medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Exploring the Pitfalls of Large Language Models: Inconsistency and Inaccuracy in Answering Pathology Board Examination-Style Questions

Shunsuke Koga [1]

[1] Department of Pathology and Laboratory Medicine, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania, USA

Running title: Pitfalls of Large Language Models

Corresponding Author:
Shunsuke Koga, MD, PhD
Department of Pathology and Laboratory Medicine
Hospital of the University of Pennsylvania,
3400 Spruce Street, Philadelphia
Pennsylvania, 19104, USA
Tel: (267) 872-3856
Email: shunsuke.koga@pennmedicine.upenn.edu

Abbreviations: AI - Artificial Intelligence; LLMs - Large Language Models

Keywords: artificial intelligence, large language models, pathology, ChatGPT, Google Bard, inconsistency, comparative study, application in medicine.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Abstract

In the rapidly advancing field of artificial intelligence, large language models (LLMs) such as ChatGPT and Google Bard are making significant progress, with applications extending across various fields, including medicine. This study explores their potential utility and pitfalls by assessing the performance of these LLMs in answering 150 multiple-choice questions, encompassing 15 subspecialties in pathology, sourced from the PathologyOutlines.com Question Bank, a resource for pathology examination preparation. Overall, ChatGPT outperformed Google Bard, scoring 122 out of 150, while Google Bard achieved a score of 70. Additionally, we explored the consistency of these LLMs by applying a test-retest approach over a two-week interval. ChatGPT showed a consistency rate of 85%, while Google Bard exhibited a consistency rate of 61%. In-depth analysis of incorrect responses identified potential factual inaccuracies and interpretive errors. While LLMs have potential to enhance medical education and assist clinical decision-making, their current limitations underscore the need for continued development and the critical role of human expertise in the application of such models.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Introduction

Over the past decade, artificial intelligence (AI) has made significant progress, particularly in the development of large language models (LLMs). These models use extensive text data for training, enabling them to generate human-like text, understand context, respond to queries, and facilitate language translation [1].

ChatGPT, including its advanced versions GPT-3.5 and GPT-4, is an LLM developed by OpenAI that has been extensively used in a range of applications, such as writing assistance and programming support. Another notable LLM is Google Bard developed by Google. Importantly, Google Bard has the unique ability to access real-world, current information through Google Search. This capability renders it especially beneficial for answering queries that require the latest data.

LLMs, including ChatGPT, have demonstrated their potential in medical education [2, 3]. ChatGPT not only exhibited promising performance in the United States Medical Licensing Exam [4], but also in equivalent medical exams in Japan and Italy [5, 6]. Current research is examining how LLMs perform in particular medical disciplines. For instance, GPT-4 outperformed other LLMs like GPT-3.5 and Google Bard in a neurosurgery oral boards examination[7]. In another study comparing ChatGPT and Google Bard's responses to lung cancer inquiries, both models showed accuracy, though not without flaws [8].

While AI and machine learning techniques, particularly image analysis [9-11], have been explored in pathology, the evaluation of LLMs has not been extensively investigated [12, 13]. This study seeks to address this gap by assessing the performance of ChatGPT and Google Bard within the field of pathology. The focus is on evaluating and comparing their accuracy and consistency in answering board examination-style questions in pathology.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Methods

Question Selection and Evaluation

The study compared the performance of two LLMs, ChatGPT (GPT-4) and Google Bard, using questions from the PathologyOutlines.com Question Bank (https://www.pathologyoutlines.com/review-questions), a resource for pathology examination preparation.

The question bank had 3365 questions across pathology subspecialties, but for this study, we selected 150 multiple-choice questions, with 10 from each of 15 subspecialties, to ensure a balanced dataset: Autopsy & forensics, Bone, joints & soft tissues, Breast, Dermatopathology, Gastrointestinal & liver, Genitourinary & adrenal, Gynecological, Head & Neck, Hematopathology, Informatics & Digital Pathology, Medical renal, Neuropathology, Stains & CD markers/Immunohistochemistry, Thoracic, and Clinical Pathology. Each question was presented in a single best answer, multiple-choice format. Both LLMs were presented with the same set of questions. No additional context or hints were provided to the models apart from the questions themselves, to simulate real-world application. Questions containing images were excluded from the question bank as ChatGPT is not capable of processing image data.

To evaluate the consistency of both LLMs, a test-retest approach was implemented. The same set of 150 questions was posed to the models on two separate occasions, with a two-week interval. By comparing the answers on both performances, we aimed to assess the consistency in their responses.

Statistical analyses

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Statistical analyses were performed using R version 4.3.1 (R Foundation for Statistical Computing, Vienna, Austria). A Chi-square test was conducted to compare the performance of ChatGPT and Google Bard. Statistical significance was set at $p < 0.05$.

Results

Overall test scores

The performance of the two LLMs, ChatGPT and Google Bard, was evaluated across 15 subspecialties in pathology. Overall, ChatGPT significantly outperformed Google Bard across all subspecialties; ChatGPT achieved a total score of 122 out of 150, compared to Google Bard's score of 70 ($p < .001$). Detailed performance outcomes of each LLM across all subspecialties are presented in **Table 1**.

Assessment of Consistency

In the assessment of consistency of both LLMs, test scores were largely consistent between the first and second sessions. The scores of ChatGPT were 122 and 126 out of 150 in the first and second tests, respectively, while Google Bard scored 70 and 69 in the same tests. Despite the relative stability in test scores, a detailed inspection revealed significant changes; identical answers in both sessions were present in only 85% (127/150) of ChatGPT's responses and a lower 61% (92/150) of Google Bard's (**Table 2**). ChatGPT initially provided 28 incorrect answers. In the retest, it corrected 11 of these but also altered 7 correct answers to incorrect ones. Among the initial errors, ChatGPT repeated 5. Google Bard exhibited a similar but more pronounced pattern. Starting with 80 incorrect responses, it corrected 19 in the retest, made

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

new errors in 20 previously correct answers, and repeated 19 of its original mistakes in the retest.

Evaluation of Incorrect Responses

We identified incorrect answers from both LLMs in our study. One example was a question about Lynch syndrome: "Lynch syndrome usually arises from a germline mutation in a gene coding for a mismatch repair protein. A germline mutation in which of the following genes could also cause Lynch syndrome?" with options A. BRAF, B. CDH1, C. EPCAM, D. MUTYH. The correct answer was C. EPCAM [14]. However, Google Bard incorrectly chose option D. MUTYH and justified its answer by associating MUTYH mutations with Lynch syndrome, which is a factual inaccuracy as MUTYH mutations cause a different type of hereditary colorectal cancer known as MUTYH-associated polyposis [15]. In contrast, ChatGPT correctly selected the answer as C. EPCAM.

Another example was a question about nemaline myopathy: “In which gene are de novo mutations most commonly associated with nemaline myopathy?” with options A. NEB, B. KLHL40, C. TPM3, D. ACTA1, and E. TNNT1. The correct answer was D. ACTA1, but both models selected A. NEB. Specifically, ChatGPT justified its answer by stating that NEB is the most commonly involved gene in cases of nemaline myopathy, while Google Bard indicated that de novo mutations in the NEB gene are the most common cause of this condition. These responses exhibit factual and interpretative inaccuracies, as NEB mutations, while common in nemaline myopathy, are typically inherited in an autosomal recessive manner, not de novo [16].

Discussion

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Our study further underscores the strengths and weaknesses of LLMs in medicine. While ChatGPT consistently surpassed Google Bard in accuracy and consistency, neither model answered all questions correctly, suggesting knowledge or comprehension gaps. Additionally, retesting after two weeks revealed inconsistencies in both LLMs' responses to the same questions, highlighting potential reliability issues [17].

In assessing LLMs' comprehension of medical queries, our study identified two error types. Firstly, Google Bard displayed factual inaccuracies, incorrectly linking MUTYH to Lynch syndrome despite vast data access. Secondly, ChatGPT exhibited interpretation errors. While answering a question about "de novo" mutations in nemaline myopathy, it correctly identified NEB as a common cause but overlooked the specific "de novo" context, highlighting LLMs' potential for nuanced misunderstandings.

Another important consideration in the application of LLMs in medical fields is their consistency or reliability, defined as the models' ability to provide the same answer to identical prompts when asked on different occasions. Our assessment of consistency revealed a suboptimal consistency rate for both LLMs (i.e., 85% in ChatGPT and 61% in Google Bard), which are consistent with the results of another study that evaluated ChatGPT's responses to surgical case questions [18]. Such inconsistencies underline the current limitations of LLMs and highlights the necessity for further development and refinement to improve their consistency for effective use in the medical field.

There are some limitations of our study. Firstly, there was no direct comparison with human performance. While our results shed light on the capabilities of LLMs in answering complex medical questions, understanding how their performance compares directly to medical students or professionals remains crucial. Additionally, our focus was largely on pathology questions in English language. To generalize our findings, future studies should encompass

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

different medical specialties and languages. Lastly, the challenge of incorporating images into our evaluation also presents a significant limitation.

In conclusion, our study indicates that LLMs have the potential to assist in clinical decision-making in the future. Both models demonstrated inconsistencies and inaccuracies, emphasizing the need for their further development and rigorous validation. While the potential of these AI models is promising, human oversight and expertise remain crucial in the medical field.

Acknowledgement

This manuscript was edited and proofread by ChatGPT (GPT-4, OpenAI), and the author verified the final content.

Disclosure Statement

None.

Author Contributions

Conception and design of the study: SK

Acquisition and analysis of data: SK

Drafting the manuscript: SK

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

References

1 Chen M, Tworek J, Jun H*, et al.* Evaluating Large Language Models Trained on Code. 2021; arXiv:2107.03374.

2 Koga S. The Potential of ChatGPT in Medical Education: Focusing on USMLE Preparation. *Ann Biomed Eng*. 2023.

3 Feng S, Shen Y. ChatGPT and the Future of Medical Education. *Acad Med*. 2023; **98**: 867-68.

4 Kung TH, Cheatham M, Medenilla A*, et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023; **2**: e0000198.

5 Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. *JMIR Med Educ*. 2023; **9**: e48002.

6 Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How Does ChatGPT Perform on the Italian Residency Admission National Exam Compared to 15,869 Medical Graduates? *Ann Biomed Eng*. 2023.

7 Ali R, Tang OY, Connolly ID*, et al.* Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery*. 2023.

8 Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology*. 2023; **307**: e230922.

9 Komura D, Ishikawa S. Machine learning approaches for pathologic diagnosis. *Virchows Arch*. 2019; **475**: 131-38.

10 Koga S, Ikeda A, Dickson DW. Deep learning-based model for diagnosing Alzheimer's disease and tauopathies. *Neuropathol Appl Neurobiol*. 2022; **48**: e12759.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

11 Kim M, Sekiya H, Yao G*, et al.* Diagnosis of Alzheimer Disease and Tauopathies on Whole-Slide Histopathology Images Using a Weakly Supervised Deep Learning Algorithm. *Lab Invest*. 2023; **103**: 100127.

12 Schukow C, Smith SC, Landgrebe E*, et al.* Application of ChatGPT in Routine Diagnostic Pathology: Promises, Pitfalls, and Potential Future Directions. *Adv Anat Pathol*. 2023.

13 Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathol*. 2023: e13207.

14 Tutlewska K, Lubinski J, Kurzawski G. Germline deletions in the EPCAM gene as a cause of Lynch syndrome - literature review. *Hered Cancer Clin Pract*. 2013; **11**: 9.

15 Nielsen M, Morreau H, Vasen HF, Hes FJ. MUTYH-associated polyposis (MAP). *Crit Rev Oncol Hematol*. 2011; **79**: 1-16.

16 Laitila J, Wallgren-Pettersson C. Recent advances in nemaline myopathy. *Neuromuscul Disord*. 2021; **31**: 955-67.

17 Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? 2023; arXiv:2307.09009.

18 Beaulieu-Jones BR, Shah S, Berrigan MT, Marwaha JS, Lai S-L, Brat GA. Evaluating Capabilities of Large Language Models: Performance of GPT4 on Surgical Knowledge Assessments. *medRxiv*. 2023: 2023.07.16.23292743.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 1: Performance scores of ChatGPT and Google Bard across pathology subspecialties

Subspeciality	**ChatGPT**	**Google Bard**
Autopsy & Forensics	9	6
Bone, Joints & Soft Tissues	7	3
Breast	7	3
Dermatopathology	8	5
Gastrointestinal & Liver	9	6
Genitourinary & Adrenal	6	4
Gynecological	9	7
Head & Neck	9	3
Hematopathology	9	5
Informatics & Digital Pathology	9	7
Medical Renal	9	3
Neuropathology	8	5
Stains & CD markers/Immunohistochemistry	8	3
Thoracic	8	5
Clinical Pathology	7	5
Total	**122**	**70**

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293401; this version posted August 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Table 2: Consistency of Model Responses

Outcome	ChatGPT	Google Bard
No change in response	127 (85%)	92 (61%)
Correct to incorrect response	7 (5%)	20 (13%)
Incorrect to another incorrect response	5 (3%)	19 (13%)
Incorrect to correct response	11 (7%)	19 (13%)