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

Shunsuke Koga, MD, PhD,[1]

1) Department of Pathology and Laboratory Medicine, Hospital of the University of Pennsylvania, Philadelphia, PA 19104

Correspondence to:

Shunsuke Koga, MD, PhD

Hospital of the University of Pennsylvania

3400 Spruce Street

Philadelphia, PA 19104

Phone: 904-953-7137

Email: shunsuke.koga@pennmedicine.upenn.edu

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 sourced from the PathologyOutlines.com Question Bank, a well-established resource for pathology examination preparation.

The assessment, encompassing 15 subspecialties in pathology, evaluated the accuracy and consistency of responses by these LLMs. Overall, ChatGPT outperformed Google Bard, scoring 122 out of 150, while Google Bard achieved a score of 70. In addition to accuracy, 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 lower consistency rate of 61%. In-depth analysis of incorrect responses identified potential factual inaccuracies and interpretive errors, underscoring the need for ongoing model refinement and human oversight.

In conclusion, 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.

Keywords: Artificial Intelligence, Large Language Models, Pathology, ChatGPT, Google Bard, Inconsistency

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.

The application of these LLMs in medicine has gained increasing attention recently. A study conducted by Kung et al. demonstrated that without specialized training, ChatGPT performed at or near the passing threshold on all three parts of the United States Medical Licensing Exam [2]. The research scope was not confined to English language exams, with several studies extending to non-English evaluations. For instance, GPT-4 was assessed using the Japanese Medical Licensing Examination, where it exhibited an acceptable performance, despite possible constraints due to a relative paucity of high-quality Japanese medical data [3]. An Italian study assessed GPT-4 using the Italian Residency Admission National Exam, producing promising results which scored the top 98.8th percentile among 15,869 medical graduates [4].

The scope of research is now focusing on comparing LLMs in specific medical fields. One study evaluated the performance of LLMs, including GPT-3.5, GPT-4, and Google Bard, using a question set specifically designed for a neurosurgery oral boards examination, where GPT-4 notably achieved the highest score at 82.6% [5]. Another investigation compared the accuracy and consistency of responses provided by ChatGPT and Google Bard to non-expert questions about lung cancer. The study found that, while both models displayed considerable accuracy, their responses were not uniformly flawless [6].

Building on previous research, our study focuses on the assessment of ChatGPT and Google Bard within the field of pathology, an area that warrants further exploration [7,8]. We aimed to evaluate and compare their accuracy and consistency of these models in answering board examination-style questions in pathology.

Methods

Question Selection and Evaluation

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

While the question bank consisted of 3365 questions across subspecialties of pathology, a total of 150 multiple-choice questions were selected for this study to maintain a manageable and balanced dataset. These comprised 10 questions from

each of the following 15 subspecialties: 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 (IHC), 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.

Each LLM's responses were compared to the correct answers as defined in the question bank. Correct answers were given a score of one, and incorrect answers were assigned a score of zero. The total score for each model was calculated by summing the individual scores for all 150 questions.

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

Statistical analyses were performed using R version 4.3.1 (R Foundation for Statistical Computing, Vienna, Austria) and the Rcmdr package for R. 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**.

ChatGPT achieved its highest score (9/10) in seven subspecialties: Autopsy & Forensics, Gastrointestinal & Liver, Gynecological, Head & Neck, Hematopathology, Informatics & Digital Pathology, and Medical Renal. Its lowest score was 6/10, achieved in the Genitourinary & Adrenal subspecialty. In contrast, Google Bard scored highest (7/10) in the Gynecological and Informatics & Digital Pathology subspecialties, and scored lowest (3/10) in the Bone, Joints & Soft Tissues, Breast, Head & Neck, Medical Renal, and Stains & CD Markers/IHC subspecialties.

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 out of 150 in the same tests. Despite the relative stability in test scores, a detailed inspection revealed significant changes; stable responses — 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**).

Furthermore, the pattern of response alterations was noteworthy. Initially, ChatGPT provided 28 incorrect answers. In the retest, it rectified 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 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. The question asked: "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 [9]. 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 [10]. In contrast, ChatGPT correctly selected the answer as C. EPCAM.

Another example was a question about nemaline myopathy. The question was: “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 [11].

Discussion

The results of our study provide additional insights into the capabilities and current limitations of LLMs like ChatGPT and Google Bard in medical settings. As with previous study [5], ChatGPT consistently outperformed Google Bard in our evaluation, which included measures of correct response rates and consistency. However, it is important to note that neither LLM was able to provide correct answers to all questions, indicating gaps in their medical knowledge or comprehension. Furthermore, when we conducted a second test two weeks after the initial test, we observed inconsistencies in the responses of both LLMs to identical questions. This points towards inherent limitations in their ability to reliably reproduce correct information.

To further understand the potential limitations and sources of errors for LLMs in comprehending medical queries, we focused on two specific examples from our study. The first example underscores the potential for factual inaccuracies [12]. A question about a germline mutation that could also cause Lynch syndrome was inaccurately answered by Google Bard, which chose MUTYH as the correct response. Despite Google Bard's access to an extensive pool of online data, it mistakenly linked MUTYH to Lynch syndrome, demonstrating that factual errors can occur even with the wealth of up-to-date information available.

The second example highlights the more complex issue of interpretation errors, particularly when it comes to nuanced and context-specific queries. The question was about the gene most commonly associated with "de novo" mutations in nemaline myopathy. In genetics, a "de novo" mutation is one that occurs spontaneously, for the first time, rather than being inherited from parents. The correct answer to this question is the *ACTA1* gene; however, ChatGPT answered *NEB*, which, while not incorrect in a broader context, failed to address the specificity of the question. NEB mutations are indeed the most common cause of nemaline myopathy, but these are typically inherited in an "autosomal recessive" manner, requiring a copy of the mutation from both parents for the disease to occur. In this scenario, it seems that ChatGPT may have overlooked the "de novo" aspect of the question, focusing instead on the broader, more common cause of the condition (NEB) instead of the more specific "de novo" mutation (ACTA1). This example highlights the potential for misinterpretation by LLMs when confronted with complex, nuanced queries. It serves as a reminder of the need for continual refinement of these models, and the importance of maintaining human oversight in their

application, particularly in critical fields like medicine [12]. This underlines the challenges of AI in understanding and responding accurately to the intricacies and complexities inherent in medical language and context.

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 less than ideal consistency rate for both LLMs. ChatGPT exhibited a consistency rate of 85%, and Google Bard demonstrated a lower consistency rate of only 61%. These findings are consistent with the results of another study that evaluated ChatGPT's responses to surgical case questions and reported an inconsistency rate of 36.4% [13]. Such inconsistencies underline the current limitations of LLMs 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 different medical specialties and languages. Lastly, the challenge of incorporating images into our evaluation also presents a significant limitation. Although Google Bard possesses the ability to process image-based information, we found its accuracy on image-based questions to be notably low, which

was consistent with its overall performance in our study, leading us to exclude this feature from our evaluation.

In conclusion, our study indicates that LLMs have the potential to enhance medical education and possibly 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.

References

1 Chen M, Tworek J, Jun H, Yuan Q, Ponde de Oliveira Pinto H, Kaplan J *et al.* Evaluating Large Language Models Trained on Code. arXiv:2107.03374 (2021). <https://ui.adsabs.harvard.edu/abs/2021arXiv210703374C>.

2 Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C *et al.* 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 (2023).

3 Takagi S, Watari T, Erabi A, and Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. *JMIR Med Educ* **9**, e48002, doi:10.2196/48002 (2023).

4 Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, and Egro FM. How Does ChatGPT Perform on the Italian Residency Admission National Exam

Compared to 15,869 Medical Graduates? *Ann Biomed Eng*, doi:10.1007/s10439-023-03318-7 (2023).

5 Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL *et al.* Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery*, doi:10.1227/neu.0000000000002551 (2023).

6 Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, and Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology* **307**, e230922, doi:10.1148/radiol.230922 (2023).

7 Schukow C, Smith SC, Landgrebe E, Parasuraman S, Folaranmi OO, Paner GP *et al.* Application of ChatGPT in Routine Diagnostic Pathology: Promises, Pitfalls, and Potential Future Directions. *Adv Anat Pathol*, doi:10.1097/PAP.0000000000000406 (2023).

8 Koga S. Evaluating the Performance of Large Language Models: ChatGPT and Google Bard in Generating Differential Diagnoses in Clinicopathological Conferences of Neurodegenerative Disorders". *Brain Pathol*, doi:10.1111/bpa.13207 (2023). *In Press*

9 Tutlewska K, Lubinski J, and Kurzawski G. Germline deletions in the EPCAM gene as a cause of Lynch syndrome - literature review. *Hered Cancer Clin Pract* **11**, 9, doi:10.1186/1897-4287-11-9 (2013).

10 Nielsen M, Morreau H, Vasen HF, and Hes FJ. MUTYH-associated polyposis (MAP). *Crit Rev Oncol Hematol* **79**, 1-16, doi:10.1016/j.critrevonc.2010.05.011 (2011).

11 Laitila J andWallgren-Pettersson C. Recent advances in nemaline myopathy. *Neuromuscul Disord* **31**, 955-967, doi:10.1016/j.nmd.2021.07.012 (2021).

12 Borji A. A Categorical Archive of ChatGPT Failures. arXiv:2302.03494 (2023). <https://ui.adsabs.harvard.edu/abs/2023arXiv230203494B>.

13 Beaulieu-Jones BR, Shah S, Berrigan MT, Marwaha JS, Lai S-L, and Brat GA. Evaluating Capabilities of Large Language Models: Performance of GPT4 on Surgical Knowledge Assessments. *medRxiv*, 2023.2007.2016.23292743, doi:10.1101/2023.07.16.23292743 (2023).

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**

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%)