frontiers | Frontiers in Artificial Intelligence

TYPE Brief Research Report
PUBLISHED 05 August 2024
DOI 10.3389/frai.2024.1379297

OPEN ACCESS

EDITED BY
Olawande Daramola,
University of Pretoria, South Africa

REVIEWED BY
Sunyoung Jang,
The Pennsylvania State University, United States
Hanqing Zhao,
Hebei University, China

*CORRESPONDENCE
Eoin Daniel O'Sullivan
✉ eoin.osullivan@health.qld.gov.au

RECEIVED 30 January 2024
ACCEPTED 23 July 2024
PUBLISHED 05 August 2024

CITATION
Khan MP and O'Sullivan ED (2024) A comparison of the diagnostic ability of large language models in challenging clinical cases. *Front. Artif. Intell.* 7:1379297. doi: 10.3389/frai.2024.1379297

A comparison of the diagnostic ability of large language models in challenging clinical cases

Maria Palwasha Khan[1] and Eoin Daniel O'Sullivan[1,2]*

[1]Kidney Health Service, Metro North Hospital and Health Service, Brisbane, QLD, Australia, [2]Institute of Molecular Bioscience, University of Queensland, St Lucia, QLD, Australia

Introduction: The rise of accessible, consumer facing large language models (LLM) provides an opportunity for immediate diagnostic support for clinicians.

Objectives: To compare the different performance characteristics of common LLMS utility in solving complex clinical cases and assess the utility of a novel tool to grade LLM output.

Methods: Using a newly developed rubric to assess the models' diagnostic utility, we measured to models' ability to answer cases according to accuracy, readability, clinical interpretability, and an assessment of safety. Here we present a comparative analysis of three LLM models—Bing, Chat GPT, and Gemini—across a diverse set of clinical cases as presented in the New England Journal of Medicines case series.

Results: Our results suggest that models performed differently when presented with identical clinical information, with Gemini performing best. Our grading tool had low interobserver variability and proved a reliable tool to grade LLM clinical output.

Conclusion: This research underscores the variation in model performance in clinical scenarios and highlights the importance of considering diagnostic model performance in diverse clinical scenarios prior to deployment. Furthermore, we provide a new tool to assess LLM output.

KEYWORDS

artificial intelligence, machine learning, clinical medicine, LLM, diagnostics

1 Introduction

Accurate diagnosis is a fundamental step in high quality clinical care. The potential for large language models (LLMs) to provide clinical support and to improve diagnostic abilities of clinicians is increasingly appreciated and the subject of much research interest (Cascella et al., 2023; Chen et al., 2023; Chirino et al., 2023; Huang et al., 2023; Kleesiek et al., 2023). While it is becoming apparent that publicly available LLMs can produce impressive results in clinical vignettes, it is not known which model is most currently most useful to a working clinician, nor is there a reproducible way to compare LLMS output (Kung et al., 2023).

Here, we propose a simple, rapidly deployed and actionable grading rubric to compare the clinical utility of the output of LLMs. We use this rubric to grade the output of three publicly available models, ChatGPT (GPT3.5, April 2024 OpenAI, San Francisco, United States), Bing (GPT4, April 2024 Microsoft, Redmond, United States) and Gemini (Pathways Language Model -PaLM v1.5, April 2024, Google, Mountain View, USA) when asked provide clinically

appropriate diagnosis and differential to a clinical vignette. To simulate challenging cases where a competent clinician may realistically have need of diagnostic support, we selected a range of clinical cases presented in the New England Journal of Medicines (NEJM) Cases series. These cases have sufficient complexity to move beyond the shorter, "classic" vignettes of simple medical examination and provide challenge to a post graduate clinician. To measure the ability of the models to provide clinically useful support, we designed a simple rubric to grade output. This focused on both the ability to provide an accurate and understandable diagnosis, appropriate differentials as well as providing clinically safe output, free of hallucinations or dangerous suggestions and presented in a readable, understandable manner.

2 Methods

Ten distinct clinical cases were selected from clinical cases in the NEJM case series, covering a spectrum of medical conditions, from ovarian masses to toxic shock. The case discussions in the NEJM are presented initially as a vignette of the history of the presenting complaint, relevant initial investigations, and examination findings. Thereafter, an expert discussion follows, explaining the clinical reasoning behind relevant differentials. This initial vignette was the input provided to models. Specifically, the three models under investigation, Bing, ChatGPT4 and Gemini were asked to "please provide a diagnosis and relevant differential diagnosis to the following case," and the unedited text (and data tables where relevant) was provided. The free, consumer facing versions of the models were used, as open to the general public and accessed via the respective websites. No other programs or plugins were used. In this manner, models were prompted using identical text input. The specific cases chosen were: Ruptured ovarian cyst, Systemic Lupus Erythematosus, Insulinoma, Posterior Reversable Encephalopathy Syndrome, Renal Cryptococcosis, Necrotizing Anterior Scleritis, B12 deficiency, HNF1b mutation, SARS-ARDS and Septic shock syndrome.

A Rubric was created to grade the output of the models based on real world clinical interpretability and utility. The correct diagnosis and relevant differentials are provided by the NEJM in the case, so this was used as the ground truth for evaluation of accuracy. However, other metrics such as readability and potentially dangerous suggestion as subject by nature, so multiple medical clinicians from a range of specialties, countries and levels of seniority provided scores to allow generalisability. The grading rubric developed is presented in Figure 1.

The performance metrics the rubric focused on were correct diagnosis, differential diagnosis, hallucination, readability, explainability, incorrect statements, and overall subjective assessment of potentially risk suggestions. The grading rubric was applied independently by 10 physicians from 3 different countries and a range of seniorities and specialties, and an average score calculated.

Scores were compared between models using ANOVA test and Tukey's *post hoc* testing. *p*-values below 0.05 were considered statistically significant. Analysis performed in R version 4.3.2. Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC) to measure the consistency of scores assigned by multiple raters. A two-way consistency model (ICC type: consistency) was applied to evaluate agreement among raters. The ICC was computed based on the average scores provided by each rater.

Criteria	Score	Descriptor
Diagnosis	5	Correct diagnosis, accurately described
	3	A reasonable diagnosis that would appear in the top 3 differential diagnoses of a competent attending physician.
	2	A differential diagnosis that is technically correct but would not be in the top 3 differentials of a competent attending physician.
Differential	3	Accurate, concise list of at least 2 relevant differentials.
	2	Incomplete (missing at least 1 differential that would be in the top three of a competent physician) or excessive (over 5 differentials)
	1	Inaccurate or inappropriate suggestion, or missing at least 2 appropriate suggestions
Hallucination	2	No hallucinations in text
Readability	3	Readable to a layperson
	2	Readable to a doctor, concise and focused text
	1	Any of the following: Incoherent or excessive, distracting, or irrelevant information.
Logical Flow	2	Clear presentation of logical flow, understandable by domain novice
	1	logical flow to answer that is mostly complete, or gaps in logical can be filled in by a competent attending physician
	0	Any one of the following: Minimal or absent logical flow, incorrect logic displayed, inappropriate assumptions or relationships suggested.
Factually correct	1	All information presented as a fact is correct
Dangerous	-5	Any clinically dangerous suggestion – a differential or suggestion that could foreseeably lead to patient harm.

FIGURE 1
The grading rubric to assess LLM output when presented with the case vignette.

Statistical analysis was performed using the icc() function from the irr package in R. The ICC value, along with its 95% confidence interval (CI), was calculated to quantify the level of agreement among raters. A significance test (F-test) was conducted to determine whether the ICC significantly deviated from zero, indicating reliable agreement among raters.

3 Results

Total scores for each question are shown in Figure 2A. The B12 Deficiency case, and SLE case were found to have significantly lower means scores compared against the total mean score of 11.4, suggesting the LLM output was of lower quality in these cases. In contrast, the Insulinoma case and SARS-ARDS case both had statistically higher scores, suggesting the LLM output was of higher subjective quality, and the models were more adept at these cases.

The cases used, and LLM output are provided in Supplementary material S1, and the observers scoring is provided in Supplementary material S2. The mean total scores per model were Bing 10.4, ChatGPT 11.5 and Gemini 12.3. The distribution of scores by model is shown in Figure 2B. The analysis of total scores among different models show significant differences. Specifically, the mean total score was 1.09 points higher for ChatGPT output compared to Bing (95% CI [0.249, 1.931], $p = 0.007$), and 1.87 points higher for Gemini compared to Bing (95% CI [1.029, 2.711], $p < 0.001$). However, the difference between Gemini and ChatGPT was not statistically significant (mean difference = 0.78, 95% CI [−0.061,

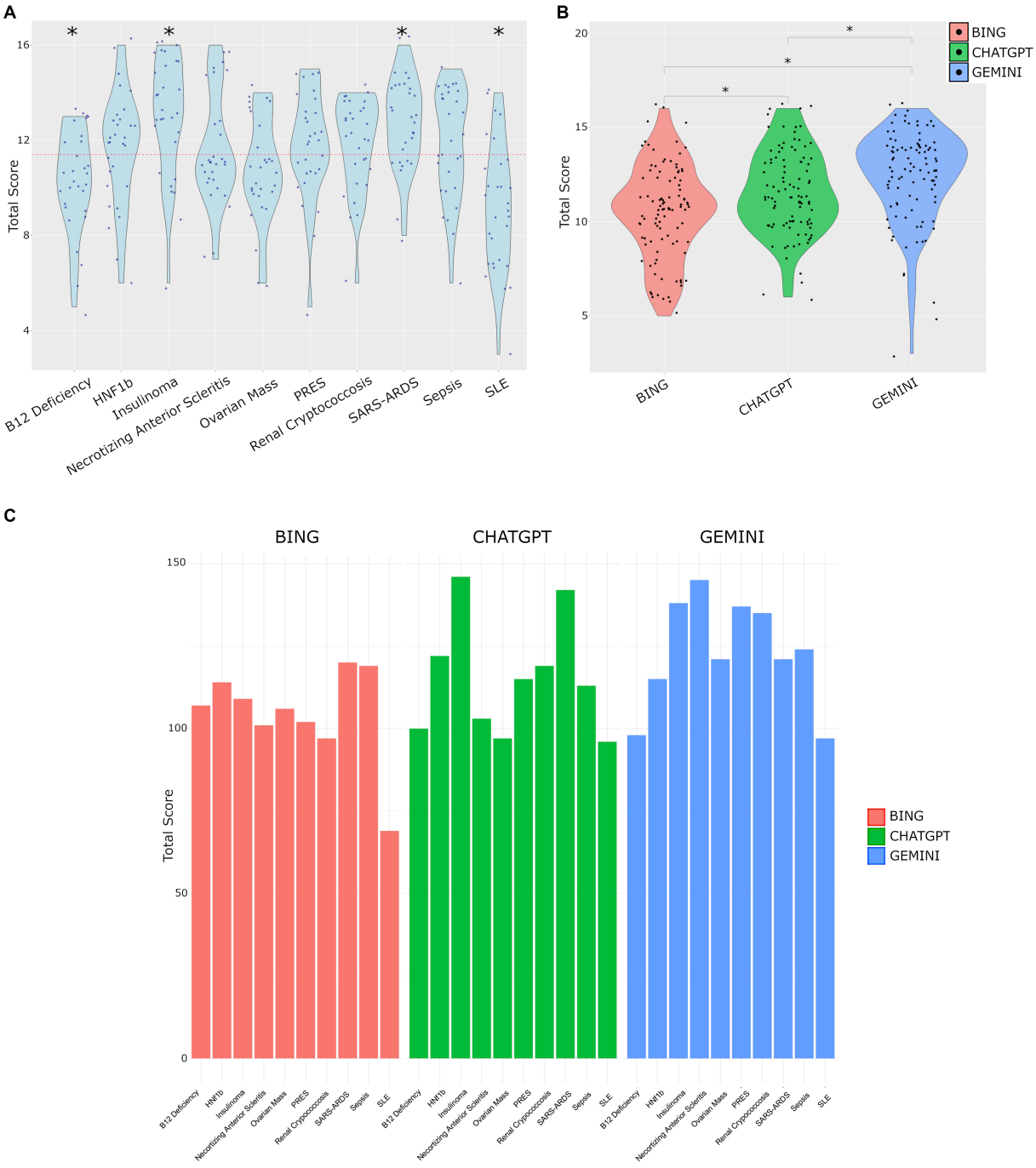

FIGURE 2
(A) Total scores across all cases. Red dotted line represents the mean score (11.4). Significant differences tested by *T*-test vs. mean score, $p < 0.05$. **(B)** Distribution of total scores by LLM. Tukey's *post hoc* test following ANOVA revealed significant differences in total scores among models with mean differences of 1.09 (95% CI [0.249, 1.931], $p = 0.007$) for CHATGPT vs. BING, 1.87 (95% CI [1.029, 2.711], $p < 0.001$) for GEMINI vs. BING, and 0.78 (95% CI [−0.061, 1.621], $p = 0.075$) for GEMINI vs. CHATGPT. Adjusted *p*-values were used, with significance set at $p < 0.05$. **(C)** Distribution of the sum of total scores for each question on a per model basis.

1.621], $p = 0.075$). These findings highlight distinct performance characteristics among the models, with Gemini showing the highest mean total score.

The performance of each model on each question is shown in Figure 2C. The total scores representing the sum of each observer's score was highly variable across questions and models.

The analysis of average scores using the two-way consistency model demonstrated strong agreement among raters. The intraclass correlation coefficient (ICC) was calculated to be 0.892 (95% CI, 0.823–0.941), indicating a high level of consistency in scoring across the 30 subjects rated by 10 raters [$F(29,261) = 9.23$, $p = 3.76\text{e-}26$]. This finding supports the reliability of the scoring rubric.

4 Discussion

This report used readily available, consumer facing models and complex cases with sufficient red herrings and distractors so as to challenge a physicians. As a whole, the models performed best on relatively straightforward cases (SARS-ARDS and Insulinoma where there were few differentials), and performed least well in the most complex cases of B12 deficiency and SLE, which were the most undifferentiated presentations. This suggests a strength of LLM models may be in "rare" diseases such as insulinomas which, while rare, have distinct features, as opposed to vaguer, multisystem diseases such as SLE. The Bing model, while competent in easier cases, exhibited limitations in correctly diagnosing challenging cases. Furthermore, Bing is limited to a finite number of questions per day, impacting on its reliability and utility in a real-world environment.

We found that ChatGPT and Gemini both outperformed Bing, a finding that has been consistent when tested in a range of clinical scenarios including haematology cases, physiology cases vignettes, surgical decision making and dentistry (Dhanvijay et al., 2023; Giannakopoulos et al., 2023; Kumari et al., 2023; Lim et al., 2023; Gomez-Cabello et al., 2024; Lee et al., 2024a,b).

Our grading tool was easily understandable and quick to deploy, and demonstrated high interobserver reliability suggesting it may be of use to other researchers when assessing the diagnostic output of LLM models.

Limitations include the number of cases used which may impact generalisability, as well as assessing answers in a non-clinical environment, rather than assessing the outputs utility in clinical work setting in real time. Additionally, we restricted our assessment to English language output limiting generalisability to other languages.

The use of LLMS in clinical diagnostics presents significant ethical and data security challenges. Patient confidentiality is at risk when using such models. While it is possible to anonymise input to protect patient identity, it is entirely conceivable that identifiable data could be inadvertently inputted, or with enough contextual data (location, time, identity of user) the identity of a patient could be compromised. The models tested are not appropriate for real patient data due to the lack of data security.

Further, relying on LLMs to solve diagnostic cases raises questions around accountability in the event of errors, and without existing regulatory and ethical frameworks to support users, such tools may not be ready for formal integration into care pathways. However, as we have shown, the standard, consumer facing models available via their websites perform well, and the temptation to deploy such models remains – highlighting the urgency of the required frameworks.

Thus the exact role of such tools in a clinican's work remains uncertain, and the need for oversight, as well as potential deskilling of staff (in particular junior staff) due to potential for overreliance in challenging cases which could impede learning.

Future work should systemically assess a greater number of cases to assess broader generalisability outside of internal medicine. These cases were selected from challenging scenarios with final diagnoses presented at the conclusion of the case which facilitated a comparison of output to a "ground truth." An important translational step for future work will be to compare models in a clinical setting, with pragmatic analysis of applicability to real world cases assessed in real time, and where there is ambiguity as to the "true" diagnosis.

This research underscores the variation in model performance in clinical vignettes and highlights the importance of considering diagnostic model performance in diverse clinical scenarios. The findings suggest that model effectiveness varies based on the complexity of presented cases, and here we provide the community with a tool to help assess this output.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

MK: Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. EO'S: Conceptualization, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to thank Ben Strimaitis, Aoife O'Sullivan, Elizabeth Leong, Stephane Chang, Yi Chen Lim, Sahar Samini, Izza Shahid, and Daniel O'Sullivan for their independent review of cases and scoring the LLM output.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1379297/full#supplementary-material

References

Cascella, M., Montomoli, J., Bellini, V., and Bignami, E. (2023). Undefined. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J. Med. Syst.* 47:33. doi: 10.1007/s10916-023-01925-4

Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., and Cui, L. (2023). LLM-empowered Chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv*:13614. doi: 10.48550/arXiv.2305.13614

Chirino, A., Wiemken, T., Furmanek, S., Mattingly, W., Chandler, T., Cabral, G., et al. (2023). High consistency between recommendations by a pulmonary specialist and ChatGPT for the management of a patient with non-resolving pneumonia. *Norton Health Care Med. J.* 1. doi: 10.59541/001c.75456

Dhanvijay, A. K. D., Pinjar, M. J., Dhokane, N., Sorte, S. R., Kumari, A., and Mondal, H. (2023). Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology. *Cureus* 15:e42972. doi: 10.7759/cureus.42972

Giannakopoulos, K., Kavadella, A., Aaqel Salim, A., Stamatopoulos, V., and Kaklamanos, E. G. (2023). Evaluation of the performance of generative AI large language models ChatGPT, Google bard, and Microsoft Bing chat in supporting evidence-based dentistry: comparative mixed methods study. *J. Med. Internet Res.* 25:e51580. doi: 10.2196/51580

Gomez-Cabello, C. A., Borna, S., Pressman, S. M., Haider, S. A., and Forte, A. J. (2024). Large language models for intraoperative decision support in plastic surgery: a comparison between ChatGPT-4 and Gemini. *Med. Kaunas Lith.* 60:957. doi: 10.3390/medicina60060957

Huang, H., Zheng, O., Wang, D., Yin, J., Wang, Z., Ding, S., et al. (2023). ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. *Int. J. Oral Sci.* 15:29. doi: 10.1038/s41368-023-00239-y

Kleesiek, J., Wu, Y., Stiglic, G., Egger, J., and Bian, J. (2023). An opinion on ChatGPT in health care—written by humans only. *J. Nucl. Med.* 64, 701–703. doi: 10.2967/jnumed.123.265687

Kumari, A., Kumari, A., Singh, A., Singh, SK., Juhi, A., Dhanvijay, AKD., et al. (2023). Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google bard, and Microsoft Bing. *Cureus* 15:e43861. doi: 10.7759/cureus.43861

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., de Leon, L., Elepaño, C., et al. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit. Health* 2:e0000198. doi: 10.1371/journal.pdig.0000198

Lee, Y., Shin, T., Tessier, L., Javidan, A., Jung, J., Hong, D., et al. (2024a). Harnessing artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and bard in generating clinician-level bariatric surgery recommendations. *Surg. Obes. Relat. Dis. Off. J. Am. Soc. Bariatr. Surg.* 20, 603–608. doi: 10.1016/j.soard.2024.03.011

Lee, Y., Tessier, L., Brar, K., Malone, S., Jin, D., McKechnie, T., et al. (2024b). Performance of artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and bard in the American Society for Metabolic and Bariatric Surgery textbook of bariatric surgery questions. *Surg. Obes. Relat. Dis. Off. J. Am. Soc. Bariatr. Surg.* 20, 609–613. doi: 10.1016/j.soard.2024.04.014

Lim, Z. W., Pushpanathan, K., Yew, S. M. E., Lai, Y., Sun, C. H., Lam, J. S. H., et al. (2023). Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google bard. *EBioMedicine* 95:104770. doi: 10.1016/j.ebiom.2023.104770