Research Letter

Evaluating the Diagnostic Performance of Large Language Models on Complex Multimodal Medical Cases

Wan Hang Keith Chiu[1]; Wei Sum Koel Ko[1]; William Chi Shing Cho[2], PhD; Sin Yu Joanne Hui[3]; Wing Chi Lawrence Chan[4], PhD; Michael D Kuo[5], MD

[1]Department of Diagnostic and Interventional Radiology, Queen Elizabeth Hospital, Hong Kong, China (Hong Kong)

[2]Department of Clinical Oncology, Queen Elizabeth Hospital, Hong Kong, China (Hong Kong)

[3]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China (Hong Kong)

[4]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[5]Ensemble Group, Scottsdale, AZ, United States

Corresponding Author:
Michael D Kuo, MD
Ensemble Group
10541 E Firewheel Drive
Scottsdale, AZ, 85259
United States
Phone: 1 4084512341
Email: mikedkuo@gmail.com

Abstract

Large language models showed interpretative reasoning in solving diagnostically challenging medical cases.

(J Med Internet Res 2024;26:e53724) doi: 10.2196/53724

KEYWORDS

large language model; hospital; health center; Massachusetts; statistical analysis; chi-square; ANOVA; clinician; physician; performance; proficiency; disease etiology

Introduction

Large language models (LLMs) have demonstrated a surprising performance in radiological examinations [1]. However, their proficiency in real-world medical reasoning, especially when integrating multimodal data remains uncertain [2]. This study evaluates the ability of 3 commonly used LLMs—Google Bard (subsequently rebranded Gemini), Claude 2, and GPT-4—to generate differential diagnoses (ddx) from complex multimodality diagnostic cases.

Methods

Overview

Consecutive case records of the Massachusetts General Hospital from July 2020 to June 2023 were selected [3]. The cases were diagnostically challenging, but a final diagnosis was provided. Only the case presentation and a simple prompt asking for the top 5 ddx were used as input. Each case was run independently to prevent the model from being influenced by prior cases. To evaluate the stability of the results, all cases were reinputted into each LLM. To enable objective assessment, all diagnoses were mapped to their corresponding *International Classification of Diseases, Tenth Revision* (*ICD-10*) codes, with higher-level codes used in case an exact code could not be assigned (Figure 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 the ddx and the final diagnosis as well as their similarity to each other, measured at the *ICD-10* chapter level. Chi-square and ANOVA tests were used to compare categorical data between the LLMs. Statistical analyses were performed using Prism 10 (GraphPad Software).

Figure 1. (A) Standardized prompt used for each case to generate differential diagnoses (ddx). (B) An example of *International Classification of Diseases, Tenth Revision* (*ICD-10*) code hierarchy structure; the first character (an alphabetical letter) denotes the chapter, and when combined with the next 2 digits, it forms the *ICD-10* category code. (C) An example of a large language model (LLM)–generated ddx and the corresponding *ICD-10* codes (case 34); in this case, none of the 3 LLMs included the final diagnosis (high-grade B-cell lymphoma, not otherwise specified; C83.30) in their ddx. For Bard, 3 of the 5 ddx belonged to the same chapter as the final diagnosis (chapter II: C22.0, C85.9, and C79.9). For Claude 2, only 1 of the 5 ddx belonged to the same chapter as the final diagnosis (chapter II: C85.9). For GPT-4, only 1 of the 5 ddx belonged to the same chapter as the final diagnosis (chapter II: C79.9).

(A) The prompt used at the outset of each case was as follows:

"Assuming I am an internist, please assume the role of an expert radiologist and provide the top 5 differential diagnoses for my patient based on the given presentation and investigative findings."

(B) An example of the *ICD-10* code hierarchy:

Chapter (A00-B99): Certain infectious and parasitic diseases
Section (A00-A09): Certain infectious and parasitic diseases
Category (A00): Cholera

(C) Example of final diagnosis (*ICD-10* category), and LLMs ddx with their corresponding *ICD-10* codes:

Final diagnosis (corresponding *ICD-10* code)	Top 5 ddx (corresponding *ICD-10* codes) by Bard	Top 5 ddx (corresponding *ICD-10* codes) by Claude 2	Top 5 ddx (corresponding *ICD-10* codes) by GPT4
High-grade B-cell lymphoma, not otherwise specified (C83.30)	1. Primary sclerosing cholangitis (K83.01) 2. Hepatocellular carcinoma (C22.0) 3. Lymphoma (C85.9) 4. Metastases (C79.9) 5. Sarcoidosis (D86.9)	1. Lymphoma (C85.9) 2. Sarcoidosis (D86.9) 3. Mycobacterial infection (A31.9) 4. Whipple's disease (K90.81) 5. Autoimmune disorder (M35.9)	1. Lymphatic disorders, such as Lymphangiectasia (I89.0) 2. Metastatic malignancy (C79.9) 3. Chronic granulomatous disease (eg, sarcoidosis; D86.9) 4. Chronic inflammatory bowel disease (eg, Crohn's disease; K50.90) 5. Infections, such as tuberculosis (A18.32)

Ethics Approval

Approval from an institutional review board was not required due to the use of publicly available nonidentifiable data.

Results

The diagnostic accuracy on 104 evaluated cases based on the first set of answers by the LLMs was 27.9% for Bard, 30.8% for Claude 2, and 31.7% for GPT-4. Accuracy significantly improved at the *ICD-10* chapter (body site or system) level, reaching 65.4% for Bard, 66.3% for Claude 2, and 71.2% for GPT-4. The mean number of the same ddx generated in each case in the repeatability testing was 2.3 (SD 1.1) for Bard, 2.4 (SD 1.2) for Claude 2, and 2.4 (SD 1.2) for GPT-4.

All 3 LLMs showed evidence of interpretive reasoning, as they tended to generate sets of ddx whose member diagnoses were often related to each other. The mean number of ddx per case belonging to the same *ICD-10* chapter as each other was 2.6 (SD 1.1) for Bard, 2.7 (SD 1.1) for Claude 2, and 2.4 (SD 0.9) for GPT-4. Interestingly, these related diagnosis "clusters" were often unrelated to the final diagnosis. The mean number of ddx belonging to the same *ICD-10* chapter as the final diagnosis was 1.2 (SD 1.3) for Bard, 1.4 (SD 1.4) for Claude 2, and 1.4 (SD 1.2) for GPT-4. These two findings were irrespective of whether the LLMs could include the final diagnosis in their ddx. Furthermore, the performance of the LLMs varied by disease etiology, although this difference was not statistically significant (Table 1).

Table 1. Performance of individual large language models (LLMs).

Characteristics	Bard	Claude 2	GPT4	*P* value
Accuracy by *ICD-10*[a] hierarchy level, %				
Category	27.9	30.7	30.7	<.001[b]
Chapter	65.4	66.3	71.2	<.001[b]
Accuracy by *ICD-10* etiology (top 5 by frequency), n (%)				
Certain infectious and parasitic diseases (chapter I: A00-B99)	20 (35.0)	45.0	50.0	.62[c]
Neoplasm (chapter II C00-D48)	19 (52.6)	63.2	57.9	.75[c]
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (chapter III: D50-D89)	8 (12.5)	25.0	12.5	.74[c]
Endocrine, nutritional, and metabolic diseases (chapter IV: E00-E90)	9 (33.3)	33.3	33.3	>.99[c]
Diseases of the musculoskeletal system and connective tissue (chapter XIII: M00-M99)	11 (36.4)	72.7	63.6	.20[c]
Number of diagnoses per ddx[d] per case generated by LLMs belonging to the same hierarchical chapter as the final diagnosis based on assigned *ICD-10* codes, mean (SD)	1.2 (1.3)	1.4 (1.4)	1.4 (1.2)	—[e]
Number of diagnoses per ddx per case generated by LLMs belonging to the same hierarchical chapter based on assigned *ICD-10* codes, mean (SD)	2.6 (1.1)	2.7 (1.1)	2.4 (0.9)	—
Number of the same ddx per case generated by LLMs on repeatability testing, mean (SD)	2.3 (1.1)	2.4 (1.2)	2.4 (1.2)	—

[a]*ICD-10*: *International Classification of Diseases, Tenth Revision*.

[b]Comparison of each LLM's performance at the *ICD-10* category level versus the chapter level.

[c]Comparison of each LLM's performance across different *ICD-10* etiologies. *P* values were not significant.

[d]ddx: differential diagnoses.

[e]Not applicable.

Discussion

This study rigorously evaluated 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 accuracy of all 3 models at the *ICD-10* category level, coupled with a mean of >3 out of 5 diagnoses located in a chapter outside the final diagnosis chapter, collectively suggest either a knowledge or reasoning gap in current LLMs. Although performance differences are observed between different types of disease etiology (eg, 12.5% for Chapter III vs 63.6% for Chapter XIII in GPT4), the small numbers and unequal distribution of etiologies preclude adequate analysis; however, this area warrants further investigation. Conversely, the moderate number of LLM-generated ddx belonging to the same body site or system (chapter) implies these models can integrate and reason across complex clinical findings.

This study has limitations, including the low reproducibility of the ddx generated by the LLMs. The generative nature of these models and their continuous updates may lead to performance drifts and contradictory results. Further research and validation are necessary to generate consistent and explainable results as well as explore the relationships between performance and repeatability. Second, we did not assess whether human-artificial intelligence interaction or prompt engineering would affect diagnostic accuracy. Nevertheless, attempts to "overengineer" general LLMs toward a desired output could cloud real-world applicability, detracting from the ease of use that makes 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 warrants further exploration, as it could potentially hamper patient management [6].

In conclusion, LLMs may have a role in enhancing physician diagnosis of complex, multimodal clinical cases when applied judiciously.

Conflicts 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. Jun 01, 2023;307(5):e230582. [doi: 10.1148/radiol.230582] [Medline: 37191485]
2. Jamshidi N, Feizi A, Sirlin CB, Lavine JE, Kuo MD. Multi-modality, multi-dimensional characterization of pediatric non-alcoholic fatty liver disease. Metabolites. Aug 08, 2023;13(8):929. [FREE Full text] [doi: 10.3390/metabo13080929] [Medline: 37623872]
3. Dougan M, Anderson MA, Abramson JS, Fitzpatrick MJ. Case 14-2022: a 57-year-old man with chylous ascites. N Engl J Med. May 12, 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. Jul 03, 2023;330(1):78-80. [FREE Full text] [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
5. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology. Sep 2023;308(3):e231362. [doi: 10.1148/radiol.231362] [Medline: 37724963]
6. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res. Aug 22, 2023;25:e48659. [FREE Full text] [doi: 10.2196/48659] [Medline: 37606976]

Abbreviations

Ddx: differential diagnoses
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
LLM: large language model

Edited by T de Azevedo Cardoso, G Eysenbach; submitted 17.10.23; peer-reviewed by R Yang, L Zhu, H Mondal; comments to author 05.02.24; revised version received 22.02.24; accepted 23.04.24; published 13.05.24

Please cite as:
Chiu WHK, Ko WSK, Cho WCS, Hui SYJ, Chan WCL, Kuo MD
Evaluating the Diagnostic Performance of Large Language Models on Complex Multimodal Medical Cases
J Med Internet Res 2024;26:e53724
URL: https://www.jmir.org/2024/1/e53724
doi: 10.2196/53724
PMID: 38739441