Large Language Models in Pathology: A Comparative Study on Multiple Choice Question Performance with Pathology Trainees

Wei Du, MD, PhD,[1] Jaryse Carol Harris, MD,[1] Alessandro Brunetti, MD,[1] Olivia Leung, MD,[1] Xingchen Li, MD,[2] Selemon Walle, MD,[1] Qing Yu, MD, PhD,[2] Xiao Zhou, MD, PhD,[2] Fang Bian, MD, PhD,[2] Kajanna McKenzie, MD,[1] Manita Kanathanavanich, MD,[1] Xueting Jin, MD,[1] Farah El-Sharkawy, MD,[1] and Shunsuke Koga, MD, PhD[1]

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

Running title: Large Language Models in Pathology

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

Word: 2792 References: 22

Figures: 4 Tables: 4

Keywords: artificial intelligence, large language models, pathology, ChatGPT, Bard, inconsistency, comparative study, resident, medical education.

Abstract

Aims: Large language models (LLMs), such as ChatGPT and Bard, have shown potential in various medical applications. This study aims to evaluate the performance of LLMs, specifically ChatGPT and Bard, in pathology by comparing their performance with that of pathology residents and fellows, and to assess the consistency of their responses.

Methods: We selected 150 multiple-choice questions covering 15 subspecialties, excluding those with images. Both ChatGPT and Bard were tested on these questions three times, and their responses were compared with those of 14 pathology trainees from two hospitals. Questions were categorized into easy, intermediate, and difficult based on trainee performance. Consistency and variability in LLM responses were analyzed across three evaluation sessions.

Results: ChatGPT significantly outperformed Bard and trainees, achieving an average total score of 82.2% compared to Bard's 49.5% and trainees' 50.7%. ChatGPT's performance was notably stronger in difficult questions (61.8%-70.6%) compared to Bard (29.4%-32.4%) and trainees (5.9%-44.1%). For easy questions, ChatGPT (88.9%-94.4%) and trainees (75.0%-100.0%) showed similar high scores. Consistency analysis revealed that ChatGPT showed a high consistency rate of 85%-80% across three tests, whereas Bard exhibited greater variability with consistency rates of 61%-54%.

Conclusion: ChatGPT consistently outperformed Bard and trainees, especially on difficult questions. While LLMs show significant potential in pathology education and practice, ongoing development and human oversight are essential for reliable clinical application.

Introduction

Over the past decade, artificial intelligence (AI) has made significant progress, particularly in the development of large language models (LLMs). These models, trained on extensive text data, can generate human-like text, understand context, respond to queries, and facilitate language translation.[1] Notable examples include ChatGPT by OpenAI, which utilizes the Generative Pre-trained Transformer (GPT)-3.5 and GPT-4 models, and Bard by Google, which is based on the Pathways Language Model (PaLM) 2. Both applications have been widely used for various purposes, such as writing assistance and complex question-answering tasks. These applications are also quite accessible and can be readily used by people without extensive knowledge in AI or computer science.

LLMs have been evaluated on various tasks in the medical field, demonstrating impressive capabilities.[2-4] For instance, ChatGPT has shown promising performance by achieving passing scores on the United States Medical Licensing Exam (USMLE).[5] Additionally, LLMs have been tested in specialized medical board examinations across different disciplines, achieving levels comparable to those of medical professionals.[6] Moreover, LLMs have proven capable of generating differential diagnoses based on patient chief complaints and medical histories, showcasing their potential in clinical decision-making support. Singhal et al. further demonstrated that Flan-PaLM and Med-PaLM, advanced variants of the PaLM, achieved state-of-the-art performance on multiple medical question-answering benchmarks, significantly surpassing previous models.[7]

AI and machine learning, particularly in image analysis, have been extensively explored in pathology and have shown promise in tasks such as automated image analysis and diagnostic support [8-10]; however, the evaluation and application of LLMs in pathology remain limited.[11 12] In one study, ChatGPT has been used to generate multiple-choice questions for

pathology board examinations, although expert review and refinement were necessary. Geetha et al. assessed the ability of ChatGPT to answer pathology MCQs and found that its performance was lower than the average peer performance, achieving an accuracy of 56.98% compared to 62.81% for residents.[13] This suggests that while ChatGPT has potential in medical education, it may not yet surpass human trainees in knowledge. In contrast, our previous research demonstrated high performance by ChatGPT in answering board examination-style questions in pathology, but we did not compare this with human test takers, such as residents.[14] To address this gap, the present study expands on our previous work by directly comparing the performance of these LLMs with 14 pathology trainees from two hospitals in the United States. This comparative analysis aims to provide insights into the relative strengths and limitations of LLMs in pathology education and practice.

Methods

Question Selection and Evaluation

The study compared the performance of two LLMs, ChatGPT (GPT-4) and Bard, as well as pathology trainees, including residents and fellows, using multiple-choice questions from the PathologyOutlines.com Question Bank (https://www.pathologyoutlines.com/review-questions), a resource for pathology examination preparation. The question bank contained 3365 multiple-choice questions across pathology subspecialties. For this study, 150 questions were selected, with 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, Thoracic, and Clinical Pathology. Questions containing images were excluded as ChatGPT was not capable of processing image

data at the time of the study. Each question was presented in a single best answer, multiple-choice format, and both LLMs were given the same set of questions without additional context or hints.

Comparison with Pathology Trainees

The performance of ChatGPT and Bard was compared with that of 14 pathology trainees on the same set of questions. The pathology trainees included 10 from the Hospital of the University of Pennsylvania— composed of five PGY-1 residents, two PGY-3 residents, and three fellows— and 4 from Pennsylvania Hospital — composed of three PGY-1 residents and one PGY-2 resident—. In this study, Junior trainees are defined as PGY-1 residents (N = 8) and Senior trainees as PGY-2, PGY-3, and fellows (N = 6). All questions were listed in a single document file. Participants were asked to complete all 150 questions within 2 hours without using any external resources.

Assessment of Question Difficulty

To evaluate the performance of LLMs based on question difficulty, we categorized the questions into three levels: easy, intermediate, and difficult. The difficulty level was defined based on the number of correct answers provided by 14 pathology trainees. Questions correctly answered by 10 to 14 trainees were categorized as easy, those answered correctly by 5 to 9 trainees as intermediate, and those answered correctly by 0 to 4 trainees as difficult.

Consistency Evaluation

To assess the consistency of LLM’s performance, each model was presented with the same set of 150 questions three times: the initial test, a follow-up two weeks later, and a final test 32 weeks after the second. We compared the changes in total scores and the breakdown of response changes among the three tests. Heatmaps were generated using Python, specifically

the *matplotlib* and *seaborn* packages, to visually represent the consistency and variability in responses across the three evaluations.

Statistical analyses

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

Results

Overall test scores

The performance of ChatGPT, Bard, and trainees was evaluated across 15 subspecialties in pathology. Overall, ChatGPT significantly outperformed Bard in all subspecialties, achieving an average total score of 123.3 ± 2.3 (mean ± standard deviation; 82.2%) across three tests. In comparison, the average score of Bard was 74.3 ± 8.4 (49.5%) across three tests. The average score of the 14 trainees was 76.1 ± 14.4 (50.7%), which was significantly lower than that of ChatGPT ($p<0.001$) (**Figure 1**). Among the trainees, Junior trainees had an average score of 67.6 ± 8.9 (45.1%), while Senior trainees scored higher with an average of 87.5 ± 15.7 (58.3%). Detailed performance outcomes across all subspecialties are presented in **Table 1**.

Assessment of Consistency

In the assessment of consistency of both LLMs, test scores were largely consistent among the three sessions. The scores of ChatGPT were 122, 126 and 122 out of 150 in the first, second, and third tests, respectively. Despite the relative stability in test scores, a closer examination

revealed significant changes; identical answers between the first and second sessions were present in 85% (127/150) of ChatGPT's responses. This consistency slightly decreased to 82% (123/150) between the first and third sessions and to 80% (120/150) between the second and third sessions (**Table 2**). Changes in ChatGPT's responses included 7% to 10% shifting from incorrect to correct answers, 5% to 7% shifting from correct to incorrect, and 3% to 4% shifting from one incorrect answer to another (**Table 2**).

Bard exhibited a more pronounced variability in its responses (**Table 3**). The total scores for the three tests were 70, 69, and 84, respectively. Identical answers between the first and second sessions were present in 61% (92/150) of responses, dropping to 54% (81/150) between the first and third sessions, and 55% (82/150) between the second and third sessions. Changes in Bard's responses included 11% to 13% shifting from correct to incorrect answers, 11% to 21% shifting from incorrect to correct answers, and 13% to 14% shifting from one incorrect answer to another. These response variations of LLMs across three evaluations were visualized as a heatmap as shown in **Figure 2**.

Assessment of the difficulty level

To evaluate whether the difficulty level for LLMs and trainees is comparable, we first categorized questions based on the number of trainees who answered each question correctly. The distribution of these questions is visualized in the histogram (**Figure 3**), showing a range from 3 questions that no trainees answered correctly to 8 questions that all trainees answered correctly.

Based on the trainees' responses, we categorized 36 questions as easy, 80 as intermediate, and 34 as difficult. For the easy questions, ChatGPT scored 88.9%, 94.4%, and 88.9% across three tests, while Bard scored 63.9%, 69.4%, and 86.1%. The trainees averaged 84.6%, with scores ranging from 75.0% to 100.0%. In the intermediate category, ChatGPT

scored 86.3%, 90.0%, and 80.0%, while Bard scored 46.3%, 41.3%, and 53.8%. The trainees had an average score of 50.7%, ranging from 35.0% to 73.8%. For difficult questions, ChatGPT scored 61.8%, 58.8%, and 70.6%, compared to Bard's scores of 29.4%, 32.4%, and 29.4%. The trainees averaged 20.0%, with scores ranging from 5.9% to 44.1%.

Examples from these difficulty levels are presented in **Table 4**. An example of the easiest question is from the autopsy category, asking about the most likely cause of death in a 25-year-old male patient. Clues such as a low respiratory rate, pinpoint pupils, and multiple needle tracks on the arm indicate opioid overdose. An example of an intermediate question is from the bone, joints & soft tissues category, which asks about the immunohistochemical stains that help distinguish retroperitoneal myolipoma from a well-differentiated liposarcoma invading smooth muscle. The correct answer is CDK4, as it is overexpressed in well-differentiated liposarcomas but not in benign myolipomas, making it a key distinguishing marker. Seven out of 14 trainees correctly answered this question. On the other hand, an example of a hardest question is from the Clinical Pathology category. The question describes a gram-negative rod isolated from the sputum of a homeless, alcoholic patient with a chronic cough. The isolate forms a mucoid colony that turns pink on MacConkey agar and produces a blue spot when treated with indole. The correct organism is Klebsiella oxytoca, identified by its indole positivity, which distinguishes it from the more commonly associated Klebsiella pneumoniae. ChatGPT consistently selected Klebsiella pneumoniae in all three tests. Bard selected Klebsiella oxytoca in the first test and Klebsiella pneumoniae in the second and third tests. All 14 pathology trainees chose Klebsiella pneumoniae.

Overall, ChatGPT consistently outperformed Bard and closely matched the higher-performing trainees, especially on easy and intermediate questions. Bard exhibited more variability and lower accuracy across all difficulty levels, while trainees showed a wide range of performance, particularly with difficult questions.

Discussion

The present study expands our prior work by directly comparing the performance of LLMs on pathology questions with that of pathology trainees at different stages of their training. The results demonstrated that ChatGPT outperformed the trainees, while Bard's scores were comparable to those of the trainees. The superiority of ChatGPT was more pronounced on difficult questions. Despite their strengths, LLMs did not achieve perfect accuracy and consistency. The responses of LLMs were unstable; even with the same prompt, the responses varied, leading to changes in answers and scores. Our findings illustrate the necessity of improved understanding and awareness of these pitfalls, while also highlighting the importance of careful supervision when using LLMs.[15]

Recent studies demonstrate that LLMs are increasingly performing better than residents and physicians in medical examinations across various specialties. Katz et al. systematically compared GPT-4 with 849 physicians across five core medical disciplines, finding that GPT-4 ranked higher than most physicians in psychiatry (74.7% median percentile) and performed similarly to the median physician in general surgery and internal medicine (44.4% and 56.6% median percentiles, respectively). Although its performance was lower in pediatrics and obstetrics/gynecology, GPT-4 still exceeded a considerable fraction of practicing physicians in these areas, ultimately passing the board residency exam in four out of five specialties. Another study evaluated GPT-4 against Family Medicine residents on a multiple-choice medical knowledge test.[16] GPT-4 significantly outperformed the residents, achieving a score of 82.4%, higher than the top-performing resident. GPT-4 correctly answered 89 out of 108 questions and provided rationales for its responses in 86.1% of cases. The average resident performance was 56.9%. Wang et al. investigated the performance of GPT-3.5 and GPT-4 using novel pathology-

specific questions written by an international group of pathologists.[17] GPT-4 scored higher than both GPT-3.5 and a practicing pathologist on 12 of 15 questions, demonstrating its capacity to meet or exceed trained pathologist performance. Collectively, these studies highlight the impressive capabilities of GPT-4 in medical education and suggest its potential to enhance clinical training and decision-making.

Consistent with these findings, our study demonstrated that ChatGPT significantly outperformed both Bard and pathology trainees in answering pathology questions. This supports the notion that LLMs can achieve performance levels comparable to or exceeding those of human medical trainees. However, our findings contrast with those of Geetha et al., who found that ChatGPT's performance on pathology questions was lower than that of pathology residents (56.98% vs. 62.81%).[13] This discrepancy could be attributed to differences in study design, question sets, and the specific prompts used in the studies. Geetha's study was conducted before June 2023, while our study started in July 2023; therefore, ChatGPT may have been updated and improved during this timeframe, potentially enhancing its performance in our study. Additionally, in our study, 8 out of 14 trainees were PGY-1 residents with less than one year of training, which may have contributed to the lower scores observed in our trainee group. The varied outcomes underscore the need for further research to better understand the capabilities and limitations of LLMs in different medical contexts.

One intriguing aspect of our findings is the trend that both LLMs and human trainees showed lower accuracy on questions categorized as difficult based on trainee performance, which was consistent with the findings of Geetha’s study.[13] This similarity suggests that the factors contributing to question difficulty may affect both humans and LLMs, though possibly for different reasons. For trainees, difficult questions often require higher-order thinking, integration of knowledge across different domains, and application of nuanced clinical judgment. For LLMs, these questions likely present challenges because they may involve complex medical concepts,

rare conditions, or ambiguous language that the model has encountered less frequently during training. LLMs are trained on vast amounts of text data, but this data may not always include the specific, detailed medical knowledge needed to answer certain high-level pathology questions accurately. Moreover, the training process of LLMs, which relies on pattern recognition and statistical associations, may not fully capture the intricate reasoning and contextual understanding required for the most challenging questions. Therefore, while the difficulty levels appear to align for both humans and AI, the underlying reasons for difficulty may differ.

A key consideration in the use of LLMs in medical applications is their reliability, which refers to the models' ability to provide consistent answers to identical prompts when tested multiple times. Our study found notable inconsistencies in the responses across three evaluation sessions, where Bard exhibited greater variability than ChatGPT. Such changes can be attributed to several factors, including the inherent randomness in LLMs' output generation processes and potential updates or changes in the underlying models over time. These findings align with previous research that highlighted similar inconsistencies in LLMs' responses.[18] The variability in answers highlights the limitations of LLMs, underscoring the need for ongoing development to improve reliability. Consistent performance is crucial for their effective use in medical education and clinical decision-making; therefore, human oversight remains essential to verify and interpret LLM outputs accurately.

While the present study successfully addressed the limitation of our prior research by directly comparing the performance of LLMs with that of pathology trainees, several limitations remain. First, similar to the previous studies,[13 14] this research utilized multiple-choice questions without images because ChatGPT was unable to process uploaded images at the time of the study's initiation. This capability was only introduced in November 2023 with the release of GPT-4Vision.[19] Although integrating image-based questions could provide a more comprehensive assessment of LLMs in pathology, recent literature indicates that the accuracy

of medical image analysis by these models remains suboptimal.[20-22] Thus, incorporating image-based questions might not yet yield reliable comparisons at this point and could require further technological advancements and validations. Second, the relatively small number of trainee participants is a notable limitation. To mitigate this, we recruited trainees from two hospitals: the Hospital of the University of Pennsylvania and Pennsylvania Hospital. Despite this effort, the sample size of 14 trainees may still be insufficient to generalize the findings across the broader population of pathology trainees. Future studies should aim to include a larger and more diverse cohort of trainees from multiple institutions to enhance the robustness and generalizability of the results. Finally, this study exclusively analyzed pathology questions in the English language. Pathology trainees and professionals globally operate in various languages, and the performance of LLMs may vary based on linguistic and cultural contexts.

In conclusion, our study demonstrates the capability of LLMs in answering a wide range of questions in pathology, outperforming the residents and fellows. Although these results support their potential in medical applications including medical education and clinical decision-making, both models showed inconsistencies and inaccuracies, emphasizing the need for 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.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930-1940. doi: 10.1038/s41591-023-02448-8.
2. Feng S, Shen Y. ChatGPT and the Future of Medical Education. Acad Med 2023;**98**(8):867-68 doi: 10.1097/ACM.0000000000005242.
3. Koga S. The Potential of ChatGPT in Medical Education: Focusing on USMLE Preparation. Ann Biomed Eng 2023;**51**(10):2123-24 doi: 10.1007/s10439-023-03253-7.
4. Koga S. The Integration of Large Language Models Such as ChatGPT in Scientific Writing: Harnessing Potential and Addressing Pitfalls. Korean J Radiol 2023;**24**(9):924-25 doi: 10.3348/kjr.2023.0738.
5. 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**(2):e0000198 doi: 10.1371/journal.pdig.0000198.

6. 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 doi: 10.1227/neu.0000000000002551.

7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;**620**(7972):172-80 doi: 10.1038/s41586-023-06291-2.

8. Komura D, Ishikawa S. Machine learning approaches for pathologic diagnosis. Virchows Arch 2019;**475**(2):131-38 doi: 10.1007/s00428-019-02594-w.

9. Koga S, Ikeda A, Dickson DW. Deep learning-based model for diagnosing Alzheimer's disease and tauopathies. Neuropathol Appl Neurobiol 2022;**48**(1):e12759 doi: 10.1111/nan.12759.

10. 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**(6):100127 doi: 10.1016/j.labinv.2023.100127.

11. 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 doi: 10.1097/PAP.0000000000000406.

12. 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 doi: 10.1111/bpa.13207.

13. Geetha SD, Khan A, Khan A, Kannadath BS, Vitkovski T. Evaluation of ChatGPT pathology knowledge using board-style questions. Am J Clin Pathol 2024;**161**(4):393-98 doi: 10.1093/ajcp/aqad158.

14. Koga S. Exploring the pitfalls of large language models: Inconsistency and inaccuracy in answering pathology board examination-style questions. Pathol Int 2023;**73**(12):618-20 doi: 10.1111/pin.13382.

15. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? 2023. https://ui.adsabs.harvard.edu/abs/2023arXiv230709009C (accessed July 01, 2024).

16. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of Resident and AI Chatbot Performance on the University of Toronto Family Medicine Residency Progress Test: Comparative Study. JMIR Med Educ 2023;**9**:e50514 doi: 10.2196/50514.

17. Wang AY, Lin S, Tran C, et al. Assessment of Pathology Domain-Specific Knowledge of ChatGPT and Comparison to Human Performance. Arch Pathol Lab Med 2024 doi: 10.5858/arpa.2023-0296-OA.

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 doi: 10.1101/2023.07.16.23292743.

19. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. 2023. https://ui.adsabs.harvard.edu/abs/2023arXiv230308774O (accessed July 01, 2024).

20. Koga S, Du W. From text to image: challenges in integrating vision into ChatGPT for medical image interpretation. Neural Regeneration Research 2024 doi: 10.4103/NRR.NRR-D-24-00165.

21. Laohawetwanit T, Namboonlue C, Apornvirat S. Accuracy of GPT-4 in histopathological image detection and classification of colorectal adenomas. J Clin Pathol 2024 doi: 10.1136/jcp-2023-209304.

22. Miao J, Thongprayoon C, Cheungpasitporn W, Cornell LD. Performance of GPT-4 Vision on kidney pathology exam questions. Am J Clin Pathol 2024 doi: 10.1093/ajcp/aqae030.

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

Subspecialty	**ChatGPT**	**Bard**	**Trainees**
Autopsy & Forensics	8.0 ± 1.0	5.7 ± 0.6	6.1 ± 1.5
Bone, Joints & Soft Tissues	7.3 ± 0.6	2.7 ± 1.5	4.6 ± 2.0
Breast	7.3 ± 0.6	3.0 ± 0.0	5.6 ± 1.9
Dermatopathology	8.3 ± 0.6	5.7 ± 1.2	4.3 ± 1.8
Gastrointestinal & Liver	9.0 ± 0.0	5.3 ± 1.2	6.1 ± 1.5
Genitourinary & Adrenal	7.3 ± 1.5	5.0 ± 1.0	5.0 ± 1.5
Gynecological	9.0 ± 0.0	7.0 ± 0.0	6.3 ± 2.1
Head & Neck	9.0 ± 1.0	4.3 ± 1.5	4.4 ± 2.2
Hematopathology	9.3 ± 0.6	5.0 ± 0.0	5.3 ± 1.6
Informatics & Digital Pathology	9.3 ± 0.6	8.3 ± 1.5	6.6 ± 1.5
Medical Renal	9.7 ± 0.6	6.3 ± 3.1	5.6 ± 2.3
Neuropathology	7.7 ± 0.6	4.0 ± 1.0	3.4 ± 1.3
Stains & CD markers/Immunohistochemistry	8.3 ± 0.6	3.3 ± 1.5	4.6 ± 2.1
Thoracic	8.0 ± 1.0	4.3 ± 2.1	4.7 ± 1.8
Clinical Pathology	6.7 ± 0.6	4.3 ± 1.2	3.4 ± 1.3
Total	**123.3 ± 2.3**	**74.3 ± 8.4**	**76.1 ± 14.4**

Table 2: Consistency of ChatGPT's responses

Outcome	1st to 2nd	1st to 3rd	2nd to 3rd
No change in response	127 (85%)	123 (82%)	120 (80%)
Correct to incorrect response	7 (5%)	11 (7%)	15 (10%)
Incorrect to another incorrect response	5 (3%)	6 (4%)	4 (3%)
Incorrect to correct response	11 (7%)	11 (7%)	11 (7%)

Table 3: Consistency of Bard's responses

Outcome	1st to 2nd	1st to 3rd	2nd to 3rd
No change in response	92 (61%)	81 (54%)	82 (55%)
Correct to incorrect response	20 (13%)	17 (11%)	17 (11%)
Incorrect to another incorrect response	19 (13%)	21 (14%)	19 (13%)
Incorrect to correct response	19 (13%)	31 (21%)	32 (21%)

Table 4: Examples of easy, intermediate, and difficult questions

Difficulty	Easy	Intermediate	Difficult
Question	A 25 year old man is found unresponsive on the bedroom floor of his secure residence. He is brought to the nearest emergency department. He has shallow breathing and a respiratory rate of 5 breaths per minute. He has pinpoint pupils and multiple needle tracks on his right arm. He eventually dies despite treatments. What is the most likely cause of death?	Which of the following immunohistochemical stains would help distinguish retroperitoneal myolipoma from a well-differentiated liposarcoma invading smooth muscle?	The clinical microbiology laboratory has isolated a gram-negative rod from the sputum of a homeless, alcoholic patient with a chronic cough. The isolate forms a mucoid colony that turns pink on MacConkey agar and produces a blue spot when placed on filter paper and treated with indole. What is the most likely organism?
Option	A. Accidental air embolism following intravenous drug injection B. Acute subarachnoid hemorrhage C. Benzodiazepines toxicity D. Cocaine intoxication E. Opiate overdose	A. CDK4 B. HMB45 C. S100 D. Smooth muscle actin E. Vimentin	A. Chlamydia pneumoniae B. Klebsiella oxytoca C. Klebsiella pneumoniae D. Pseudomonas aeruginosa
Answer	E	A	B
ChatGPT	Correct x3	Correct x3	Incorrect x3
Bard	Correct x3	Incorrect x3	Incorrect x2 & Correct x1
Trainees	14 trainees correct	7 trainees correct	No trainees correct

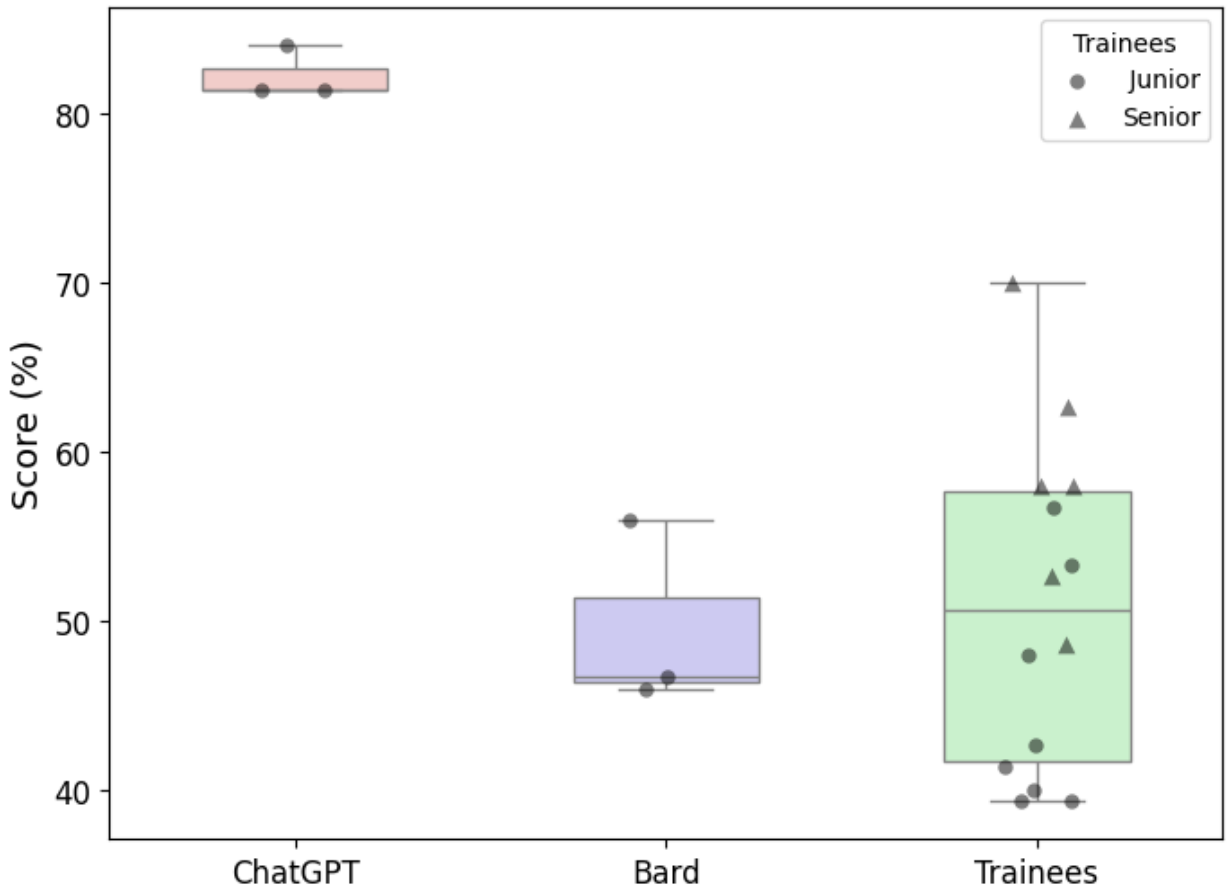

Figure 1: Comparison of the total scores between ChatGPT, Bard, and 14 trainees. The box plot illustrates that ChatGPT consistently achieves higher scores compared to Bard and trainees. Trainees are categorized into Junior (indicated by circles) and Senior (indicated by triangles).

Figure 2: Heatmap of each response. Each row in the heatmap corresponds to a question, with varying colors denoting answer choices (A to E) in each test. Each column represents the different testing rounds and the leftmost column, labeled "GT," denotes the correct answers.

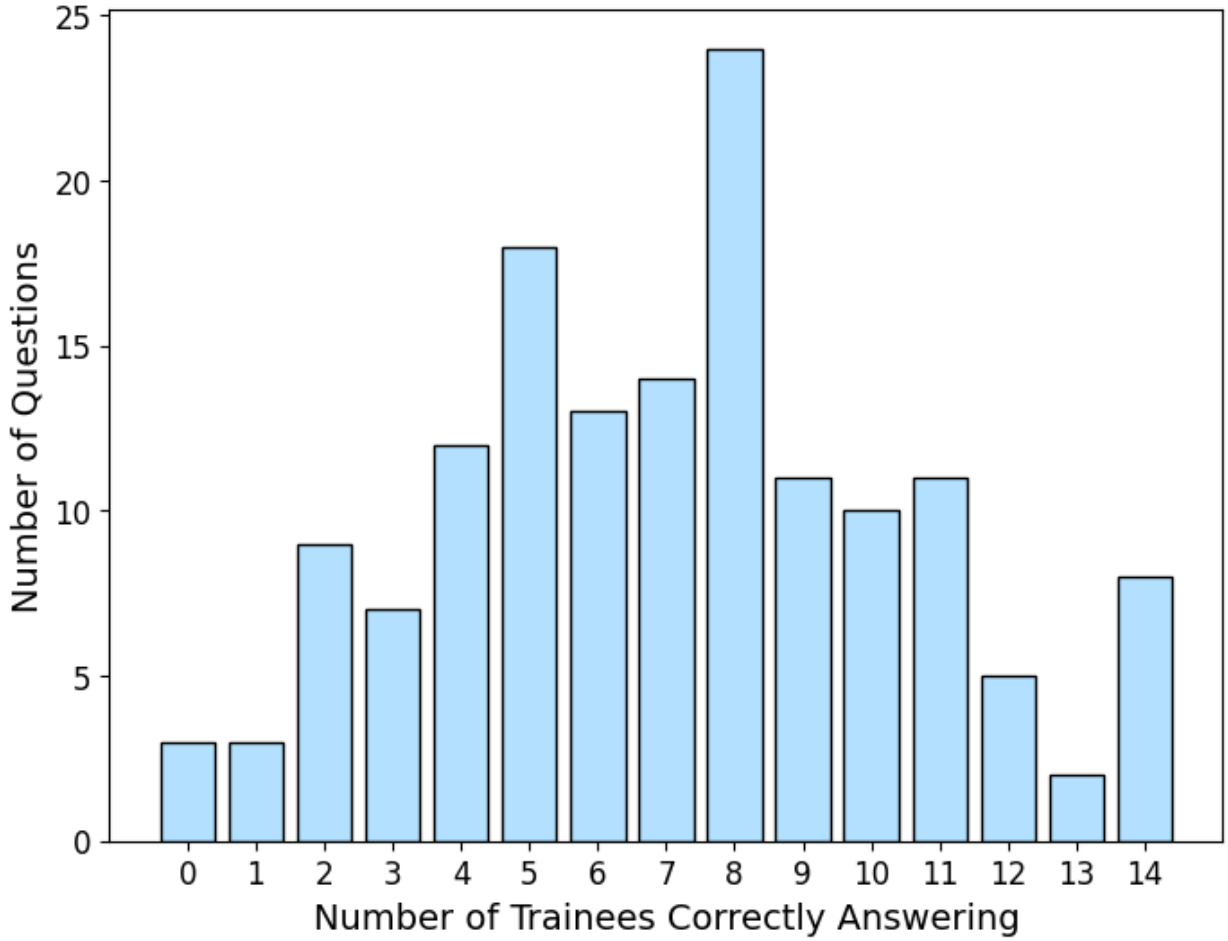

Figure 3: Distribution of Question Difficulty Based on Trainee Performance. The histogram illustrates the distribution of questions based on the number of trainees who correctly answered them. The x-axis represents the number of trainees who answered correctly, ranging from 0 (hardest questions, no trainees answered correctly) to 14 (easiest questions, all trainees answered correctly).

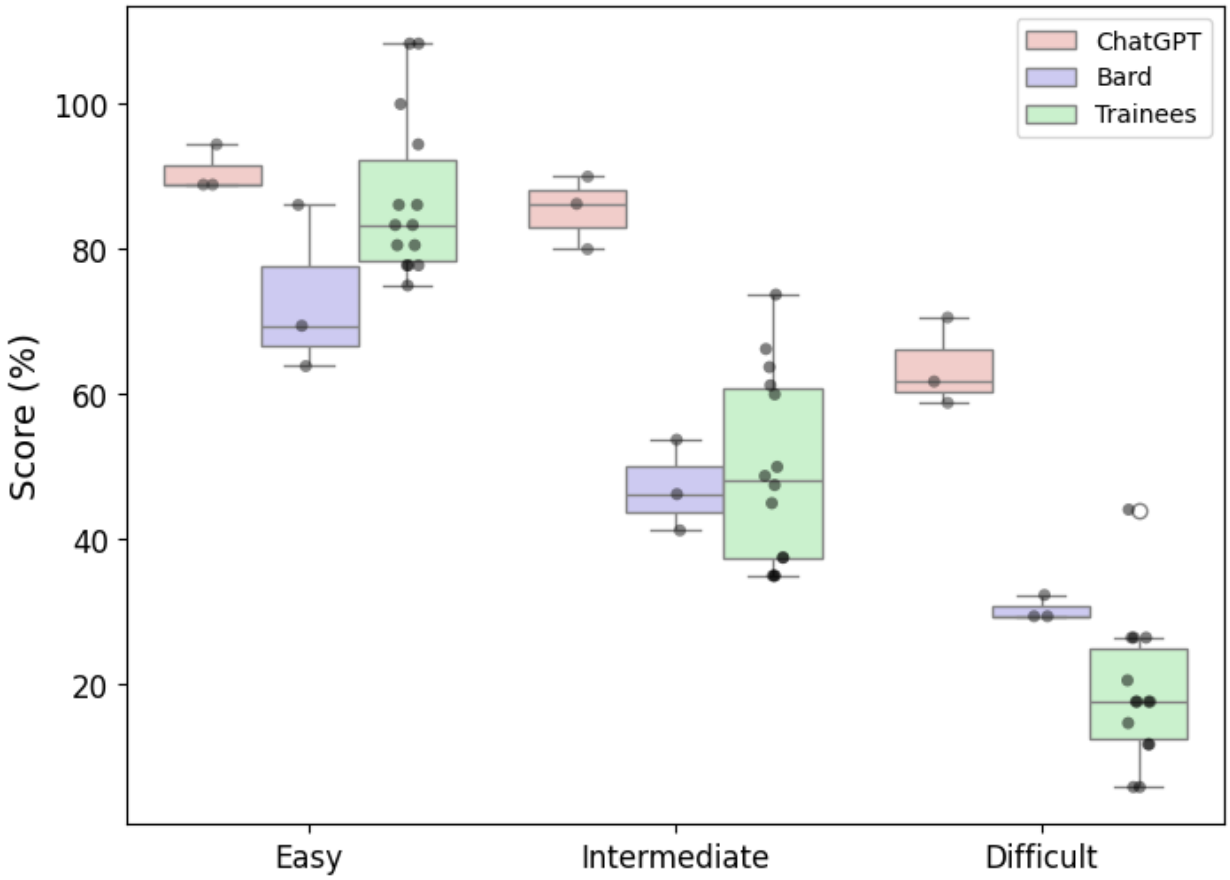

Figure 4: Box plot showing the performance of ChatGPT, Bard, and trainees categorized by question difficulty (Easy, Intermediate, Difficult). ChatGPT consistently outperforms Bard and trainees and show higher accuracy across all difficulty levels. Both LLMs and trainees display a similar trend, with higher scores on easy questions and lower scores on difficult questions. Individual trainee scores are indicated by dots, illustrating variability among the trainees.