

Performance of Language Models on the Family Medicine In-Training Exam

Rana E. Hanna, BS^a; Logan R. Smith, BA^a; Rahul Mhaskar, PhD^b; Karim Hanna, MD^{a,c}

AUTHOR AFFILIATIONS:

^aMorsani College of Medicine, University of South Florida, Tampa, FL

^bDepartment of Medical Education, Morsani College of Medicine, University of South Florida, Tampa, FL

^cDepartment of Family Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL

CORRESPONDING AUTHOR:

Karim Hanna, Morsani College of Medicine, University of South Florida, Tampa, FL, khanna@usf.edu

HOW TO CITE: Hanna RE, Smith LR, Mhaskar R, Hanna K. Performance of Language Models on the Family Medicine In-Training Exam. *Fam Med*. 2024;56(X):1-6.

doi: [10.22454/FamMed.2024.233738](https://doi.org/10.22454/FamMed.2024.233738)

PUBLISHED: 12 August 2024

KEYWORDS: artificial intelligence, board exams, family medicine, intraining exams

© Society of Teachers of Family Medicine

ABSTRACT

Background and Objectives: Artificial intelligence (AI), such as ChatGPT and Bard, has gained popularity as a tool in medical education. The use of AI in family medicine has not yet been assessed. The objective of this study is to compare the performance of three large language models (LLMs; ChatGPT 3.5, ChatGPT 4.0, and Google Bard) on the family medicine in-training exam (ITE).

Methods: The 193 multiple-choice questions of the 2022 ITE, written by the American Board of Family Medicine, were inputted in ChatGPT 3.5, ChatGPT 4.0, and Bard. The LLMs' performance was then scored and scaled.

Results: ChatGPT 4.0 scored 167/193 (86.5%) with a scaled score of 730 out of 800. According to the Bayesian score predictor, ChatGPT 4.0 has a 100% chance of passing the family medicine board exam. ChatGPT 3.5 scored 66.3%, translating to a scaled score of 400 and an 88% chance of passing the family medicine board exam. Bard scored 64.2%, with a scaled score of 380 and an 85% chance of passing the boards. Compared to the national average of postgraduate year 3 residents, only ChatGPT 4.0 surpassed the residents' mean of 68.4%.

Conclusions: ChatGPT 4.0 was the only LLM that outperformed the family medicine postgraduate year 3 residents' national averages on the 2022 ITE, providing robust explanations and demonstrating its potential use in delivering background information on common medical concepts that appear on board exams.

INTRODUCTION

Artificial intelligence (AI) has grown in popularity recently with increased application in many fields, including medicine. Large language models (LLMs) are deep learning models that aim to generate humanlike responses; LLMs are pretrained on a vast amount of information, and unlike search engines, they produce de novo responses to the inputs they receive. ChatGPT and Bard are publicly available chat-based generative AI developed by OpenAI and Google, respectively. The newest model, ChatGPT 4.0, has been shown to outperform ChatGPT 3.5 and other LLMs on most exams taken, including the bar exam, LSAT, SAT, Medical Knowledge Self-Assessment Program, and many others.¹ Interestingly, other researchers have investigated ChatGPT's performance on ophthalmology² and neurosurgery³ board review questions; however, LLM performance on family medicine board exams has not been evaluated. Given that ChatGPT 4.0 is trained using larger parameters than previous models, this LLM scored in the 90th percentile on a sample bar exam, while ChatGPT 3.5 scored in the bottom 10%.¹ ChatGPT 4.0 has limitations similar to previous models, yet fewer hallucinations. A hallucination is a term used in the AI field to refer to a coherent yet untrue AI-generated response.¹

Unlike its predecessor, ChatGPT 4.0 utilizes computer vision to analyze images uploaded by users. For instance, AI also can help diagnose pathologies like diabetic retinopathy and skin lesions;⁴ computer vision is useful in medicine because physical exam findings drive many diagnoses. The addition of computer vision to ChatGPT 4.0 enables this LLM to generate responses to ITE questions that use images in the question stem.¹ Computer vision enhances ChatGPT's test-taking abilities in image-based questions. However, ChatGPT has limitations that prevent users from utilizing AI for diagnostics that would best be left to trained clinicians.

AI has demonstrated the potential to analyze a clinical presentation, generate differential diagnoses, and develop a clinical workup plan and treatment options.⁵ This study investigates whether LLMs can perform as well as a postgraduate year 3 (PGY-3) resident on the family medicine ITE, highlighting their potential and limitations as supplementary tools for exam preparation rather than asserting their usefulness as teaching or review tools. A 2023 study evaluated ChatGPT 3.5's performance on the United States Medical Licensing Exams Step 1, Step 2CK, and Step 3.⁶ These exams are taken during the second and third years of medical school and the

first year of residency, respectively. ChatGPT 3.5 scored at or near the passing threshold (60%) for all 3 exams, with Step 1 being the lowest score.⁶ ChatGPT 3.5's explanations also provided significant insight, defined as a "novel, nonobvious and accurate response," in 88.9% of questions for all exams in the study, including open-ended or multiple-choice formatted questions; this finding demonstrates AI's potential as a tool for students working toward a medical license because it answers and details the reasoning behind the question and correct answer.⁶

AI's ability to pass several medical and nonmedical exams inspired this study, which aims to determine whether AI can be a reliable revision tool for family medicine residents studying for their board exams using the ITE. The American Board of Family Medicine (ABFM) writes the ITE. The same organization administers the official board exams, making it a dependable predictor of success in terms of board exam scores. Further, because LLM platforms differ, this study aimed to determine which LLM is most reliable and which can be best used to the advantage of residents: ChatGPT 3.5, ChatGPT 4.0, or Bard.

METHODS

Data Collection

The performance of ChatGPT 3.5, ChatGPT 4.0, and Bard was evaluated using the 2022 family medicine ITE. The ITE is a 200-question multiple-choice exam, with a scaled score out of 800, used to gauge residents' progress throughout their training; the Bayesian score predictor estimates the probability of passing the family medicine certification exam.⁷

Of the 200 questions, 193 were used for the study because seven questions (questions 21, 63, 97, 99, 138, 157, 166) were eliminated by the ABFM due to psychometric or content reasons. These 193 questions were copied and pasted verbatim into the chat box individually. Questions were preceded by the following prompt: "Answer the question and pick the correct answer choice." After each question, a new session was started to limit learning from previous questions. ChatGPT 3.5 accepts only text input; thus, to maintain consistency across the three LLMs, no images were included in the questions with corresponding images. The LLMs were asked to generate an answer to only the text portion of these questions. The correct answer was designated as 1, incorrect was 0. For any incorrect answer choice, the incorrect letter chosen was recorded. Any unanswered question was noted and counted incorrectly.

The performance of each of the LLMs was scaled using the ABFM raw-to-scaled score conversion table. The scaled score then was input into the Bayesian score predictor, assuming the LLM was performing at a PGY-3 level, to estimate the probability of passing the board exam.⁷

Statistical Analysis

The association between the performance of ChatGPT 3.5, ChatGPT 4.0, and Bard and the question categories was assessed using the χ^2 test to investigate the relationship between the specialty categories and the correct versus incorrect responses for each LLM. The agreement across

responses (correct/incorrect) from ChatGPT 3.5, ChatGPT 4.0, and Bard was investigated via Cohen's κ coefficient, reporting the coefficient and the 95% confidence intervals (CIs).⁸ The strength of agreement was interpreted using the scale shown in Table 1.⁹ *P* values less than .05 denoted statistical significance. All analyses were conducted by MedCalc software version 22.017 (MedCalc Software Ltd).

TABLE 1. Cohen's κ Interpretation *

Value of κ	Strength of agreement
<0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

*Cohen's κ values and corresponding strength of agreement ranging from poor, fair, moderate, good and very good.

RESULTS

Out of the 193 questions, ChatGPT 4.0 scored the highest, followed by ChatGPT 3.5 and Bard. ChatGPT 4.0 scored 167/193 (86.5%) with a scaled score of 730. According to the Bayesian score predictor, assuming the LLM was performing at the level of a PGY-3 resident, ChatGPT 4.0 has a 100% chance of passing the family medicine board exam. ChatGPT 3.5 scored 128/193 (66.3%), translating to a scaled score of 400 and an 88% chance of passing the family medicine board exam. Bard correctly answered 124 out of 193 (64.2%) questions, resulting in a scaled score of 380; according to the Bayesian score predictor, this result confers an 85% chance of passing the boards. Of Bard incorrect questions, 16 were unanswered (8.3%); of ChatGPT 4.0 incorrect questions, one was unanswered (0.5%). Compared to the national average of PGY-3 residents, only ChatGPT 4.0 surpassed the residents' mean of 132/193 (68.4) with a scaled average of 433 and a Bayesian prediction of a 93% chance of passing.

The 193 questions were divided into 16 categories based on topic, and the performance of the three LLMs by topic was evaluated (Table 2 ;Figure 1). Except for the male reproductive and neurologic categories, Bard consistently scored lower in raw scores compared to ChatGPT 3.5 or ChatGPT 4.0. This observation is based on the absolute scores without a statistical test for significance. The χ^2 analysis demonstrated no significant difference in performance by topic for ChatGPT 3.5, ChatGPT 4.0, or Bard (*P*=.569, .763, and .129, respectively; Table 2).

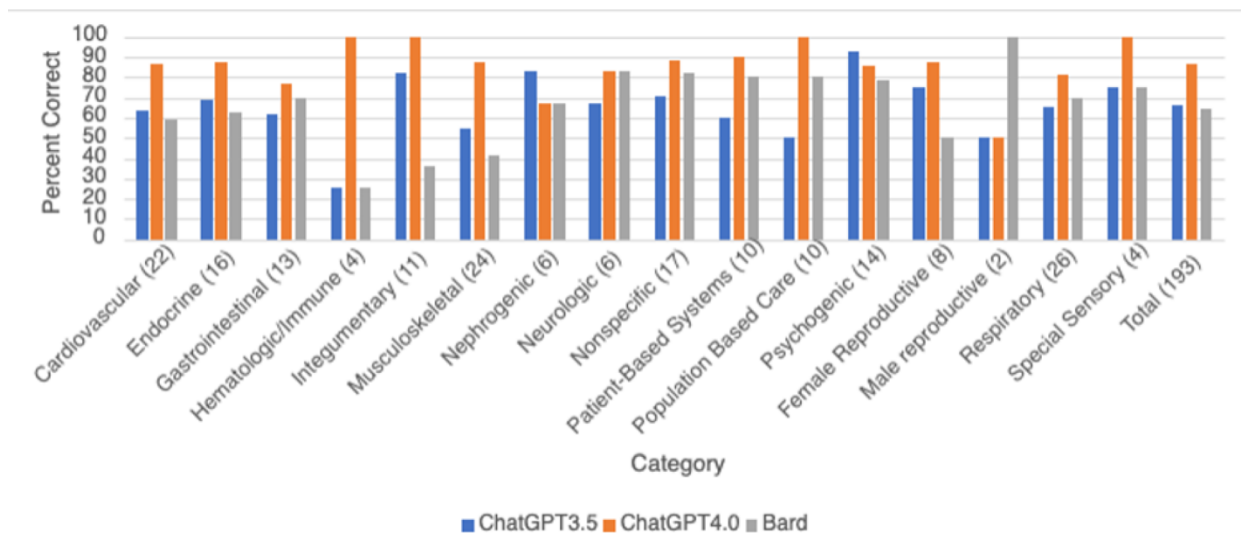
For interrater agreement overall, using Cohen's κ on the three LLMs revealed that ChatGPT 3.5 and Bard had the highest interrater agreements (weighted κ =0.383, 95% CI=0.247–0.519), indicating a fair-to-moderate agreement (Table 3). Meanwhile, ChatGPT 4.0 and Bard had the lowest interrater

TABLE 2. χ^2 Analysis of ChatGPT 3.5, ChatGPT 4.0, and Bard by Category *

Category	Number of questions, n (%)	ChatGPT 3.5 correct, n (%)	ChatGPT 4.0 correct, n (%)	Bard correct, n (%)
1. Cardiovascular	22 (11.4)	14/22 (63.6)	19/22 (86.4)	13/22 (59.1)
2. Endocrine	16 (8.3)	11/16 (68.7)	14/16 (87.5)	10/16 (62.5)
3. Gastrointestinal	13 (6.7)	8/13 (61.5)	10/13 (76.9)	9/13 (69.2)
4. Hematologic/ immune	4 (2.1)	1/4 (25.0)	4/4 (100.0)	1/4 (25.0)
5. Integumentary	11 (5.7)	9/11 (81.8)	11/11 (100.0)	4/11 (36.4)
6. Musculoskeletal	24 (12.4)	13/24 (54.2)	21/24 (87.5)	10/24 (41.7)
7. Nephrogenic	6 (3.1)	5/6 (83.3)	4/6 (66.7)	4/6 (66.7)
8. Neurologic	6 (3.1)	4/6 (66.7)	5/6 (83.3)	5/6 (83.3)
9. Nonspecific	17 (8.8)	12/17 (70.6)	15/17 (88.2)	14/17 (82.4)
10. Patient-based systems	10 (5.2)	6/10 (60.0)	9/10 (90.0)	8/10 (80.0)
11. Population-based care	10 (5.2)	5/10 (50.0)	10/10 (100.0)	8/10 (80.0)
12. Psychogenic	14 (7.3)	13/14 (92.9)	12/14 (85.7)	11/14 (78.6)
13. Female reproductive	8 (4.1)	6/8 (75.0)	7/8 (87.5)	4/8 (50.0)
14. Male reproductive	2 (1.0)	1/2 (50.0)	1/2 (50.0)	2/2 (100.0)
15. Respiratory	26 (13.5)	17/26 (65.4)	21/26 (80.8)	18/26 (69.2)
16. Special sensory	4 (2.1)	3/4 (75.0)	4/4 (100.0)	3/4 (75.0)
Total	193 (100)	128/193 (66.3)	167/193 (86.5)	124/193 (64.2)
P value	–	.569	.763	.129

*Performance of ChatGPT 3.5, ChatGPT 4.0, and Bard on the 2022 family medicine in-training exam. Categorical performance is reported, and the association between the specialty categories and the correct versus incorrect artificial intelligence responses were evaluated with a χ^2 test.

FIGURE 1. Categorical Performance of ChatGPT 3.5, ChatGPT 4.0, and Bard



Histogram portraying the performance of ChatGPT 3.5 (blue), ChatGPT 4.0 (orange), and Bard (gray) on the 2022 family medicine in-training exam overall and per category.

agreement (weighted $\kappa=0.280$, 95% CI=0.153–0.407), indicating poor-to-fair agreement (Table 3). Regarding incorrect answers, the interrater agreement revealed that if ChatGPT 4.0 chose a specific incorrect answer, it was substantially likely that ChatGPT 3.5 and Bard chose the same incorrect answer, with a moderate-to-very good agreement. (Table 3).

In addition to the differences in raw scores, the quality of answers differed among the three LLMs (Figure 2). While ChatGPT 4.0 and Bard defended the selected answer by providing background information on the question topic and addressing the incorrect answers, ChatGPT 3.5 often just stated the correct answer, providing little context.

DISCUSSION

ChatGPT 4.0 scored 86.5% on the family medicine ITE, providing valuable background information and explanations. While not suitable as a primary study tool due to a 13% error rate, it demonstrates potential as a supplementary resource for clarifying complex concepts. ChatGPT 3.5 and Bard performed below the mean score of PGY-3 residents and thus are not as good of a resource as ChatGPT 4.0 for resident use; these LLMs have a higher possibility of providing misinformation due to an increased rate of hallucinations.

Over 75% of students reported inadequate general knowledge about AI in health care, and 67% harbored positive attitudes about AI implementation in health care.¹⁰ Given the 13% error rate, cross-referencing LLM responses with verified sources is crucial for users. LLMs should not replace traditional study materials but can provide useful supplementary explanations and insights. At the time of this study, ChatGPT 4.0 required a paid monthly subscription; when deciding on study resources, the LLMs' significant difference in scores may factor into buyers' decisions on whether to purchase ChatGPT 4.0 or use the free, yet less accurate models like Bard and ChatGPT 3.5. Of note, ChatGPT 4.0, the highest scoring LLM in this study, still did not achieve a perfect score and left one question unanswered; so using LLMs as a supplement to other board prep materials is proposed because their valuable explanations may clarify confusing concepts. AI is rapidly advancing; while most AI platforms perform exceptionally well on the ITE, a lot of improvement is necessary and a very low rate of hallucinations must be reached before an AI can be used as an independent teaching tool rather than a supplementary resource. Of note, medical doctors often do not obtain 100% accuracy on exams yet are very competent in treating their patients. Despite AI's imperfect performance, disqualifying it as an ideal teaching tool, AI, particularly ChatGPT 4.0, retains its value because it can provide insight on difficult questions and topics.

Moreover, besides the male reproductive category, whose data set was too small to reveal significant information ($n=2$), and the neurologic category, Bard's categorical raw score was consistently below that of ChatGPT 3.5 or ChatGPT 4.0, or both. However, while Bard had the lowest score (64.2%), 16 of the incorrect questions were left unanswered (8.2%), stating, "I am

a text-based AI and can't assist with that." The rate of incorrect answers was used to gauge the rate of hallucinations among the AI models. Bard achieved the lowest score on the 2022 ITE, thus it is more prone to generate hallucinations than ChatGPT 3.5, making it the least reliable resource of those evaluated in this study. Users should be cautious and aware of the risk of misinformation when using any LLM, especially Bard, which has proved to have the highest rate of hallucinations among these three LLMs.¹¹

Further, there's a discrepancy in the quality of responses between ChatGPT 4.0 and Bard versus ChatGPT 3.5. ChatGPT 4.0 and Bard outlined detailed background information and discredited the other options with supporting evidence, providing the user with broader knowledge. ChatGPT 3.5 most often only stated the correct answer, including a short explanation, if any, in support of its choice and failing to mention why other options were not selected.

The interrater agreements demonstrated that the incorrect answers picked by ChatGPT 4.0 substantially agreed with those chosen by both ChatGPT 3.5 and Bard; on the other hand, ChatGPT 3.5 and Bard's incorrect answers only moderately agreed with each other. Learning why incorrect answers are wrong is just as important as choosing the correct one; because no data are available on how residents performed per question and the most common answer picked is unknown, the incorrect answer chosen by ChatGPT 4.0 can be used to identify and avoid common pitfalls residents may fall into when selecting an answer.

As a supplementary resource, AI can answer the practice question and then, to eliminate the effect of hallucinations, students and residents should cross reference the results with the answer key to make sure it is the proper answer. Once the answer is known to be correct, the AI's reasoning and background information can likely be useful to review material previously learned. Moreover, AI can reword or rework its response into a table, bullet points, or other formats per user preference. As AI improves, this study, using a different ITE exam, can be replicated to gauge the progress of LLMs and their ability to aid in resident and student revision of practice exams.

Passing the ITE and board exams cannot be equated with the ability of a health care provider and does not take away from the crucial role physicians play in the care of their patients. This study sheds light on how LLMs can be used as a supplement to residents' exam preparation; however, further studies evaluating LLMs' capability to follow medical protocol and guidelines in patient care are necessary to determine how LLMs will impact the medical field.

LIMITATIONS

This study had several limitations, including the evaluation of LLMs on multiple-choice questions only. The high error rate limits the extrapolation of AI's performance as an effective teaching tool. Future studies should investigate the quality and accuracy of AI's explanations in medical education.

TABLE 3. Interrater Agreement Overall and for Incorrect Answer Choices for ChatGPT 3.5, ChatGPT 4.0, and Bard *

		Bard and ChatGPT 3.5	Bard and ChatGPT 4.0	ChatGPT 3.5 and ChatGPT 4.0
Agreement: overall performance	Weighted κ	0.383	0.280	0.360
	95% confidence interval	0.247–0.519	0.153–0.407	0.229–0.491
Agreement: incorrect answer choice	Weighted κ	0.400	0.689	0.649
	95% confidence interval	0.150–0.650	0.435–0.943	0.389–0.910

*Interrater agreements between Bard and ChatGPT 3.5, Bard and ChatGPT 4.0, and ChatGPT 3.5 and ChatGPT 4.0. This table reflects the level of agreement of the three large language models on the in-training exam overall and the selected incorrect answer choice.

FIGURE 2. Examples of Responses for ChatGPT 4.0, Bard, and ChatGPT 3.5

A

You
answer the following question and pick the correct answer choice.
Which one of the following regimens is recommended for the treatment of hypertension in a patient with stage 3 chronic kidney disease and proteinuria?
A) A loop diuretic and a β -blocker
B) An ACE inhibitor and an angiotensin receptor blocker
C) An ACE inhibitor and a thiazide diuretic
D) A calcium channel blocker and a thiazide diuretic
E) A potassium-sparing diuretic and a thiazide diuretic

ChatGPT
The recommended treatment for hypertension in patients with chronic kidney disease (CKD), especially those with proteinuria, is to target the renin-angiotensin-aldosterone system (RAAS) because of its beneficial effects on reducing proteinuria and potentially slowing the progression of CKD.

Choice B, "An ACE inhibitor and an angiotensin receptor blocker," represents a combination therapy targeting the RAAS. However, it's important to note that while both ACE inhibitors and angiotensin receptor blockers (ARBs) can individually be used in this setting, combining them can lead to adverse effects such as hyperkalemia, worsening kidney function, and hypotension. Therefore, they are generally not used together unless there's a specific indication and under careful monitoring.

Choice C, "An ACE inhibitor and a thiazide diuretic," offers an antihypertensive effect through the ACE inhibitor (targeting the RAAS) and a diuretic effect from the thiazide. This combination is more typical for treating hypertension in CKD patients with proteinuria. Thiazide diuretics, however, might be less effective in stage 3 CKD due to reduced glomerular filtration rate, but they can still be used.

Based on current guidelines and the information given, the most appropriate answer would be: C) An ACE inhibitor and a thiazide diuretic.

However, individual patient factors, as well as frequent monitoring of kidney function and potassium, are essential when choosing and adjusting antihypertensive therapy.

B

You
answer the following question and pick the correct answer choice.
Which one of the following regimens is recommended for the treatment of hypertension in a patient with stage 3 chronic kidney disease and proteinuria?
A) A loop diuretic and a β -blocker...

The correct answer is C. An ACE inhibitor and a thiazide diuretic.

ACE inhibitors and ARBs are the first-line treatment for hypertension in patients with CKD and proteinuria. They work by blocking the renin-angiotensin-aldosterone system (RAAS), which is a major driver of hypertension and kidney damage in CKD. Thiazide diuretics are also effective in lowering blood pressure and reducing proteinuria in CKD.

The other answer choices are not as well-supported by evidence. Loop diuretics are effective in lowering blood pressure, but they can also worsen kidney function. β -blockers are not as effective as ACE inhibitors or ARBs in lowering blood pressure in CKD patients. Calcium channel blockers are generally not recommended as first-line treatment for hypertension in CKD patients. Potassium-sparing diuretics can be used in combination with other antihypertensive medications, but they can increase the risk of hyperkalemia, especially in patients with CKD.

Here is a summary of the recommended treatment regimen for hypertension in patients with stage 3 CKD and proteinuria:

• **First-line therapy:** ACE inhibitor or ARB + thiazide diuretic
• **Second-line therapy:** Add a mineralocorticoid receptor antagonist (MRA) or a non-dihydropyridine calcium channel blocker (CCB)
• **Third-line therapy:** Add a centrally acting antihypertensive agent, such as clonidine or guanfacine

It is important to note that this is just a general guideline. The best treatment regimen for a particular patient will depend on their individual needs and comorbidities.

C

You
answer the following question and pick the correct answer choice.
Which one of the following regimens is recommended for the treatment of hypertension in a patient with stage 3 chronic kidney disease and proteinuria?
A) A loop diuretic and a β -blocker
B) An ACE inhibitor and an angiotensin receptor blocker
C) An ACE inhibitor and a thiazide diuretic
D) A calcium channel blocker and a thiazide diuretic
E) A potassium-sparing diuretic and a thiazide diuretic

ChatGPT
The recommended regimen for the treatment of hypertension in a patient with stage 3 chronic kidney disease and proteinuria is:

C) An ACE inhibitor and a thiazide diuretic

This combination helps to control blood pressure and also has potential renal protective effects.

Screenshots of responses of the three LLMs to capture the difference in the quality of answers and explanations between ChatGPT 4.0 (A), Bard (B), and ChatGPT 3.5 (C) to the same multiple-choice question about hypertension treatment in a patient with chronic kidney disease. A. ChatGPT 4.0 gives background information on targeting the renin-angiotensin-aldosterone system in managing hypertension. Not only does ChatGPT 4.0 clearly explain why it chose the correct answer (ACE inhibitor and Thiazide), but it goes on to explain why the other options are incorrect. B. Bard's method is similar to that of ChatGPT 4.0; it also provides details of the desired mechanism of action to target the proper pathway and states why other answers are not the first-line treatment. C. On the other hand, ChatGPT 3.5 provides no contextual information on the underlying mechanisms of hypertension and why other answers are incorrect.

ChatGPT 3.5 does not accept images. Therefore, questions with associated figures were answered using text-only information for all three LLMs, which may have affected their performance. Further studies should evaluate AI's capability to interpret visual data. Medicine is a visual field; another study evaluating AI's ability to identify physical presentations, such as certain skin lesions and rashes, would be beneficial. Moreover, due to continual improvement and learning of AI, the LLMs may have performed better toward the end of the weeklong data collection period. This change in performance may be minimized by collecting the data within the same 24-hour period.

Also, many categories had small sample sizes that were insufficient for determining statistical significance, resulting in a lack of P values for the categorical analysis. Lastly, no published information was available on residents' performance per category or question; thus, we could not determine the statistical significance between residents' and the LLMs' performance.

CONCLUSIONS

In conclusion, ChatGPT 4.0 was the only LLM that outperformed the family medicine PGY-3 residents' national averages on the 2022 ITE. While ChatGPT 4.0 demonstrates potential as a supplementary resource, it should not be relied upon as a primary study tool due to its error rate. Users should be cautious and cross-reference AI-generated information with reliable sources. While ChatGPT 3.5 and Google Bard had an 85% chance or higher of passing the board exams, they both scored below the PGY-3 national average. In addition, ChatGPT 4.0 scored 86.5% on the 2022 ITE and is, therefore, not always accurate. While AI can be used as a supplementary resource for residents, all users need to be aware of LLMs' limitations and to use them with caution to avoid learning false information.

REFERENCES

1. GPT-4 Technical Report. *Open AI*. 2023. <https://cdn.openai.com/papers/gpt-4.pdf>.
2. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141(8):798–800.
3. Hopkins BS, Nguyen VN, Dallas J. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. *J Neurosurg*. 2023;139(3):904–911.
4. Lin S. A clinician's guide to artificial intelligence (AI): why and how primary care should lead the health care AI revolution. *J Am Board Fam Med*. 2022;35(1):175–184.
5. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J*. 2019;6(2):94–98.
6. Kung TH, Cheatham M, Medenilla A. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023;2(2):198.
7. Bayesian score predictor. *American Board of Family Medicine*. 2024. <https://rtm.theabfm.org/bayesian/predictor>.
8. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213–220.
9. Altman DG. *Practical Statistics for Medical Research*. Chapman & Hall/CRC; 1991.
10. Busch F, Hoffmann L, Truhn D. Medical students' perceptions towards artificial intelligence in education and practice: a multinational, multicenter cross-sectional study. *medRxiv*. 2023(December).
11. Ali R, Tang OY, Connolly ID. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurg*. 2023;93(5):90–91.