Received: 26 March 2024 | Revised: 16 May 2024 | Accepted: 19 May 2024

DOI: 10.1002/lrh2.10438

BRIEF REPORT

Learning Health Systems

Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases

Geoffrey W. Rutledge

HealthTap, Sunnyvale, California, USA

Correspondence
Geoffrey W. Rutledge, HealthTap, 209 E. Java Dr. #61987, Sunnyvale, CA 94088, USA.
Email: geoffrey.rutledge@healthtap.com

Abstract

Introduction: Large language models (LLMs) have a high diagnostic accuracy when they evaluate previously published clinical cases.

Methods: We compared the accuracy of GPT-4's differential diagnoses for previously unpublished challenging case scenarios with the diagnostic accuracy for previously published cases.

Results: For a set of previously unpublished challenging clinical cases, GPT-4 achieved 61.1% correct in its top 6 diagnoses versus the previously reported 49.1% for physicians. For a set of 45 clinical vignettes of more common clinical scenarios, GPT-4 included the correct diagnosis in its top 3 diagnoses 100% of the time versus the previously reported 84.3% for physicians.

Conclusions: GPT-4 performs at a level at least as good as, if not better than, that of experienced physicians on highly challenging cases in internal medicine. The extraordinary performance of GPT-4 on diagnosing common clinical scenarios could be explained in part by the fact that these cases were previously published and may have been included in the training dataset for this LLM.

KEYWORDS
challenging clinical cases, differential diagnosis, GPT-4, large language models

1 | INTRODUCTION

The release of generative AI based on very large language models (LLMs) has created renewed interest in using AI chatbots for health care.[1] Prior studies reported that the GPT-4 LLM achieved a top 10 diagnostic accuracy of 57% on challenging clinical cases that were derived from published NEJM Clinical Pathological Case (CPC) reports versus 36% achieved by NEJM readers.[2] Another report showed that the Palm-2 LLM, tuned for performance on the differential diagnosis task, achieved top 10 diagnostic accuracy of 59.1% on NEJM CPC cases versus physicians who achieved an accuracy of 33.6%.[3]

2 | QUESTIONS OF INTEREST

Previous evaluations were based on cases whose details were published in NEJM and available for use in training data for medically tuned LLMs. To remove the possibility that the dataset used for training the LLMs included the test data and validate the diagnostic accuracy of GPT-4, we undertook to evaluate its diagnostic performance on a set of highly challenging clinical cases whose details and correct diagnoses were not previously published.

We also tested GPT-4 against a set of more common clinical scenarios for which the diagnostic performance of doctors was previously measured and reported.

3 | METHODS

We evaluated the capability of GPT-4 to generate useful differential diagnoses for two sources of clinical scenarios.

1. A set of 45 clinical vignettes that represent common clinical conditions and contain enough information for an experienced clinician to make an accurate diagnosis. This set reported in Semigran et al.[4] includes 15 cases that warrant urgent evaluation, 15 cases that warrant outpatient physician evaluation, and 15 cases that are less acute and suitable for self-care. These cases were previously published, and the details of each case and the correct diagnoses are available online.[5]
2. A set of 36 challenging cases in internal medicine. These scenarios were previously created by Friedman et al. as diagnostic challenges to test internal medicine physicians and decision support systems. They were carefully selected by three experienced internists from unpublished actual clinical cases that have a known correct diagnosis, but the features are not sufficiently definitive for experts to identify with certainty the correct diagnosis. That is, the cases do not contain pathognomonic or definitive diagnostic information. They were designed to challenge doctors to identify the possible causes for each presentation, and the evaluation metric was whether or not the differential diagnosis list constructed included the correct diagnosis. The authors of these cases were careful not to allow publication of the details of each case or the correct diagnosis for each case. Further details of the case creation methodology, assessment of the level of difficulty, and the selection method for these curated cases are in the reference.[6] Two example cases from this set were published as a supplement to a paper published in BMJ Quality and Safety.[7]

We accessed chatGPT-4 in January 2024 via the commercially available chat.openai.com service. The prompt used was "For all the following clinical cases, list the most likely diagnosis and all other likely diagnoses in JSON format:" followed by the full text of each case. No images, graphics, or other media were included in the inputs to GPT-4.

This commercially available GPT-4 service has no parameters that the user can set, and OpenAI does not explicitly disclose the parameter settings for the chat.openai.com service. In response to the query "What are the values of all parameters when someone accesses GPT4 using chat.openai.com?" The GPT-4 interface describes the relevant internal parameter settings as follows:

- Temperature: For chat applications, a lower temperature (e.g., around 0.5–0.7) is common to ensure responses are coherent and less random.
- Top_p: This might be set in a range to support diversity but maintain relevance, possibly around 0.9.

4 | RESULTS

For the 45 previously published clinical vignettes, the top 1 diagnosis from GPT-4 was 96% (43/45) correct, and the top 3 diagnoses were 100% (45/45) correct. In previous studies, physicians scored the top 1 diagnosis 72.1% (797/1105) correct, and the top 3 diagnoses 84.3% (932/1105) correct[4] (see Table 1).

For the 36 challenging internal medicine cases, the top 6 diagnoses from GPT-4 were 61.1% (22/36) correct. In previous studies, physicians during their residency training achieved the top 6 correct in 43.7% (283/648), and physicians on faculty achieved the top 6 diagnoses correct in 49.1% (314/639) of their evaluations[6] (see Table 2).

5 | DISCUSSION

Experienced clinicians included the correct diagnosis in their top 3 diagnoses for the common clinical vignettes 84% of the time, and they included the correct diagnosis in the top 6 list for the most challenging internal medicine cases 44% (residents) and 49% (faculty) of the time.[4]

By contrast, compared to the physicians, GPT-4 achieved a significantly better top 3 diagnostic accuracy for the common clinical scenarios (100%) and a better top 6 diagnostic accuracy for the most challenging cases (61.1%).

The diagnostic performance of GPT-4 was measured as better than that of internal medicine residents in training and as good as or better than the performance of experienced faculty physicians. This result for the LLM that has not been trained specifically on the medical domain is remarkable and suggests that with additional training, it should be possible for LLMs to achieve performance that is consistently better than that of physicians.

It is perhaps not surprising that GPT-4 would perform so well for more common clinical scenarios. These common clinical

TABLE 2 GPT-4 diagnostic performance on previously unpublished challenging internal medicine cases.[5]

	Top 6 diagnostic accuracy %		95% CI	*p*-value
GPT-4	61.1%	(22/36)	[0.43, 0.77]	
Residents	43.7%	(283/648)	[0.40, 0.48]	0.04
Faculty	49.1%	(314/639)	[0.45, 0.53]	0.13

TABLE 1 GPT-4 diagnostic performance on previously published clinical vignettes.[4,5]

Percent of correct diagnoses	Diagnosis in Top 1	CI Top 1	*p*-value	Diagnosis in Top 3	CI Top 3	*p*-value
GPT-4	95.6% (43/45)	[0.82, 0.99]		100% (45/45)	[0.92, 1]	
Physicians	72.1% (797/1105)	[0.69, 0.75]	0.002	84.3% (932/1105)	[0.82, 0.86]	<0.001

scenarios are freely available online, and a Google search for the text of each case returns the URL of the case descriptions and their correct diagnoses from the web service of a well-known and highly regarded medical publication. Thus, these cases were almost certainly included in the training set for GPT-4. This highlights that the evaluation of LLMs should be carefully constructed to avoid the possibility of contamination of the test data in the training dataset for the LLM.

6 | CONCLUSION

GPT-4 demonstrates a remarkable ability to generate accurate differential diagnosis lists for both common and highly challenging cases in internal medicine, with a performance that is at least as good as, if not better than that of experienced physicians.

The testing of LLM-based generative AI models against previously published clinical cases should be interpreted with caution because previously published cases are likely to be included in the training data for the LLMs.

ACKNOWLEDGMENTS

The author would like to thank Dr. Charles P. Friedman for sharing the details of the previously compiled and unpublished challenging cases in internal medicine.

CONFLICT OF INTEREST STATEMENT

The author is an employee of HealthTap and may own stock as part of the standard compensation.

ORCID

Geoffrey W. Rutledge https://orcid.org/0000-0002-1238-9389

REFERENCES

1. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388:1233-1239. doi:10.1056/NEJMsr2214184
2. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical case. *NEJM AI*. 2023;1(1):1-3. doi:10.1056/AIp2300031
3. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. arXiv:2312.00164v1. https://arxiv.org/pdf/2312.00164.pdf
4. Semigran HL, Levine DM, Nundy S, et al. Comparison of physician and computer diagnostic accuracy. *JAMA Intern Med*. 2016;176(12):1860-1861. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2565684
5. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self-diagnosis and triage: audit study. *BMJ*. 2015;351:h3480. https://www.bmj.com/content/351/bmj.h3480.long
6. Friedman CP, Elstein AS, Wolf FM, et al. Enhancement of Clinicians' diagnostic reasoning by computer-based consultation a multisite study of 2 systems. *JAMA*. 1999;282(19):1851-1856. https://jamanetwork.com/journals/jama/fullarticle/192106
7. Sibbald M, Monteiro S, Sherbino J, et al. Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of Isabel. *BMJ Qual Saf*. 2022;31:426-433. https://qualitysafety.bmj.com/content/31/6/426

How to cite this article: Rutledge GW. Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases. *Learn Health Sys*. 2024;8(3):e10438. doi:10.1002/lrh2.10438