World Journal of Methodology

Submit a Manuscript: https://www.f6publishing.com

World J Methodol 2024 December 20; 14(4): 92802

DOI: 10.5662/wjm.v14.i4.92802

ISSN 2222-0682 (online)

ORIGINAL ARTICLE

Observational Study

Comparative evaluation of artificial intelligence systems' accuracy in providing medical drug dosages: A methodological study

Swaminathan Ramasubramanian, Sangeetha Balaji, Tejashri Kannan, Naveen Jeyaraman, Shilpa Sharma, Filippo Migliorini, Suhasini Balasubramaniam, Madhan Jeyaraman

Specialty type: Medical laboratory technology

Provenance and peer review: Invited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's classification

Scientific Quality: Grade C

Novelty: Grade B

Creativity or Innovation: Grade B

Scientific Significance: Grade B

P-Reviewer: Viswanathan VK

Received: February 6, 2024

Revised: May 29, 2024

Accepted: June 25, 2024

Published online: December 20, 2024

Processing time: 171 Days and 3.9 Hours

Swaminathan Ramasubramanian, Sangeetha Balaji, Tejashri Kannan, Department of Orthopaedics, Government Medical College, Omandurar Government Estate, Chennai 600002, Tamil Nadu, India

Naveen Jeyaraman, Madhan Jeyaraman, Department of Orthopaedics, ACS Medical College and Hospital, Dr MGR Educational and Research Institute, Chennai 600077, Tamil Nadu, India

Shilpa Sharma, Department of Paediatric Surgery, All India Institute of Medical Sciences, New Delhi 110029, India

Filippo Migliorini, Department of Life Sciences, Health, Link Campus University, Rome 00165, Italy

Filippo Migliorini, Department of Orthopaedic and Trauma Surgery, Academic Hospital of Bolzano (SABES-ASDAA), Teaching Hospital of the Paracelsus Medical University, Bolzano 39100, Italy

Suhasini Balasubramaniam, Department of Radio-Diagnosis, Government Stanley Medical College and Hospital, Chennai 600001, Tamil Nadu, India

Corresponding author: Madhan Jeyaraman, MS, PhD, Assistant Professor, Research Associate, Department of Orthopaedics, ACS Medical College and Hospital, Dr MGR Educational and Research Institute, Velappanchavadi, Chennai 600077, Tamil Nadu, India.
madhanjeyaraman@gmail.com

Abstract

BACKGROUND

Medication errors, especially in dosage calculation, pose risks in healthcare. Artificial intelligence (AI) systems like ChatGPT and Google Bard may help reduce errors, but their accuracy in providing medication information remains to be evaluated.

AIM

To evaluate the accuracy of AI systems (ChatGPT 3.5, ChatGPT 4, Google Bard) in providing drug dosage information per Harrison's Principles of Internal Medicine.

METHODS
A set of natural language queries mimicking real-world medical dosage inquiries was presented to the AI systems. Responses were analyzed using a 3-point Likert scale. The analysis, conducted with Python and its libraries, focused on basic statistics, overall system accuracy, and disease-specific and organ system accuracies.

RESULTS
ChatGPT 4 outperformed the other systems, showing the highest rate of correct responses (83.77%) and the best overall weighted accuracy (0.6775). Disease-specific accuracy varied notably across systems, with some diseases being accurately recognized, while others demonstrated significant discrepancies. Organ system accuracy also showed variable results, underscoring system-specific strengths and weaknesses.

CONCLUSION
ChatGPT 4 demonstrates superior reliability in medical dosage information, yet variations across diseases emphasize the need for ongoing improvements. These results highlight AI's potential in aiding healthcare professionals, urging continuous development for dependable accuracy in critical medical situations.

Key Words: Dosage calculation; Artificial intelligence; ChatGPT; Drug dosage; Healthcare; Large language models

©The Author(s) 2024. Published by Baishideng Publishing Group Inc. All rights reserved.

Core Tip: This study reveals ChatGPT 4's superior accuracy in providing medical drug dosage information, highlighting the potential of artificial intelligence (AI) to aid healthcare professionals in minimizing medication errors. The analysis, based on Harrison's Principles of Internal Medicine, underscores the need for ongoing AI development to ensure reliability in critical medical situations. Variations in disease-specific and organ system accuracies suggest areas for improvement and continuous refinement of AI systems in medicine.

Citation: Ramasubramanian S, Balaji S, Kannan T, Jeyaraman N, Sharma S, Migliorini F, Balasubramaniam S, Jeyaraman M. Comparative evaluation of artificial intelligence systems' accuracy in providing medical drug dosages: A methodological study. *World J Methodol* 2024; 14(4): 92802
URL: https://www.wjgnet.com/2222-0682/full/v14/i4/92802.htm
DOI: https://dx.doi.org/10.5662/wjm.v14.i4.92802

INTRODUCTION

ChatGPT is a Large Language Model (LLM) developed by open artificial intelligence (AI). The GPT (Generative Pre-trained Transformer) architecture, more precisely GPT-3.5, is the foundation of this system[1]. This model is trained on a massive quantity of text data to comprehend and produce responses to user inputs that are human-like. It has potential uses in medical practice, research, and teaching[2]. Through curriculum development, simulated training, and language translation, ChatGPT could contribute to medical education. It could aid with information retrieval in research, and it could potentially enhance the accuracy and efficiency of medical recording in clinical settings[3]. The most sophisticated algorithm in Open AI, GPT-4, generates safer and more insightful responses[4]. Because of its superior general knowledge and problem-solving skills, it can more accurately tackle complex problems. It is more collaborative and innovative than ever. On activities involving creative and technical writing, it can produce, edit, and revise alongside users[5]. Google Bard is a similar experimental conversation AI service that is powered by Language Model for Dialogue Applications. It draws on information from the Web to provide fresh, high-quality responses to queries[6].

Errors in medication administration pose a serious risk to patients, especially in the emergency room. In this situation, over-dosage has been recognized as the most frequent drug mistake. Approximately 20% of IV medication orders differed by more than 10% from the suggested dose according to research done with emergency medicine residents[7]. One-third of pharmaceutical errors in hospitals involved patients under the age of 18, with newborns accounting for half of the occurrences according to a different study on numeracy errors in hospitals. Overdoses were the most common mistake type, mostly occurring *via* parenteral (IV) (77%) or oral (20%) routes[8]. Research also examined critical-care nurses' knowledge and its relationship to medication errors, identifying risk areas such as antibiotic administration intervals, high-risk medication dilution, concentration, infusion-rate errors, and administration of medications *via* nasogastric tubes [9,10]. In the context of pediatric residents, a study demonstrated that a substantial number of trainees made drug dose calculation errors, including life-threatening mistakes. Interestingly, there was no correlation between training length and the likelihood of errors[11].

LLMs, like ChatGPT and Bard, have the potential to help medical professionals minimize dosage errors by acting as knowledgeable resources that improve healthcare providers' understanding of dosage calculations and reduce errors from misinterpretation or insufficient knowledge. In this context, our study evaluates the reliability of LLMs, namely ChatGPT 3.5, ChatGPT 4, and Google's Bard, in providing accurate drug dosage information per Harrison's Principles of

Internal Medicine. The goal is to determine the credibility of these AI systems in minimizing dosage errors in clinical practice. Ultimately, the study aims to evaluate their ability to enhance patient safety and treatment efficacy by reducing dosage errors.

MATERIALS AND METHODS

To conduct a thorough and comparative evaluation of the reliability of AI systems in supplying medical drug dosages, a comprehensive methodology was followed. Initially, data regarding the use of drugs, their dosages, and routes of administration for various medical conditions were meticulously gathered from the 21st edition of Harrison's Principles of Internal Medicine[12]. This established a robust foundation for accuracy comparison. The next step involved the construction of queries in a natural language format. These queries were designed to closely mimic how medical professionals might inquire about drug dosages and administration in real-world scenarios. This approach ensured the relevance and applicability of the AI systems' responses to practical medical situations.

For the actual interaction with the AI systems, three platforms were selected: ChatGPT 3.5, ChatGPT 4, and Google's Bard. Each system was presented with an identical set of questions. The responses were then collected systematically for subsequent analysis. The evaluation of these responses was based on a 3-point Likert scale. Responses were scored as +1 for correct responses that aligned accurately with the information in Harrison's Textbook, 0 for neutral responses that were neither fully correct nor incorrect, and -1 for incorrect responses that contained misinformation or significant inaccuracies.

For data analysis, Python, along with its associated libraries such as SciPy, NumPy, and Matplotlib, was utilized. The analysis focused on several aspects. Basic statistics involved counting and categorizing the responses from each system and calculating the percentage of affirmative ('Yes') responses. The overall system accuracy was assessed by calculating weighted accuracies using the Likert scale scores. Additionally, disease-specific accuracy and body system accuracy were also analyzed and compared among the systems, providing insights into each system's strengths and weaknesses in various medical domains. The results were then presented in a user-friendly format, employing bar charts and tables for a clear visual representation of the data. This approach allowed for an effective comparative analysis of the performance of the three AI systems across different categories, thus providing a comprehensive understanding of their reliability in medical settings.

RESULTS

When examining the count of each type of response ('Neutral', 'No', 'Yes', which correspond to a score of +1, 0, and -1 respectively) for each system, the total responses for each system were consistent at 462. Among the three systems, GPT 4 demonstrated the highest percentage of 'Yes' responses with a staggering 83.77%. GPT 3.5 trailed closely behind with 77.06% 'Yes' responses, while Bard registered the lowest with 54.76%. Table 1 displays the distribution of 'Yes' responses demarcating the superior performance of GPT 4 relative to its counterparts. Table 2 succinctly represents the comparative performance of each model, indicating that ChatGPT 4 achieved the highest weighted accuracy at 0.6775, followed by ChatGPT 3.5 with 0.5519, and Bard with 0.3745. Figures 1 and 2 depict a bar chart representation of the same.

Upon analyzing the accuracy of different diseases, as illustrated in Table 3 and Figure 3, notable variations were observed. For instance, while all systems exhibited complete accuracy (1.0) in recognizing diseases like 'Acromegaly', 'Myoclonus', and 'Nephrolithiasis', they faltered on others. Notably, 'Myotonic dystrophy' saw a drastic drop in ChatGPT 3.5's accuracy to -1.0, in stark contrast to ChatGPT 4 and Bard, both of which maintained a 1.0 score. At the bottom of the performance spectrum, diseases such as 'Multiple sclerosis' and 'Hypothyroidism' registered -1.0 across both ChatGPT versions, with Bard following suit for 'Multiple sclerosis'.

Table 4 and Figure 4 depict the weighted accuracy evaluation for each organ system. Intriguing results surfaced with 'Cardiovascular system', 'Respiratory system', and 'Hematology' securing full scores across all systems. 'Infectious Diseases' and 'Immune system' showcased noticeable differences across the systems, with Bard especially lagging in the former with a score of 0.2059. The 'Gastrointestinal tract' presented consistent challenges across all systems, with no system surpassing a score of 0.5385.

DISCUSSION

The integration of LLMs like ChatGPT into medical information provision and clinical decision-making has been the focus of recent research. These studies, ranging from benchmarking LLMs against each other to comparing them with physicians and textbooks, provide insights into their potential utility and limitations in healthcare. Our study evaluated the accuracy of ChatGPT 3.5, ChatGPT 4, and Google Bard in providing drug dosage information. ChatGPT 4 achieved the highest correct response rate (83.77%) and weighted accuracy (0.6775), outperforming ChatGPT 3.5 (77.06% correct, 0.5519 accuracy) and Bard (54.76% correct, 0.3745 accuracy). Disease-specific analysis showed perfect scores for 'Acromegaly', 'Myoclonus', and 'Nephrolithiasis' across all systems, but significant variability for others like 'Myotonic dystrophy' and 'Multiple sclerosis'. Organ system analysis indicated high performance in 'Cardiovascular system, respiratory system', and 'Hematology', but lower and variable accuracy in 'Infectious Diseases' and 'Gastrointestinal tract'.

Table 1 Total count and percentage of 'Yes' responses for each Large Language Model

System	Neutral	No	Yes	Total
GPT 4	1	74	387 (83.77)	462
GPT 3.5	5	101	356 (77.06)	462
Bard	129	80	253 (54.76)	462

Data are *n* (%).

Table 2 Weighted accuracy comparison across the Large Language Model

Model	Weighted accuracy
ChatGPT 4	0.6775
ChatGPT 3.5	0.5519
Bard	0.3745

Table 3 Disease accuracy comparison

Name of disease	ChatGPT 4	ChatGPT 3.5	Bard
Acromegaly	1.0	1.0	1.0
Orthostatic hypotension	1.0	1.0	0.0
Myasthenia gravis	1.0	1.0	0.5
Myoclonus	1.0	1.0	1.0
Myotonic dystrophy	1.0	-1.0	1.0
Neonatal onset multisystem inflammatory disease	1.0	1.0	1.0
Neoplastic spinal cord compression	1.0	1.0	1.0
Nephrolithiasis	1.0	1.0	1.0
Neurological infections	1.0	1.0	0.0
Neuromyelitis optica	1.0	0.0	1.0
Thiamine deficiency	-1.0	1.0	1.0
Anaphylactic reaction	-1.0	-1.0	0.0
Reactive arthritis	-1.0	-1.0	0.0
Fibrous dysplasia	-1.0	-1.0	1.0
Hypothyroidism	-1.0	-1.0	-1.0
Multiple sclerosis	-1.0	-1.0	-1.0
Hypophosphatemia	-1.0	-1.0	1.0
Hypomagnesemia	-1.0	-1.0	0.0
Alcohol intoxication	-1.0	-1.0	0.0
Post-concussive state	-1.0	-1.0	0.0

Benchmarking studies

Lim *et al*[13] observed that ChatGPT-4.0 demonstrated superior accuracy (80.6% 'good' responses) in providing information on myopia care, outperforming both its predecessor, ChatGPT-3.5, and Google Bard. Particularly notable was its performance in 'treatment and prevention', along with its self-correction capabilities[13]. Similarly, O'Hagan *et al*[14] found that ChatGPT's responses to queries about Alopecia areata had a mean accuracy score of 4.41 out of 5, with ChatGPT 4.0 performing slightly better than 3.5. This study highlighted the variability of inappropriateness and accuracy based on the nature of the questions. Our study showed similar results, confirming ChatGPT-4.0's superior accuracy and reliability in medical information provision.

Table 4 Detailed accuracy values for each organ system across the three Large Language Model

Organ system	ChatGPT 4	ChatGPT 3.5	Bard
Cardio vascular system, respiratory system	1.0000	0.6667	0.6667
Hematology	1.0000	1.0000	-1.0000
Respiratory	1.0000	0.3333	0.3333
Respiratory system	1.0000	1.0000	0.5000
Infectious diseases	0.8039	0.7451	0.2059
Immune system	0.6752	0.4188	0.2650
Central nervous system	0.6585	0.6220	0.5610
Hematological malignancies	0.6429	0.5714	0.4286
Cardio vascular system	0.6000	0.6667	0.3333
Endocrine system	0.5556	0.4444	0.5714
Renal	0.5556	0.3704	0.5185
Gastrointestinal tract	0.5385	0.2308	0.2308

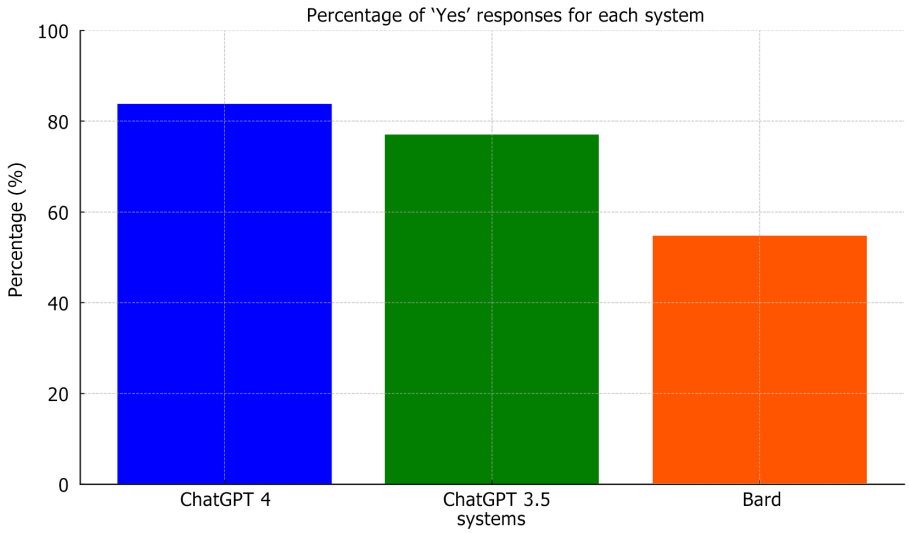


Figure 1 Percentage of 'Yes' responses for each Large Language Model.

Comparative efficacy of ChatGPT and physicians

Contrasting the responses generated by ChatGPT with those provided by medical professionals reveals favorable outcomes. However, the importance of verification with trusted sources and the need for further refinement of these models for clinical use are consistently emphasized across these studies.

Johnson *et al*[15] reported that the responses of ChatGPT to a wide range of medical questions had a median accuracy score of 5.5 on a six-point Likert scale, indicating a high level of correctness. However, they emphasized the need for ongoing research and development for clinical applications[15]. The article by Ramasubramanian S, focused on the application of ChatGPT for enhancing patient safety and operational efficiency in medical settings. The study found that a customized version of the model, text-davinci-003, had 80% accuracy compared to textbook medical knowledge, underscoring its potential as a tool for medical professionals[16].

In the field of endodontics, Mayo-Wilson *et al*[17] assessed the reliability of ChatGPT, finding an overall consistency of 85.44% in responses. The study revealed no significant variations based on question difficulty, although the accuracy rate for simpler queries was lower[17]. Hirosawa *et al*[18] evaluated the diagnostic accuracy of ChatGPT-3, noting an impressive 93.3% accuracy in differential-diagnosis lists for clinical vignettes, albeit slightly lower than the performance of physicians[18]. Hsu *et al*[19] conducted a prospective cross-sectional analysis to assess ChatGPT's accuracy and suitability in medication consultation responses. The study distinguished between real-world medication inquiries and questions regarding the interaction between traditional Chinese and Western medicines. ChatGPT showed a higher appropriateness rate for public medication consultation questions compared to those from healthcare providers in a hospital setting[19]. Finally, Rao *et al*[20] evaluated ChatGPT's effectiveness in clinical decision support, finding an overall accuracy of 71.7% across various clinical vignettes. This study highlighted the model's higher accuracy in making final

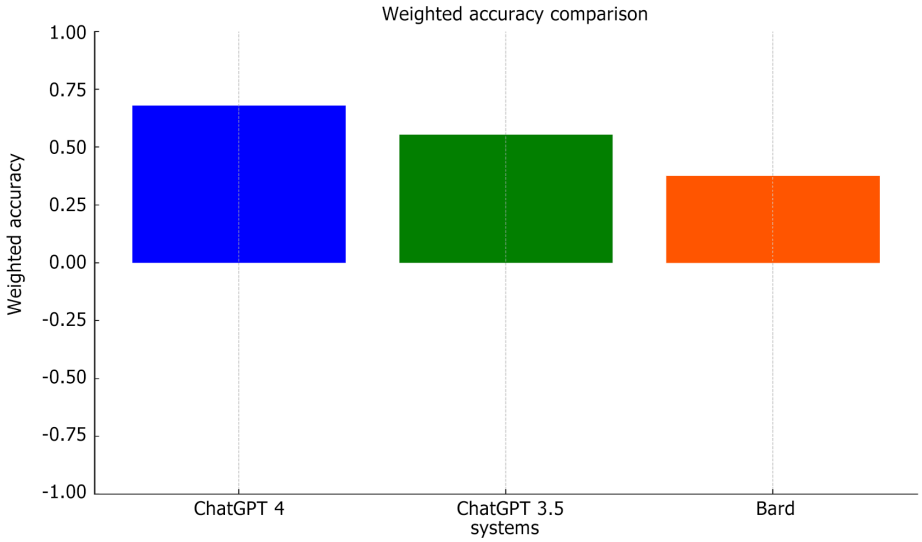


Figure 2 Weighted accuracy comparison across the Large Language Model.

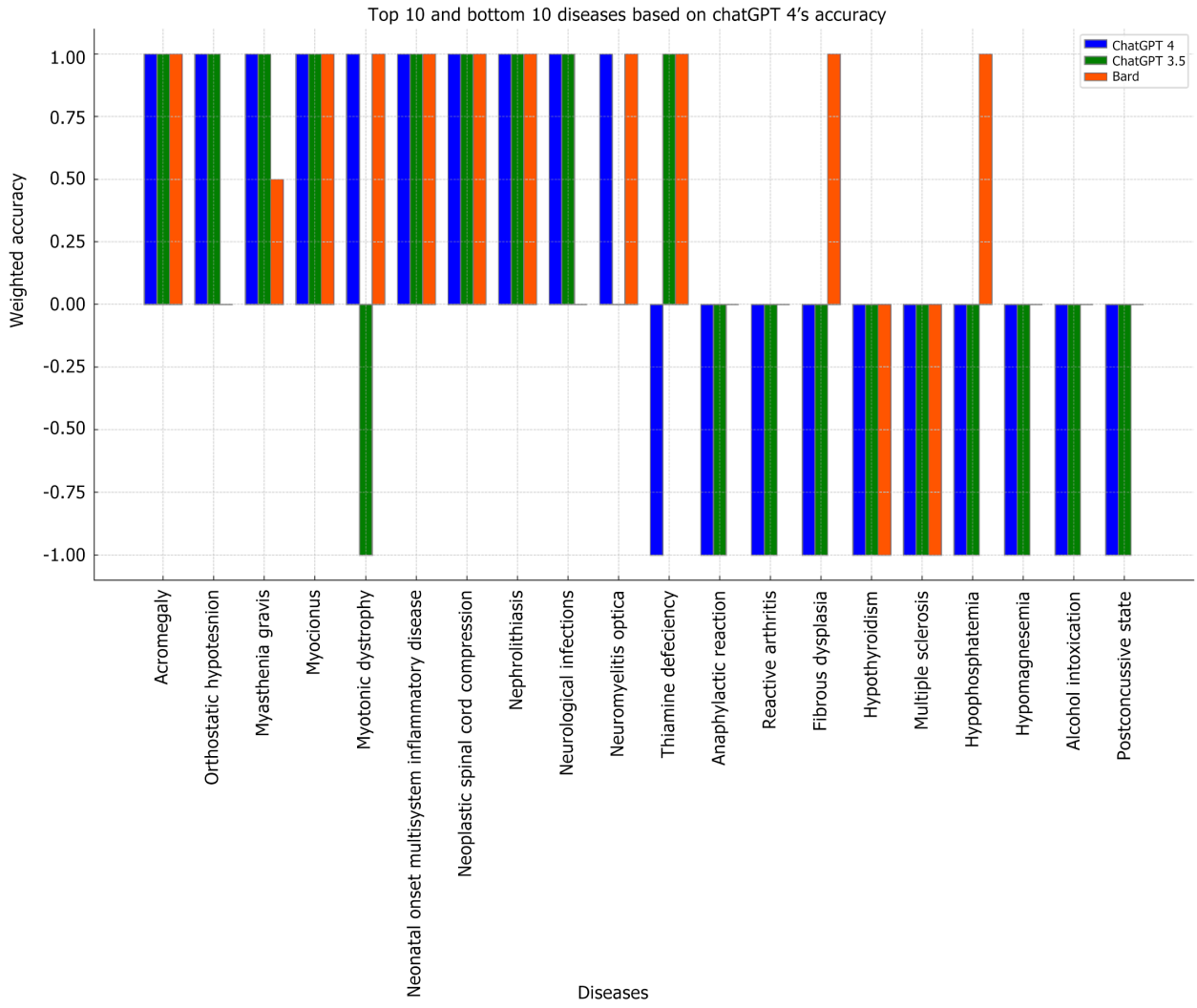


Figure 3 Weighted accuracy for each disease across the three systems (top 10 and bottom 10 based on ChatGPT 4's accuracy).

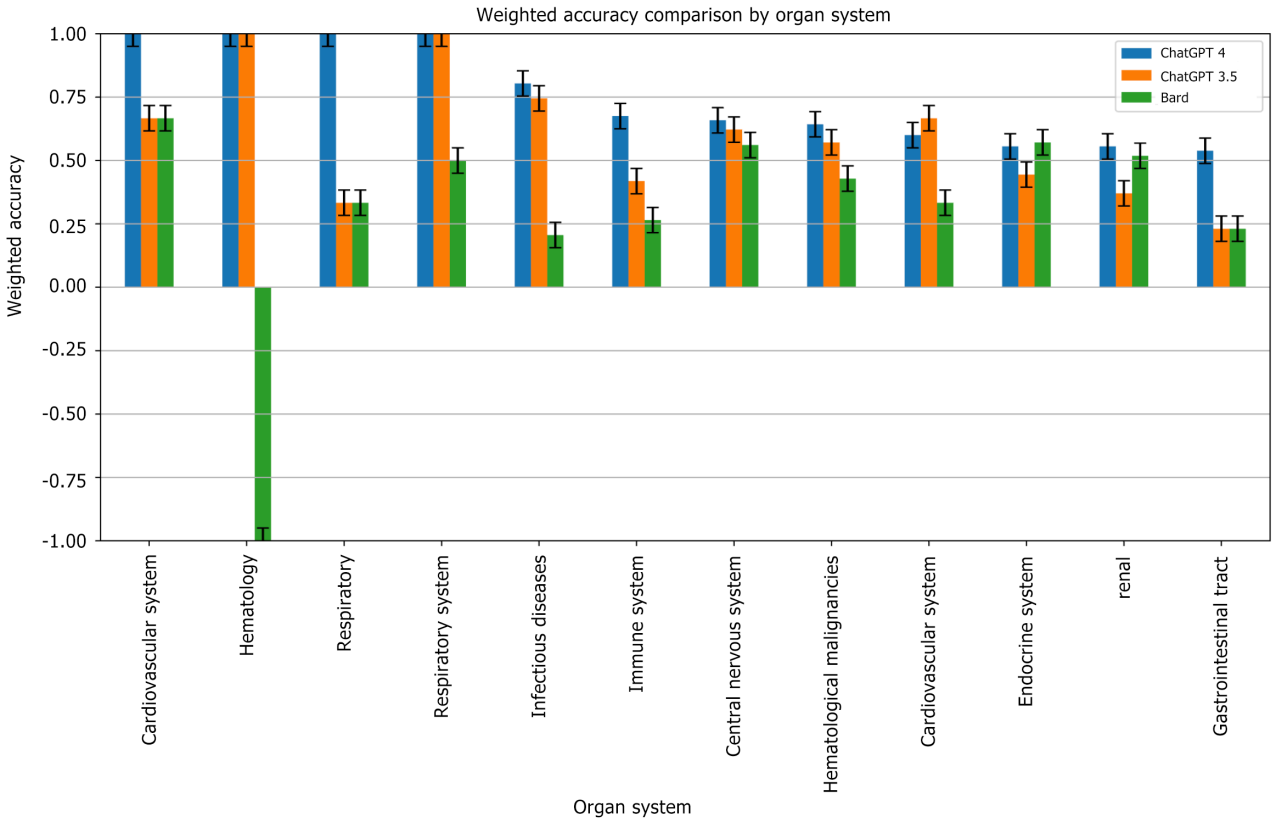


Figure 4 Weighted accuracy for each organ system across the three Large Language Model.

diagnoses compared to initial differential diagnoses and clinical management[20].

These studies collectively indicate a trend towards improved accuracy and appropriateness of responses with newer iterations of ChatGPT. They also highlight the potential of LLMs in supplementing medical professionals, especially in information provision and decision-making processes. Our study aligns with the current discourse on accuracy, showing that ChatGPT 4 achieved the highest rate of correct responses at 83.77% and demonstrated the best overall weighted accuracy of 0.6775 among other LLMs.

ChatGPT exhibits several limitations that warrant consideration (Figure 5). First, continuous improvement and ongoing evaluation are crucial for maintaining and enhancing the accuracy of LLMs, emphasizing the need for further research and model development[13,15,18,21]. Second, not all responses generated by ChatGPT are guaranteed to be accurate, underscoring the recommendation to use the model within the context of a comprehensive evaluation by a licensed healthcare professional and the importance of verifying information with trusted sources[14-16,21-23]. Third, the model may experience hallucinations, and its training data composition remains unclear, leading to errors in its outputs [19,20,24,25]. Lastly, reservations regarding ethics, legality, privacy, data accuracy, learning, and bias risks pose additional concerns, highlighting the need for vigilant consideration and responsible use of ChatGPT in various applications[24,26].

Several studies have explored the perception and application of ChatGPT in pharmacy practice, addressing pharmacists' opinions, its performance in clinical pharmacy, and its consistency in managing pharmacotherapy cases[22,24,27]. Abu Hammour *et al*[24] found mixed responses among Jordanian pharmacists, with concerns about accuracy, biases, and ethical, legal, and privacy issues[24]. Huang *et al*[27] demonstrated ChatGPT's strong performance in drug counseling but identified limitations in prescription review, patient education, adverse drug reaction recognition, and causality assessment[27]. Al-Dujaili *et al*[22] reported varying accuracy and consistency over time, suggesting potential in generating clinically relevant information but emphasizing the need for ongoing development and evaluation in pharmacy practice[22]. Despite the promise, pharmacists' concerns highlight the importance of cautious integration of AI technologies in healthcare. Wang *et al*[28] examined ChatGPT's potential in public health initiatives in low- and middle-income countries, highlighting its accessibility and benefits for health literacy and training. The study raised concerns about variable performance, accuracy, safety, and regulatory needs, emphasizing domain-specific pre-training, privacy, data protection, and ethical considerations, advocating for a cautious, expert-supported approach[28].

Recent studies highlight the potential of AI models in identifying and predicting drug-drug interactions (DDIs) and comparing their performance to human healthcare professionals. Al-Ashwal *et al*[21] found that ChatGPT-3.5 had the lowest accuracy (0.469), with slight improvements in ChatGPT-4, while Microsoft Bing AI was the most accurate, outperforming ChatGPT-3.5, ChatGPT-4, and Bard. ChatGPT models identified more potential interactions but had higher false positives, negatively affecting accuracy and specificity[21]. Juhi *et al*[29] reported that ChatGPT accurately identified 39 out of 40 DDI pairs consistently across question types but noted that responses were above the recommended reading level, suggesting a need for simplification for better patient comprehension[29].

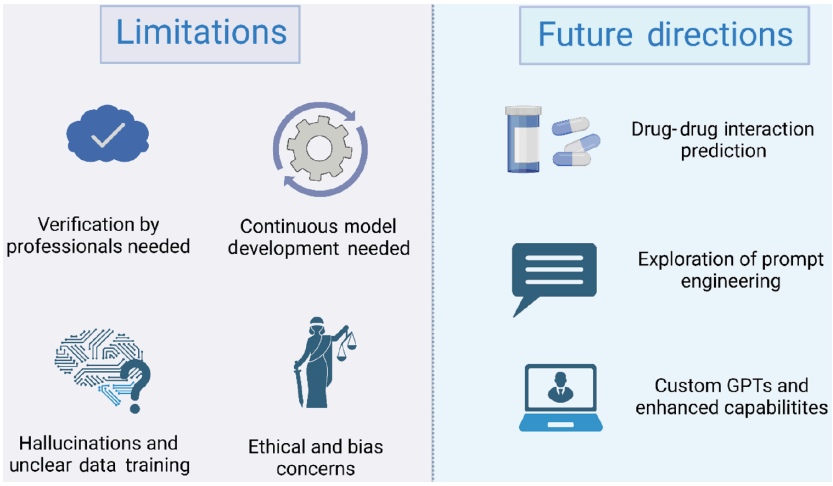


Figure 5 Limitations and future avenues of research of Large Language Model.

Shin's study systematically explores the use of appropriate prompt engineering while evaluating ChatGPT's performance in tasks related to pharmacokinetic data analysis, including report writing, code generation, and problem-solving. While ChatGPT displayed satisfactory proficiency in generating reports and code related to pharmacokinetics, it exhibited notable errors in numerical calculations involving exponential functions. The model's outputs were characterized by variability and lacked reproducibility for the same prompt. Although prompt engineering strategies were effective in reducing errors, they did not eliminate them[25].

Open AI has recently introduced a feature enabling users to create customized versions of ChatGPT by blending specific instructions, additional knowledge, and a combination of desired skills. Additionally, they have unveiled the GPT-4 Turbo model, which boasts enhanced capabilities, cost-effectiveness, and support for a 128K context window[30,31]. Google recently introduced a novel AI model named 'Gemini,' exhibiting capabilities comparable to GPT-4. Distinguished by its multimodal functionality, Gemini possesses the capacity to adeptly generalize and seamlessly comprehend, manipulate, and integrate diverse forms of information encompassing text, code, audio, images, and video [32,33].

As depicted in Figure 5, the advancing field of AI models in DDI prediction and the exploration of prompt engineering and custom GPTs represent promising avenues for further research and development. Continued investigations into these areas hold the potential to enhance the accuracy, specificity, and overall capabilities of AI tools in healthcare settings, contributing to improved patient care and safety.

CONCLUSION

Our research offers insights into the accuracy of AI systems (ChatGPT 3.5, ChatGPT 4, and Google Bard) in providing drug dosage information aligned with Harrison's Principles of Internal Medicine. ChatGPT 4 surpassed the other systems, demonstrating the highest rate of accurate responses and achieving the highest overall weighted accuracy. Nonetheless, the variability observed in disease-specific and organ system accuracies across different AI platforms highlights the ongoing need for refinement and development. While ChatGPT 4 demonstrates superior reliability, continuous improvement is essential to ensure consistent accuracy in diverse medical scenarios. These findings affirm the promising role of AI in supporting healthcare professionals, emphasizing the critical importance of enhancing AI systems' accuracy for optimal assistance in crucial medical contexts.

ACKNOWLEDGMENTS

The authors would like to acknowledge Praveena Gunasekaran, Aparajita Murali, Vaishnavi Girirajan, Grace Anna Abraham, Tejas Krishnan Ranganath, Ashvika AM, and Karthikesh Diraviam for their contributions in data collection.

FOOTNOTES

Author contributions: Ramasubramanian S, Balaji S conceived and designed the study and drafted the manuscript (original and revision); Jeyaraman M and Migliorini F supervised the study and drafted the manuscript (revision); Jeyaraman M, Migliorini F, and Sharma S drafted the manuscript (revision); Ramasubramanian S, Balaji S, Kannan T contributed to drafting (original); Ramasubramanian S, Migliorini F, Balasubramaniam S, Jeyaraman N conducted visualizations and data analysis; All authors have agreed to the final version to be published and agree to be accountable for all aspects of the work.

Institutional review board statement: This study did not require approval from the Institutional Review Board.

Informed consent statement: This study did not require informed consent forms due to the data collection process.

Conflict-of-interest statement: The authors declare no conflicts of interest related to this manuscript.

Data sharing statement: No additional data are available outside of the Supplemental Materials that are published with this article.

STROBE statement: The authors have read the STROBE Statement-checklist of items, and the manuscript was prepared and revised according to the STROBE Statement- checklist of items.

Open-Access: This article is an open-access article that was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution NonCommercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: https://creativecommons.org/Licenses/by-nc/4.0/

Country of origin: India

ORCID number: Swaminathan Ramasubramanian 0000-0001-8845-8427; Sangeetha Balaji 0000-0002-1566-1333; Tejashri Kannan 0009-0009-6400-5488; Naveen Jeyaraman 0000-0002-4362-3326; Shilpa Sharma 0000-0001-8695-8372; Filippo Migliorini 0000-0001-7220-1221; Madhan Jeyaraman 0000-0002-9045-9493.

S-Editor: Liu H
L-Editor: Filipodia
P-Editor: Zhang XD

REFERENCES

1 **OpenAI**. Introducing ChatGPT. Available from: https://openai.com/blog/chatgpt
2 **Sallam M**. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel)* 2023; **11** [PMID: 36981544 DOI: 10.3390/healthcare11060887]
3 **Biswas S**. Role of Chat GPT in Patient Care. 2023
4 **Open AI**. GPT-4 Technical Report. 2023 Preprint. Available from: bioRxiv: 2303.08774 [DOI: 10.48550/arXiv.2303.08774]
5 **Open AI**. GPT-4. Available from: https://openai.com/gpt-4
6 **Google**. Try Bard, an AI experiment by Google. Available from: https://bard.google.com
7 **Bonadio W**. Frequency of emergency medicine resident dosing miscalculations treating pediatric patients. *Am J Emerg Med* 2019; **37**: 1964-1965 [PMID: 30910526 DOI: 10.1016/j.ajem.2019.03.027]
8 **Mulac A**, Hagesaether E, Granas AG. Medication dose calculation errors and other numeracy mishaps in hospitals: Analysis of the nature and enablers of incident reports. *J Adv Nurs* 2022; **78**: 224-238 [PMID: 34632614 DOI: 10.1111/jan.15072]
9 **Wennberg-Capellades L**, Fuster-Linares P, Rodríguez-Higueras E, Fernández-Puebla AG, Llaurado-Serra M. Where do nursing students make mistakes when calculating drug doses? A retrospective study. *BMC Nurs* 2022; **21**: 309 [PMID: 36357884 DOI: 10.1186/s12912-022-01085-9]
10 **Escrivá Gracia J**, Brage Serrano R, Fernández Garrido J. Medication errors and drug knowledge gaps among critical-care nurses: a mixed multi-method study. *BMC Health Serv Res* 2019; **19**: 640 [PMID: 31492188 DOI: 10.1186/s12913-019-4481-7]
11 **Rowe C**, Koren T, Koren G. Errors by paediatric residents in calculating drug doses. *Arch Dis Child* 1998; **79**: 56-58 [PMID: 9771254 DOI: 10.1136/adc.79.1.56]
12 **Loscalzo J**, Fauci A, Kasper D, Hauser S, Longo D, Jameson JL. Harrison's principles of internal medicine. 21st ed. New York: McGraw Hill, 2022
13 **Lim ZW**, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, Chen DZ, Goh JHL, Tan MCJ, Sheng B, Cheng CY, Koh VTC, Tham YC. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* 2023; **95**: 104770 [PMID: 37625267 DOI: 10.1016/j.ebiom.2023.104770]
14 **O'Hagan R**, Kim RH, Abittan BJ, Caldas S, Ungar J, Ungar B. Trends in Accuracy and Appropriateness of Alopecia Areata Information Obtained from a Popular Online Large Language Model, ChatGPT. *Dermatology* 2023; **239**: 952-957 [PMID: 37722370 DOI: 10.1159/000534005]
15 **Johnson D**, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, Chang S, Berkowitz S, Finn A, Jahangir E, Scoville E, Reese T, Friedman D, Bastarache J, van der Heijden Y, Wright J, Carter N, Alexander M, Choe J, Chastain C, Zic J, Horst S, Turker I, Agarwal R, Osmundson E, Idrees K, Kieman C, Padmanabhan C, Bailey C, Schlegel C, Chambless L, Gibson M, Osterman T, Wheless L. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. *Res Sq* 2023 [PMID: 36909565 DOI: 10.21203/rs.3.rs-2566942/v1]
16 **Ramasubramanian S**. Maximizing patient safety with ChatGPT: A novel method for calculating drug dosage. *J Prim Care Spec* 2023; **4**: 150 [DOI: 10.4103/jopcs.jopcs_19_23]
17 **Mayo-Wilson E**, Hutfless S, Li T, Gresham G, Fusco N, Ehmsen J, Heyward J, Vedula S, Lock D, Haythornthwaite J, Payne JL, Cowley T, Tolbert E, Rosman L, Twose C, Stuart EA, Hong H, Doshi P, Suarez-Cuervo C, Singh S, Dickersin K. Integrating multiple data sources (MUDS) for meta-analysis to improve patient-centered outcomes research: a protocol. *Syst Rev* 2015; **4**: 143 [PMID: 26525044 DOI: 10.1186/s13643-015-0134-z]

18 **Hirosawa T**, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. *Int J Environ Res Public Health* 2023; **20** [PMID: 36834073 DOI: 10.3390/ijerph20043378]

19 **Hsu HY**, Hsu KC, Hou SY, Wu CL, Hsieh YW, Cheng YD. Examining Real-World Medication Consultations and Drug-Herb Interactions: ChatGPT Performance Evaluation. *JMIR Med Educ* 2023; **9**: e48433 [PMID: 37561097 DOI: 10.2196/48433]

20 **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. *J Med Internet Res* 2023; **25**: e48659 [PMID: 37606976 DOI: 10.2196/48659]

21 **Al-Ashwal FY**, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools. *Drug Healthc Patient Saf* 2023; **15**: 137-147 [PMID: 37750052 DOI: 10.2147/DHPS.S425858]

22 **Al-Dujaili Z**, Omari S, Pillai J, Al Faraj A. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: A preliminary analysis with clinical pharmacy experts worldwide. *Res Social Adm Pharm* 2023; **19**: 1590-1594 [PMID: 37696742 DOI: 10.1016/j.sapharm.2023.08.012]

23 **Suárez A**, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *Int Endod J* 2024; **57**: 108-113 [PMID: 37814369 DOI: 10.1111/iej.13985]

24 **Abu Hammour K**, Alhamad H, Al-Ashwal FY, Halboup A, Abu Farha R, Abu Hammour A. ChatGPT in pharmacy practice: a cross-sectional exploration of Jordanian pharmacists' perception, practice, and concerns. *J Pharm Policy Pract* 2023; **16**: 115 [PMID: 37789443 DOI: 10.1186/s40545-023-00624-2]

25 **Shin E**, Ramanathan M. Evaluation of prompt engineering strategies for pharmacokinetic data analysis with the ChatGPT large language model. *J Pharmacokinet Pharmacodyn* 2024; **51**: 101-108 [PMID: 37952004 DOI: 10.1007/s10928-023-09892-6]

26 **Jeyaraman M**, Balaji S, Jeyaraman N, Yadav S. Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare. *Cureus* 2023; **15**: e43262 [PMID: 37692617 DOI: 10.7759/cureus.43262]

27 **Huang X**, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: A comparative study of ChatGPT and clinical pharmacists. *Br J Clin Pharmacol* 2024; **90**: 232-238 [PMID: 37626010 DOI: 10.1111/bcp.15896]

28 **Wang X**, Sanders HM, Liu Y, Seang K, Tran BX, Atanasov AG, Qiu Y, Tang S, Car J, Wang YX, Wong TY, Tham YC, Chung KC. ChatGPT: promise and challenges for deployment in low- and middle-income countries. *Lancet Reg Health West Pac* 2023; **41**: 100905 [PMID: 37731897 DOI: 10.1016/j.lanwpc.2023.100905]

29 **Juhi A**, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The Capability of ChatGPT in Predicting and Explaining Common Drug-Drug Interactions. *Cureus* 2023; **15**: e36272 [PMID: 37073184 DOI: 10.7759/cureus.36272]

30 **OpenAI**. Introducing GPTs. Available from: https://openai.com/blog/introducing-gpts

31 **OpenAI**. New models and developer products announced at DevDay. Available from: https://openai.com/blog/new-models-and-developer-products-announced-at-devday

32 **Google**. Introducing Gemini: Google's most capable AI model yet. Available from: https://blog.google/technology/ai/google-gemini-ai/#sundar-note

33 **Google**. A family of highly capable multimodal models. Available from: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf



Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-3991568
E-mail: office@baishideng.com
Help Desk: https://www.f6publishing.com/helpdesk
https://www.wjgnet.com



© 2024 Baishideng Publishing Group Inc. All rights reserved.