Evaluating the Diagnostic Performance of Large Language Models on Complex Multi-Modal Medical Cases

Wan Hang Keith Chiu, Wei Sum Koel Ko, William Chi Shing Cho, Sin Yu Joanne Hui, Wing Chi Lawrence Chan, Michael D Kuo

Submitted to: Journal of Medical Internet Research
on: October 17, 2023

Table of Contents

Evaluating the Diagnostic Performance of Large Language Models on Complex Multi-Modal Medical Cases

Wan Hang Keith Chiu[1]; Wei Sum Koel Ko[1]; William Chi Shing Cho[1]; Sin Yu Joanne Hui[2]; Wing Chi Lawrence Chan[3]; Michael D Kuo[4]

[1]Queen Elizabeth Hospital Hong Kong HK
[2]The University of Hong Kong Hong Kong HK
[3]The Hong Kong Polytechnic University Hong Kong HK
[4]Ensemble Group Scottsdale US

Corresponding Author:
Michael D Kuo
Ensemble Group
Scottsdale
Scottsdale
US

Abstract

Background: Large language models (LLMs) have demonstrated surprising performance on radiological examinations (1). However, their proficiency in real-world medical reasoning, especially when integrating multi-modal data remains uncertain (2).

Objective: This study evaluates the ability of 3 commonly used LLMs (BARD, Claude2, and GPT-4) to generate differential diagnoses (ddx) from complex multi-modality diagnostic cases.

Methods: Consecutive “Case Records of the Massachusetts General Hospital” from 07/2020–06/2023 containing clinical, biochemical and radiological information were selected (3). The cases were diagnostically challenging with a final diagnosis provided. Only the case presentation and a simple prompt asking for top 5 ddx were used as input. Each case was run independently to prevent the model from being influenced by prior cases. To enable objective assessment, all diagnoses were mapped to their corresponding 10th revision International Classification of Diseases (ICD-10) codes, with higher-level codes used if an exact code could not be assigned (Table 1).

The primary objective was accuracy, measured by whether the final diagnosis was within the LLM-generated ddx at the ICD-10 Category level. The secondary objectives were to measure the similarity between diagnoses within a ddx and its final diagnosis, as well as their similarity to each other, measured at the ICD-10 Chapter level. Chi-square and ANOVA were used to compare categorical data between LLMs. Statistical analysis was performed using Prism 10 (GraphPad Software, USA).

Results: The diagnostic accuracies on 104 evaluated cases were 27.9%, 30.8% and 31.7% for BARD, Claude2 and GPT-4, respectively. Accuracy significantly improved at the ICD-10 Chapter (body site or system) level, reaching 65.3%, 66.3%, and 71.1% respectively.

All 3 LLMs showed evidence of interpretive reasoning as they tended to generate ddx whose member diagnoses were often related to each other (median number of ddx per case belonging to the same ICD-10 chapter as each other was 3.0 for all 3 LLMs (SD 1.1, 1.1 and 0.9 for BARD, Claude2 and GPT-4)). Interestingly, these related diagnosis “clusters” were often unrelated to the final diagnosis (median number of ddx belonging to the same ICD-10 chapter as the final diagnosis was 1.0 for all 3 LLMs, (SD 1.3, 1.4 and 1.2 for BARD, Claude2 and GPT-4)). These two findings were irrespective of whether the LLMs were able to include the final diagnosis in their ddx. Furthermore, performance of the LLMs varied by disease etiology although not statistically significant (Table 2).

Conclusions: This study rigorously evaluates the diagnostic capacity of multiple LLMs using a simple standardized prompt (4). The 3 LLMs represent state-of-the-art, general LLMs, accessible to most clinicians. The relatively low accuracies of all 3 models at the ICD-10 category level, coupled with a median of 3/5 diagnoses residing in a chapter outside of the final diagnosis chapter, together, suggests either a knowledge or reasoning gap in current LLMs. Conversely, the moderate number of LLM-generated

ddx belonging to the same body site/system (chapter) implies these models can integrate and reason across complex clinical findings.

Limitations include not assessing whether human-AI interaction or prompt engineering would affect diagnostic accuracy. Nevertheless, attempts to “overengineer” general LLMs towards a desired output could cloud real-world applicability, detracting from the ease-of-use that make current LLMs attractive to general users (5). Future work includes analyzing the rationales provided by the LLMs in reaching their ddx and asking the LLMs to quantify the likelihood of each ddx. Finally, the diversity of LLM-generated ddx warrant further exploration as it could potentially hamper patient management (6).

In conclusion, LLMs may have a role in enhancing physician diagnosis on complex, multi-modal clinical cases when applied judiciously. Clinical Trial: N/A

(JMIR Preprints 17/10/2023:53724)
DOI: https://doi.org/10.2196/preprints.53724

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Evaluating the Diagnostic Performance of Large Language Models on Complex Multi-Modal Medical Cases

Introduction: Large language models (LLMs) have demonstrated surprising performance on radiological examinations [1]. However, their proficiency in real-world medical reasoning, especially when integrating multi-modal data remains uncertain [2]. This study evaluates the ability of 3 commonly used LLMs (BARD, Claude2, and GPT-4) to generate differential diagnoses (ddx) from complex multi-modality diagnostic cases.

Methods: Consecutive "Case Records of the Massachusetts General Hospital" from 07/2020–06/2023 were selected [3]. The cases were diagnostically challenging and a final diagnosis provided. Approval from an institutional review board was not required due to the use of publicly available nonidentifiable data. Only the case presentation and a simple prompt asking for top 5 ddx were used as input. Each case was run independently to prevent the model from being influenced by prior cases. In order to evaluate the stability of the results, all cases were re-inputted into each LLM. To enable objective assessment, all diagnoses were mapped to their corresponding 10th revision International Classification of Diseases (ICD-10) codes, with higher-level codes used if an exact code could not be assigned (Table 1).

The primary objective was accuracy, measured by whether the final diagnosis was within the LLM-generated ddx at the ICD-10 Category level. The secondary objectives were to measure the similarity between diagnoses within a ddx and its final diagnosis, as well as their similarity to each other, measured at the ICD-10 Chapter level. Chi-square and ANOVA were used to compare categorical data between LLMs. Statistical analysis was performed using Prism 10 (GraphPad Software, USA).

Results: The diagnostic accuracies on 104 evaluated cases based on the first set of answers by the LLMs were 27.9%, 30.8% and 31.7% for BARD, Claude2 and GPT-4, respectively. Accuracy significantly improved at the ICD-10 Chapter (body site or system) level, reaching 65.3%, 66.3%, and 71.1% respectively. The median number of the same ddx generated in each case in the repeatability testing were 2.0 for all 3 LLMs (SD 1.1, 1.2 and 1.2 for BARD, Claude2 and GPT-4 respectively).

All 3 LLMs showed evidence of interpretive reasoning as they tended to generate ddx whose member diagnoses were often related to each other (median number of ddx per case belonging to the same ICD-10 chapter as each other was 3.0 for all 3 LLMs (SD 1.1, 1.1 and 0.9 for BARD, Claude2 and GPT-4)). Interestingly, these related diagnosis "clusters" were often unrelated to the final diagnosis (median number of ddx belonging to the same ICD-10 chapter as the final diagnosis was 1.0 for all 3 LLMs, (SD 1.3, 1.4 and 1.2 for BARD, Claude2 and GPT-4)). These two findings were irrespective of whether the LLMs were able to include the final diagnosis in their ddx. Furthermore, performance of the LLMs varied by disease etiology although not statistically significant (Table 2).

Discussion: This study rigorously evaluates the diagnostic capacity of multiple LLMs using a simple standardized prompt [4]. The 3 LLMs represent state-of-the-art, general LLMs, accessible to most clinicians. The relatively low accuracies of all 3 models at the ICD-10 category level, coupled with a median of 3/5 diagnoses residing in a chapter outside of the final diagnosis chapter, together, suggests either a knowledge or reasoning gap in current LLMs. While performance differences are seen between different types of disease etiology (e.g. 12.5% for Chapter III vs 63.6% for Chapter XIII in GPT4), the small numbers and unequal distribution of etiology precludes adequate analysis although this area warrants further investigation. Conversely, the moderate number of LLM-generated ddx belonging to the same body site/system (chapter) implies these models can integrate and reason across complex clinical findings.

This study has limitations, including the low reproducibility of the ddxs generated by the LLMs. The generative nature of these models and their continuous updates may lead to performance drift and contradictory results. Further research and validation are needed to generate consistent and explainable results, as well as exploring the relationships between performance and repeatability Secondly, we did not assess whether human-AI interaction or prompt engineering would affect diagnostic accuracy. Nevertheless, attempts to “overengineer” general LLMs towards a desired output could cloud real-world applicability, detracting from the ease-of-use that make current LLMs attractive to general users [5]. Future work includes analyzing the rationales provided by the LLMs in reaching their ddx and asking the LLMs to quantify the likelihood of each ddx. Finally, the diversity of LLM-generated ddx warrant further exploration as it could potentially hamper patient management [6].

In conclusion, LLMs may have a role in enhancing physician diagnosis on complex, multi-modal clinical cases when applied judiciously.

Word count = 750

Conflict of Interest: None declared.

References

1. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology 2023;307(5). doi: 10.1148/radiol.230582
2. Jamshidi N, Feizi A, Sirlin CB, Lavine JE, Kuo MD. Multi-Modality, Multi-Dimensional Characterization of Pediatric Non-Alcoholic Fatty Liver Disease. Metabolites 2023;13(8). doi: 10.3390/metabo13080929
3. Dougan M, Cabot RC, Rosenberg ES, Dudzinski DM, Baggett MV, Tran KM, Sgroi DC, Shepard J-AO, McDonald EK, Corpuz T, Anderson MA, Abramson JS, Fitzpatrick MJ. Case 14-2022: A 57-Year-Old Man with Chylous Ascites. New England Journal of Medicine 2022;386(19):1834-1844. doi: 10.1056/NEJMcpc2115856
4. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. Jama 2023;330(1). doi: 10.1001/jama.2023.8288
5. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, Heußel CP, Kauczor H-U, Weber TF. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. Radiology 2023;308(3). doi: 10.1148/radiol.231362
6. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, Landman A, Dreyer K, Succi MD. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. Journal of Medical Internet Research 2023;25. doi: 10.2196/48659

Supplementary Files