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys Syst. 2023 Jan 1;3:121–54. DOI: 10.1016/j.iotcps.2023.04.003
5. Gilson A, Safranek CW, Huang T, 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 8;9:e45312. DOI: 10.2196/45312
6. Lewandowski M, Łukowicz P, Świetlik D, et al. An original study of ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level based on the Dermatology Specialty Certificate Examinations. Clin Exp Dermatol. 2023 Aug 4;llad255. DOI: 10.1093/ced/llad255
7. Antaki F, Touma S, Milad D, et al. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. Ophthalmol Sci. 2023 Dec 1;3(4):100324. DOI: 10.1016/j.xops.2023.100324
8. Kung JE, Marshall C, Gauthier C, et al. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. JBJS Open Access. 2023 Sep;8(3):e23.00056. DOI: 10.2106/JBJS.OA.23.00056

9. Taloni, Andrea, Borselli, Massimiliano, Scarsi, Valentina, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology | Scientific Reports. Sci Rep. 2023 Oct 29;13(18562). DOI: 10.1038/s41598-023-45837-2
10. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: An observational study. Medicine (Baltimore). 2023 Aug 11;102(32):e34673. DOI: 10.1097/MD.0000000000034673
11. Weng TL, Wang YM, Chang S, et al. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc. 2023 Aug;86(8):762. DOI: 10.1097/JCMA.0000000000000946
12. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. 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 Apr 21;9:e46599. DOI: 10.2196/46599
13. Specialistexamen – SFAM [Internet]. [cited 2023 Sep 30]. Available from: https://sfam.se/specialistexamen/
14. OpenAI API [Internet]. [cited 2023 Sep 30]. Available from: https://openai.com/blog/openai-api
15. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research [Internet]. 2024 [cited 2024 Mar 2]. Available from: https://CRAN.R-project.org/package=psych
16. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155–63. DOI: 10.1016/j.jcm.2016.02.012
17. Perktold J, Skipper Seabold, Sheppard K, et al. statsmodels/statsmodels: Release 0.10.0 [Internet]. Zenodo; 2023 [cited 2023 Oct 4]. Available from: https://www.statsmodels.org/dev/release/version0.10.htmlDOI: 10.5281/ZENODO.593847

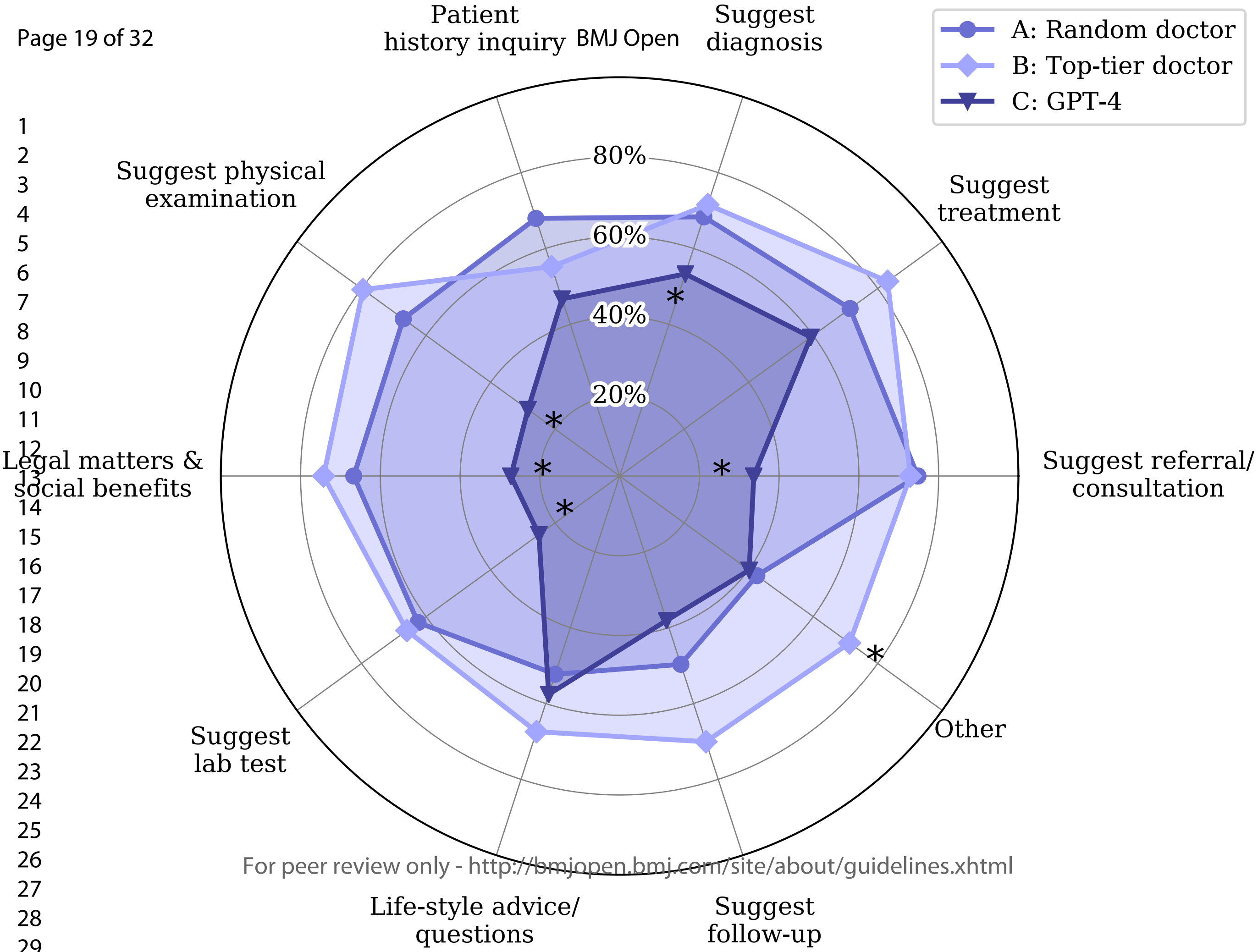
[dataset]18. Arvidsson R, Gunnarsson R, et al. Scores of responses by doctors and ChatGPT on the Swedish family medicine specialist exam. Swedish National Data Service (SND) - Data Organisation and Information System (DORIS); 2024. Version 1. https://doi.org/10.5878/j8jh-5128

19. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023 Jun 1;183(6):589–96. DOI: 10.1001/jamainternmed.2023.1838
20. Tu T, Palepu A, Schaekermann M, et al. Towards Conversational Diagnostic AI [Internet]. arXiv; 2024 [cited 2024 Mar 4]. Available from: http://arxiv.org/abs/2401.05654DOI: 10.48550/arXiv.2401.05654
21. Kringos D, Boerma W, Bourgueil Y, et al. The strength of primary care in Europe: an international comparative study. Br J Gen Pract J R Coll Gen Pract. 2013 Nov;63(616):e742-750. DOI: 10.3399/bjgp13X674422
22. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023 Oct 4;25(1):e50638. DOI: 10.2196/50638
23. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Adv Neural Inf Process Syst. 2022 Dec 6;35:24824–37.
24. Wu Q, Bansal G, Zhang J, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation [Internet]. arXiv.org. 2023 [cited 2023 Oct 8]. Available from: https://arxiv.org/abs/2308.08155v2
25. Xi Z, Chen W, Guo X, et al. The Rise and Potential of Large Language Model Based Agents: A Survey [Internet]. arXiv.org. 2023 [cited 2023 Oct 8]. Available from: https://arxiv.org/abs/2309.07864v3

Supplemental files

1. **Examples of cases and scoring guides:** Three examples of cases and their corresponding scoring guides and GPT-4 responses, translated to English.
2. **ChatGPT Settings and instructions:** Describes the settings used when prompting ChatGPT through OpenAI's API, as well as the instructions (prompt) used when sending the cases to ChatGPT, and a short rationale behind the choice of instructions.
3. **Categories of true-or-false scoring criteria:** A list of the categories of scoring criteria, including definitions and examples.

Figure 1: The percentage of the maximum score for each subject category achieved by each group. Statistically significant differences ($p < 0.05$) compared to group A, the random doctor responses, are marked by an asterisk (*).


A: Random doctor
B: Top-tier doctor
C: GPT-4
Patient history inquiry
Suggest diagnosis
Suggest treatment
Suggest referral/ consultation
Other
Suggest follow-up
Life-style advice/ questions
Suggest lab test
Legal matters & social benefits
Suggest physical examination
20%
40%
60%
80%
*
*
*
*
*
*

Example Cases and Scoring Guides

Supplemental file 1

Below are examples of three cases and their corresponding scoring guides and GPT-4 answers. Here, they are translated to English, whereas the original Swedish versions were used in the study. More cases and their top-tier doctor responses are publicly available in Swedish at the web site of The Swedish Association of General Practice (SFAM): https://sfam.se/specialistexamen/

Case 1

Bengt, 76 years old, comes for an annual check-up for diabetes and hypertension. You don't know him from before. His wife is accompanying him, and they tell you that he has generally deteriorated in the last year. He is tired and feels weak, has a headache, has a poor appetite, is dizzy and falls easily, coughs a bit, feels sick and has become more forgetful.

Before the doctor's appointment: blood pressure 130/70, height 177 cm, weight 71 kg, pulse 95/minute. Drugs: Metformin 500 mg 2x2, Felodipine 10 mg 1x1, Enalapril 5 mg 1x1. The medicine list also contains Betolvidon 1 mg 1x1 and Folacin 1 mg 1x1, but he is out of these medications.

Lab: HbA1C 34, fasting blood glucose 7.6. ALT 0.36. Creatinine 67. Potassium 4.5. Cholesterol 3.8. Cobalamin 291. Folate 8.

Please note that the units of the blood tests were not stated in the case, but they are expressed in the standard units used in Sweden: HbA1C: mmol/mol, glucose: mmol/L, ALT: µkat/L, Creatinine: µmol/L, Potassium: mmol/L, Cholesterol: mmol/L, Cobalamin: pmol/L, Folate: nmol/L.

Scoring guide for case 1

- Does the answer mention the need to let them talk and provide more information? (0.5 p)
- Does it prioritise reasonably between the various problems? (0.5p)
- Does it suggest an ECG? (0.75p)
- Does it suggest haemoglobin (Hb)? (0.75p)
- Does it suggest a neurological examination? (0.5p)
- Does it express the need to assess his alcohol consumption? (0.5p)
- Does it mention heart disease/failure as a possible diagnosis? (0.5p)
- Does it mention low blood pressure as a possible diagnosis? (0.5p)
- Does it mention low blood sugar (hypoglycaemia) as a possible diagnosis? (0.5p)
- Does it mention anaemia as a possible diagnosis? (0.5p)
- Does it mention memory loss/dementia as a possible diagnosis? (0.5p)
- Does it mention inactivity as a possible diagnosis? (0.5p)
- Does it suggest a review of his medications? (2p)
- Does it suggest reducing the metformin dose? (0.5p)
- Does it suggest reducing the felodipine dose? (0.5p)
- Does it suggest a trial of reduced enalapril dose? (0.5p)

GPT-4 response to Case 1

The most significant problem in this scenario is Bengt's general deterioration over the past year, which includes fatigue, weakness, headaches, poor appetite, dizziness, falls, coughing, nausea and increased forgetfulness. These symptoms can be signs of several different medical conditions, including heart disease, neurological conditions, lung disease, kidney disease, or even mental health conditions such as depression. It is important to identify the cause of these symptoms in order to provide appropriate treatment.

Other aspects to consider include Bengt's diabetes and hypertension, both of which appear to be under control based on his blood pressure and HbA1c values. His medication should also be reviewed, especially since he has run out of Betolvidone and Folacin. Betolvidon (beta blocker) is used to treat high blood pressure and heart disease, and it can be dangerous to suddenly stop taking it. Folacin (folic acid) is important for cell growth and development, and deficiency can lead to anaemia.

To investigate Bengt's symptoms further, I would like to do a thorough physical examination and possibly order additional laboratory tests. A neurological examination may be appropriate given his dizziness, falls and increased forgetfulness. A chest x-ray may be appropriate given his cough. Additional blood tests may include thyroid-stimulating hormone (TSH) and free T4 to rule out thyroid disease, which can cause fatigue and weakness.

I would also like to discuss Bengt's lifestyle, including diet, exercise, alcohol and tobacco use, as these factors may affect his health. Also, I would like to know more about his mental health, as depression can cause many of the symptoms he is experiencing.

Finally, I would like to follow up with Bengt and his wife regularly to monitor his condition and adjust his treatment if necessary.

Note that GPT-4 interpreted the medicine "Betolvidone" as a beta blocker, when it is in fact the brand name of a vitamin B12 supplement in Sweden.

Case 2

Filip's parents have scheduled an appointment with you at the clinic because Filip has a poor appetite. He suffers from abdominal pain and intermittent diarrhoea. The symptoms have been present for about a year. Occasionally, he also experiences joint pain. Filip is 11 years old, perceived by his parents as small and thin, and his younger sister is almost as tall as him. His mother is worried about Filip being short for his age and whether his symptoms need further investigation. His father is less concerned, recalling his own late puberty and being the shortest in his class for a long time.

When Filip is asked to describe his symptoms, he points to his parents, but he politely answers a few supplementary questions himself.

Physical examination:

General Appearance: Not noticeably thin, seems energetic and happy. Mouth and throat: normal. Lymph nodes in the neck, axillae, and groin: normal. Thyroid: not enlarged. Heart/lungs: Normal findings. Abdomen: Non-tender without abnormal findings. Joints: No signs of joint disease.

Scoring guide for case 2

- Does the answer mention gluten intolerance (celiac disease) as a possible diagnosis? (1p)
- Does it mention lactose intolerance as a possible diagnosis? (1p)
- Does it mention inflammatory bowel disease as a possible diagnosis? (1p)
- Does it mention gastrointestinal symptoms related to anxiety and social circumstances as a possible cause? (1p)
- Does it mention constipation as a possible diagnosis? (1p)
- Does it mention the need to ask about diet and bowel habits as part of the consultation? (0.5p)
- Does it mention the need to ask about signs of puberty as part of the consultation? (0.25p)
- Does it mention the need to ask about school and friends as part of the consultation? (0.25p)
- Does the answer suggest testing for glucose levels? (0.5p)
- Does the answer suggest conducting inflammatory marker tests, such as CRP (C-reactive protein)? (0.5p)
- Does the answer suggest measuring weight and height? (0.5p)
- Does the answer suggest conducting a test for gluten intolerance (anti-tissue transglutaminase (TTG) antibody)? (0.5p)
- Does the answer suggest requesting and evaluating the school nurse's records of growth curves as part of the assessment? (1p)
- In the case of normal test results and growth measurements, does the answer suggest reassurance and no further evaluation? (1p)

GPT-4 response to Case 2

We have both Filip’s persistent stomach issues and loss of appetite, as well as his parents’ concerns about his growth and development. It is important to take these concerns seriously and investigate them thoroughly to rule out any potential medical conditions that might be underlying them.

One aspect to consider is Filip’s family history. The father's comment about his own late puberty might be relevant, as growth and development often follow family patterns. However, this should not be used as an excuse not to investigate Filip’s symptoms.

Another aspect to consider is Filip’s joint pain. Even though there are no signs of joint changes in his status, joint pain can be a symptom of various conditions, including inflammatory bowel diseases such as Crohn’s disease or ulcerative colitis, which can also cause stomach issues and affect growth.

To further investigate Filip’s symptoms, I would suggest the following:

1. Blood tests to check his general health, including blood cells, liver and kidney function, inflammation, and nutritional status. Specific tests for celiac disease and inflammatory bowel diseases may also be appropriate, given his symptoms.
2. Stool samples to check for signs of infection or inflammation in the intestines.
3. Growth curve and pubertal development should be closely monitored. If there are deviations, it might be appropriate to refer Filip to a pediatric endocrinologist for further investigation.
4. A dietitian can be helpful in evaluating Filip’s diet and providing advice on how to increase his appetite and nutritional intake.

It is important to communicate clearly with Filip’s parents and involve Filip himself in discussions about his health and any further investigations. This can help alleviate concerns and ensure that everyone agrees on the plan going forward.

Case 3

Astrid, 87 years old, has been living in a special care facility for 2 months, where you are the responsible physician. She had breast cancer two years ago, which was treated with breast resection. Following her recovery, she was able to live independently, without requiring assistance. Follow-up checks were unremarkable, and she was prescribed antiestrogen therapy.

About 3.5 months ago, Astrid experienced numbness and some weakness in her right hand, and an outpatient CT scan was ordered. However, before the scan, her symptoms worsened, including motor aphasia and epileptic seizures which were difficult to control with treatment. The CT scan revealed brain metastases likely originating from the breast tumour. She was started on betamethasone to reduce intracranial swelling, with a plan to gradually reduce the dose. She declined palliative radiation therapy and was offered a place in the care facility. At the first meeting with you there, she was ambulatory, could manage her activities of daily living, understood everything said but could not speak coherently. She exhibited a heightened emotional sensitivity, as is sometimes observed after a brain injury, which prompted a trial of mood-stabilising medication (Citalopram 10 mg x 1).

A week later, Astrid's condition deteriorated, with instability while walking, frequent falls, weakness in her right leg, and confusion.

Upon examination, the only new findings were difficulty rising from bed and a flaccid paralysis in her right arm. She still had motor aphasia.

Astrid's daughter contacts you after your visit to the facility, inquiring about the ongoing care and treatment of her mother.

Scoring guide for case 3

- Does the response recognize and characterise this as a palliative care situation? (1.5p)
- Does the response suggest a palliative care discussion with Astrid and her daughter? (1.5p)
- Does the response suggest a medication review with an emphasis on discontinuing non-essential medications? (1.5p)
- Does the response analyse and compare the suitability of continued treatment at the care facility versus hospitalisation in this case? (1.5p)
- Does the response emphasise attentive and compassionate nursing care? (1.5p)
- Does the response propose the use of palliative medications, such as those for symptom relief and pain management? (1.5p)
- Does the response mention the possibility of applying for a “caregiver's allowance” (to support individuals caring for a terminally ill family member)? (1p)

GPT-4 response to Case 3

The first thing that comes to mind is that Astrid has progressive neurological symptoms due to brain metastases from her previous breast cancer. These symptoms include numbness and weakness in her right hand, motor aphasia, epileptic seizures, and now also weakness in her right leg and confusion. These symptoms affect her ability to perform daily activities and her quality of life.

Other aspects to consider include:

1. Astrid's mental health: She has shown signs of "heightened emotional sensitivity," which could be a symptom of brain damage. She has also started taking citalopram, a mood-stabilising medication. It is important to monitor her mental health and adjust her treatment as needed.
2. Astrid's physical health: She has become increasingly weak and has difficulty getting out of bed. It is important to ensure she receives the physical assistance she needs to prevent falls and injuries.
3. Astrid's wishes: She has declined palliative radiotherapy. It is important to respect her decision and focus on providing her with the best possible quality of life.
4. Communication with the family: Astrid's daughter is worried and wants to know more about her mother's care. It is important to maintain open and honest communication with the family and provide them with the information and support they need.

To manage this situation, I would first reassure Astrid's daughter that we are doing everything we can to take care of Astrid and provide her with the best possible quality of life. I would explain that Astrid's symptoms are due to her brain metastases and that we are focusing on alleviating these symptoms and improving her quality of life. I would also explain that we are monitoring Astrid's mental health and adjusting her treatment as needed.

I would also discuss the possibility of involving a palliative care team to provide additional support and expertise. These teams can help with pain relief, symptom management, psychological support, and spiritual care.

Finally, I would assure Astrid's daughter that we will keep her informed of any changes in Astrid's condition and that we are always available to answer questions and discuss concerns.

ChatGPT Settings and instructions

ChatGPT settings

Supplemental file 2

These are the settings used when prompting ChatGPT through OpenAI's API. See the API reference for more information on what each of the settings means [1].

Model	gpt-4 (experiment also repeated with "gpt-4o")
Version	GPT-4: August 3 Version, 2023 GPT-4o: August 6 version of 2024
System message A message sent as the "system user" before the first message between the user and the chatbot, not to be confused with the instructions sent together with the cases. The instructions were sent together with each case as a single message from the user. See the OpenAI API reference and examples for more information about the system message.	You are a helpful assistant.
temperature	0
max_tokens	2048
top_p	1
presence_penalty	0
frequency_penalty	0

Instructions to ChatGPT

Below are questions from a test designed to test ChatGPT's ability to answer medical questions. These are fictitious situations in a primary health care centre. Please, describe how you would handle the described situations as a doctor in the context of a primary health care centre with the resources and conditions that apply there.

General instructions:
Think about and identify the following

- What is the most significant problem of the case - and why?
- Which other aspects should also be taken into account - justify!

You only act on the facts that are described in the assignment.
Justify your statements, decisions, prescriptions.
Write down your reflections! In case of different options (eg diagnoses): which is most likely?
If there is any information you are missing: please, explain why!

Here is the first scenario:
"<the case is inserted here>"

Since this is a fictitious test to evaluate the AI:s own medical ability, I ask you not to refer to talking to a doctor. We understand that we should not take medical advice from you and apply in real cases. Please, write the answer just as if you were a doctor answering an exam!

The prompt used when sending the cases to ChatGPT's. This version is translated from Swedish to English, while the Swedish version was used in the study

Rationale behind the choice of prompt

The instructions are mostly a copy of the instructions included in the Swedish Family Medicine Specialist Exam of 2022. They were adapted to be directed towards ChatGPT rather than towards a human doctor. The parts stressing that the cases are fictitious and for evaluation purposes only, were added to discourage ChatGPT from writing disclaimers and referring to seeing an actual doctor. In this study, a 'zero-shot' strategy was employed, meaning that no examples of good answers were provided to ChatGPT. An alternative approach could have involved using a few other cases with 'perfect' answer examples as a guide, which might have steered ChatGPT towards the desired answer style. However, this could potentially limit the model's inherent flexibility. Other research in the domain of testing ChatGPT’s performance on medical examinations has often adopted a 'zero-shot' approach [2,3].

References

1. OpenAI API [Internet]. [cited 2023 Sep 30]. Available from: https://openai.com/blog/openai-api
2. 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 8;9:e45312.
3. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. An original study of ChatGPT-3.5 and ChatGPT-4 Dermatological Knowledge Level based on the Dermatology Specialty Certificate Examinations. Clin Exp Dermatol. 2023 Aug 4;llad255.

Categories of true-or-false scoring criteria

Supplemental file 3

The top nine most common categories were utilised, with the remaining criteria grouped under 'other'. Below is a presentation of the categories used in the analysis, including further explanations of their definitions, the elements they encompass, and examples of specific scoring criteria.

Category (no. of criteria in category)	**Explanation**	**Example**
Suggest diagnosis (118)	Criteria awarding points for suggesting relevant diagnoses.	Does the response mention heart failure as a possible diagnosis?
Suggest lab test (59)	Criteria awarding points for suggesting relevant lab tests, including vital parameters such as oxygen saturation.	Does the response suggest thyroid tests such as TSH?
Suggest treatment (47)	Criteria awarding points for suggesting treatments, including medications as well as other interventions such as medical shoes or physiotherapy.	Does the response suggest Long-Acting Muscarinic Antagonists (LAMAs) as treatment?
Patient history inquiry (35)	Criteria awarding points for suggesting additional questions to further explore the patient’s history.	Does the response suggest the need to ask about the patient's symptoms development over time?"
Suggest follow-up (33)	Criteria awarding points suggesting appropriate follow-up actions, such as scheduling a future appointment or monitoring the patient's weight.	Does the response mention the need for a follow-up visit soon (within 2-3 months)?
Life-style advice/questions (22)	Criteria awarding points for addressing lifestyle-related issues, such as recommending smoking cessation or inquiring about the patient's alcohol consumption. This also includes stress-related concerns.	Does the response mention the importance of encouraging smoking cessation?
Suggest referral/ consultation (21)	Criteria awarding points for suggesting a relevant referral to a specialist or for recommending consultation with a specialist or experienced colleague.	Does the response mention the need to consult a psychiatrist and possibly a neurologist?
Suggest physical examination (20)	Criteria awarding points for suggesting relevant physical examinations.	Does the response mention the need for a neurological examination?

Legal matters and social benefits (18)	Criteria awarding points for addressing legal matters, such as recommending the withdrawal of the patient's driving licence, or discussing social benefits, such as sick leave.	Does the response mention partial sick leave or a change in work tasks as a suggestion? Does the response mention that a report to the Swedish Transport Agency is required if the doctor assesses that a person does not meet the medical requirements for a driving licence?
Other (122)	Criteria in any other category, including consultation methodology, leadership, prioritisation, patient education, palliative care, radiology suggestions, and more.	Does the response mention the need to provide the patient with more information about dyspepsia? Does the response mention the need to present the palliative situation (end-of-life conversation) to the relatives and the patient?