Overconfident AI? Benchmarking LLM Self-Assessment in Clinical Scenarios

Mahmud Omar[1], Benjamin S Glicksberg[1], Girish N Nadkarni[1], Eyal Klang[1].

1- Division of Data-Driven and Digital Medicine (D3M), Department of Medicine, Icahn School of Medicine at Mount Sinai, NY.

Acknowledgment – We thank Dr. Uriel Katz and the co-authors of the paper "GPT versus Resident Physicians — A Benchmark Based on Official Board Scores, DOI: 10.1056/AIdbp2300192" for their significant contributions and for sharing the MCQ dataset which greatly facilitated our research.

Financial disclosure – This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interest – None declared.

Ethical approval was not required for this research, as it utilized already published synthetic data.

Corresponding author:

Mahmud Omar M.D.

E-mail: Mahmudomar70@gmail.com

Abstract

Background and Aim: Large language models (LLMs) show promise in healthcare, but their self-assessment capabilities remain unclear. This study evaluates the confidence levels and performance of 12 LLMs across five medical specialties to assess their ability to accurately judge their responses.

Methods: We used 1965 multiple-choice questions from internal medicine, obstetrics and gynecology, psychiatry, pediatrics, and general surgery. Models were prompted to provide answers and confidence scores. Performance and confidence were analyzed using chi-square tests and t-tests. Consistency across question versions was also evaluated.

Results: All models displayed high confidence regardless of answer correctness. Higher-tier models showed slightly better calibration, with a mean confidence of 72.5% for correct answers versus 69.4% for incorrect ones, compared to lower-tier models (79.6% vs 79.5%). The mean confidence difference between correct and incorrect responses ranged from 0.6% to 5.4% across all models. Four models showed significantly higher confidence when correct ($p<0.01$), but the difference remained small. Most models demonstrated consistency across question versions.

Conclusion: While newer LLMs show improved performance and consistency in medical knowledge tasks, their confidence levels remain poorly calibrated. The gap between performance and self-assessment poses risks in clinical applications. Until these models can reliably gauge their certainty, their use in healthcare should be limited and supervised by experts. Further research on human-AI collaboration and ensemble methods is needed for responsible implementation.

Keywords: Large Language Models (LLMs), Safe AI, AI Reliability, Clinical knowledge.

Introduction

The integration of artificial intelligence (AI) into healthcare has marked a significant evolution in the field, with LLMs at the forefront of this transformation (1). With their capacity to understand and generate human-like text, LLMs are poised to support healthcare professionals in complex decision-making and learning processes (1,2).

A wide array of LLMs is now accessible, offering solutions that cater to both the public and medical professionals (1). The efficacy of these models as educational and auxiliary tools has been demonstrated, albeit with some limitations (3,4). For instance, LLMs like Generative Pre-trained Transformers (GPT) have shown promise in providing diagnostic assistance and answering medical queries, suggesting a growing potential for these technologies in practical applications (3,5–7). Katz et al. demonstrated that GPT-4 not only improved clinically compared to its predecessor GPT-3.5, but also matched physician performance in certain areas (8). Yet, there is evidence of "hallucinations" and inaccuracies in model outputs, which could lead to harm in clinical decision-making (9,10).

One important issue is the growing need for fair AI use, focusing on developing explainable AI for safer, more knowledgeable application (11,12). However, the literature shows ambiguity in how LLMs process and output data, resembling a black box (11,13,14). Recent studies also show that LLMs tend to be overly confident when expressing confidence (15). This could be harmful, complicating the identification of errors in daily clinical decision-making.

Our goal is to benchmark widely used LLMs by assessing their confidence in answering clinical questions. We aim to determine if these models can accurately judge when to be confident in their responses.

Materials and Methods

Study Design and Data Source

This study utilizes a compiled dataset from a previous study by Katz et al., which includes 655 questions across five medical specialties: internal medicine, obstetrics and gynecology (OBGYN), psychiatry, pediatrics, and general surgery (8). Crafted from internationally recognized textbooks and guidelines, this dataset serves as a standardized framework for assessment (16–20).

To enhance benchmarking reliability, each original question was rephrased twice using GPT-4 Application Programming Interface (API) in Python, yielding 1965 questions. The prompts ensured preservation of the original medical content and correct answers (21). A random 20% sample from each field underwent manual validation by two expert physicians, confirming accuracy and consistency with the original questions.

Model Setup and Configuration

The LLMs employed in this study were prompted to return the correct answer along with a confidence score for each choice (“A”, “B”, “C”, “D”), expressed as a percentage. The models were executed using API codes in a dedicated server with 4xH100 80GB GPUs, with the corresponding codebase accessible on GitHub. We utilized Python (3.10) for data analyses. We used several Python libraries to facilitate data processing, model interaction, and analysis: NumPy (1.26.4) for numerical computations, Pandas (2.1.4) for data manipulation, Scikit-Learn (1.3.0) for statistical analysis, Hugging Face's Transformers (4.37.2) and torch (2.2.2+cu121) for accessing pre-trained NLP models, and the Json module (2.0.9) for handling JSON data formats.

We used the default hyperparameters for each model, ensuring a suitable trade-off between maintaining high output quality and introducing a controlled level of variability into our generated responses (22).

Benchmarked LLMs

We benchmarked 12 LLMs in two groups—one for the newest or largest models, and the other for older or smaller models (**Figure 1**). This includes state-of-the-art models like GPT-4o and Claude Sonnet 3.5.

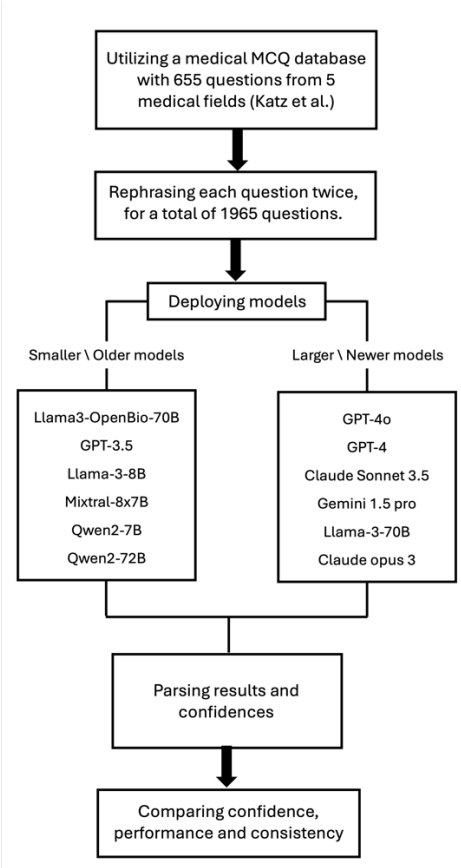


Figure 1: A flowchart representing the evaluation methodology.

Statistical analysis

Statistical analyses were conducted to evaluate model performance and confidence across medical specialties. Chi-square tests assessed overall performance differences within each field, using proportions of correct responses. Post-hoc pairwise

comparisons with Bonferroni correction identified specific inter-model differences. Confidence levels were compared between correct and incorrect responses for each model using two-sample t-tests. Mean confidence scores were calculated for higher-tier and lower-tier models, as well as across all models. Performance consistency was evaluated by comparing confidence gaps between correct and incorrect responses. All statistical tests used a significance level of $\alpha = 0.05$. Analyses were performed using R version 4.1.2 (R Core Team, 2021).

Results

Confidence analysis

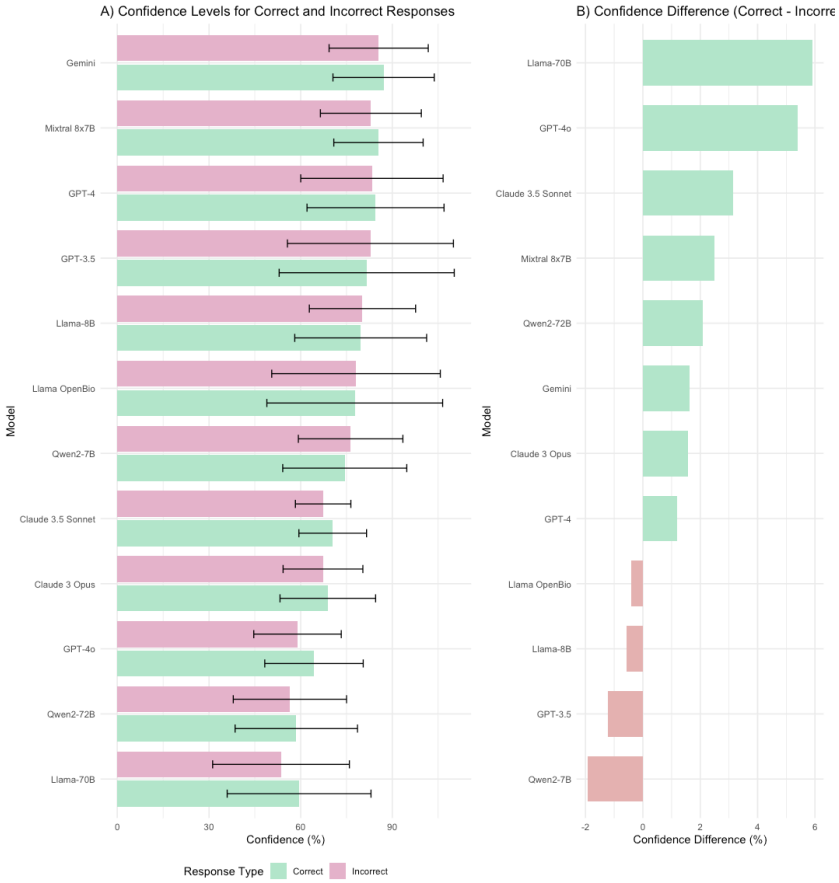
The study examined confidence levels of 12 language models when providing correct and incorrect answers. The mean confidence across all models was 76.1% when correct and 74.4% when incorrect, with an overall difference of 1.7 percentage points.

Higher-tier models showed a mean confidence of 72.5% when correct and 69.4% when incorrect, while lower-tier models displayed 79.6% confidence when correct and 79.5% when incorrect. The average difference between correct and incorrect responses was 3.1 percentage points for higher models and 0.1 for lower models (**Figure 2, Table 1**).

Four models (GPT-4.0, Llama-70B, Claude 3.5 Sonnet, and Qwen2-72B) demonstrated significantly higher confidence when correct ($p<0.01$). Gemini exhibited the highest overall confidence levels (85.6% when incorrect, 87.2% when correct), although the difference was not statistically significant ($p=0.35$).

Claude 3.5 Sonnet showed the lowest confidence levels (67.4% when incorrect, 70.5% when correct). Qwen2-7B was unique in displaying higher confidence when incorrect (76.4% vs 74.5% when correct, $p=0.01$).

GPT-3.5 and Llama-OpenBio-70B revealed minimal differences in confidence between correct and incorrect answers ($p>0.8$). The largest confidence gap was observed in GPT-4.0 (5.39 percentage points), while Llama-8B had the smallest gap (0.58 percentage points) (**Table 1**).



*(A) - Displays the average confidence and 95% confidence intervals for each model, categorized by correct answers (green) and incorrect answers (red).

(B) - Shows the differences in average confidence for each model, where green indicates higher confidence in correct answers, and red indicates higher confidence in incorrect answers.

Figure 2: LLMs confidence results between correct and incorrect answers.

Models' performances across fields

Significant differences were seen in model performance across all five medical specialties ($p < 0.001$). GPT-4o and Claude 3.5 Sonnet consistently outperformed other models. In Internal Medicine, GPT-4o (70.9%) and Claude 3.5 Sonnet (73.5%)

showed no significant difference ($p = 1.0$) but outperformed lower-tier models like qwen-7b (43.7%, $p < 0.001$). For OBGYN, Claude 3.5 Sonnet (71.0%) significantly outperformed most models, including GPT-4 (54.0%, $p < 0.001$). In Pediatrics, the top five models (GPT-4o, Llama3-70b, Claude 3.5 Sonnet, Claude Opus, GPT-4) showed no significant differences among themselves (all $p > 0.05$) but outperformed lower-tier models. Psychiatry results mirrored this pattern, with GPT-40 (84.4%) and Claude 3.5 Sonnet (82.4%) leading. In Surgery, GPT-40 (70.9%) and Claude 3.5 Sonnet (70.5%) again showed no significant difference ($p = 1.0$) but outperformed lower-performing models like qwen-7b (45.6%, $p < 0.001$) (**Tables 2-3**).

Overall benchmarking results

The highest mean for larger models was in Psychiatry (79.5%), and the lowest in OBGYN (63.4%). For lower models, the highest mean was in Psychiatry (62.3%), and the lowest in Internal Medicine (46.1%).

Across all specialties, larger models consistently outperformed lower models. The performance gap between larger and lower models was most pronounced in Internal Medicine (20.9 percentage points) and least in OBGYN (16.3 percentage points). Psychiatry showed the highest overall performance for both model groups.

Consistency results

In the consistency analysis of model performance, most models showed no significant differences across versions, indicating high consistency. Among the Larger/Newer models, GPT-4, GPT-4o, Gemini 1.5 pro, Llama-3-70b, Claude Opus 3, and Claude Sonnet 3.5 exhibited no significant differences in all fields ($p > 0.05$).

In the Smaller/Older models group, GPT-3.5, Llama Bio 70B, Mistral, Qwen-2-72B, llama8b, and qwen-7B also maintained consistent performance across versions ($p >$

0.05). However, Llama-3-70B in Psychiatry showed significant consistency issues with a p-value of 0.03.

Discussion

This study evaluated the confidence levels and performance of 12 LLMS across five medical specialties, using 1965 validated MCQs. Our findings reveal a nuanced relationship between model confidence and accuracy. The mean confidence difference between correct and incorrect responses was small, ranging from 0.6% to 5.4%.

Higher-tier models generally outperformed lower-tier models, with some showing statistically significant differences ($p<0.01$). However, the performance gap was modest, typically lower than 6% across specialties. Notably, some lower-tier models displayed higher confidence in incorrect answers. These results highlight potential risks in clinical applications, where model overconfidence, regardless of answer correctness, could lead to misinformed decisions.

While higher-tier models showed slightly better calibration between confidence and accuracy, the difference remains insufficient for reliable human interpretation. For instance, GPT-4o displayed the largest confidence gap of 5.4 percentage points between correct and incorrect answers, yet this margin is too narrow for practical clinical decision-making.

Comparing our results to current literature, we see both consistencies and divergences. Katz et al. reported GPT-4 outperforming physicians in psychiatry and performing comparably in general surgery and internal medicine (8). Our study corroborates GPT-4's strong performance, particularly in psychiatry where it achieved 84.4% accuracy. However, our findings suggest more cautious interpretation is needed, given the high confidence levels observed even for incorrect answers. Xiong

et al.'s work on LLM confidence elicitation aligns with our observations of overconfidence (15). They noted improved calibration and failure prediction as model capability increased, which parallels our finding of slightly better confidence calibration in higher-tier models. However, our results emphasize that this improvement is marginal and potentially misleading in clinical contexts.

The implications for clinical practice could be significant and warrant careful consideration. While the performance leap of newer models is promising, their inability to accurately self-assess confidence poses substantial risks in healthcare settings. We propose two primary strategies to address these challenges: developing human-AI collaboration protocols and implementing ensemble methods (23).

Human-AI collaboration protocols may offer a balanced approach to leveraging AI strengths while maintaining necessary human oversight in healthcare. Sezgin et al. emphasize the augmentative role of AI in healthcare, highlighting that AI should complement, rather than replace, healthcare providers (24). They propose a human-in-the-loop (HITL) approach, ensuring that AI systems are guided and supervised by human expertise. This approach could maintain safety and quality in healthcare services while potentially improving service quality and patient outcomes. However, effective implementation faces challenges. Careful design of user interfaces and workflows is crucial to prevent automation bias (24,25). There are also concerns about potential erosion of clinical skills if healthcare professionals become overly reliant on AI assistance (26).

Ensemble methods, which aggregate responses from multiple models, present another promising strategy (27). Mahajan et al. conducted a comprehensive review of ensemble learning techniques in disease prediction (28). Their study found that stacking, an ensemble method that combines multiple classifiers, showed the most

accurate performance in 19 out of 23 cases where it was applied. The voting approach was identified as the second-best ensemble method. These findings suggest that ensemble methods could potentially improve the accuracy and reliability of AI models in healthcare settings (28). However, ensemble methods are computationally intensive and may introduce latency in real-time clinical applications (29). Moreover, their application to medical LLMs remains under-researched. Both strategies would require extensive clinical trials for validation and the development of model-specific calibration curves for each medical specialty.

This study has several limitations. The dataset was limited to 1965 questions across five medical specialties, which may not fully represent the breadth of clinical scenarios. A combination of automatic rephrasing and manual validation could still introduce subtle biases (21). Additionally, the study used default model hyperparameters, potentially limiting performance optimization. The generalizability of results to real-world clinical settings remains uncertain due to the controlled experimental environment. The study did not explore techniques such as fine-tuning and retrieval-augmented generation (RAG), which might improve model accuracy and confidence calibration (30). Finally, the study did not consider the computational cost and time efficiency of deploying these models in practical healthcare scenarios.

Conclusion

Although newer language models show improved performance and consistency in medical knowledge tasks, their confidence levels remain poorly calibrated. This study reveals consistent overconfidence across all models, regardless of answer correctness. The gap between performance and self-assessment poses risks in clinical applications. Human-AI collaboration and ensemble methods offer potential solutions but need validation in healthcare settings.

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):1930–40.
2. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med. 2023 Oct 10;3(1):141.
3. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023 Jul 3;330(1):78–80.
4. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023 Aug 3;620(7972):172–80.
5. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq. 2023 Feb 28;rs.3.rs-2566942.
6. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. BMJ Neurol Open. 2023 Jun 15;5(1):e000451.
7. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Oto-Rhino-Laryngol Off J Eur Fed Oto-Rhino-Laryngol Soc EUFOS Affil Ger Soc Oto-Rhino-Laryngol - Head Neck Surg. 2023 Sep;280(9):4271–8.
8. Katz U, Cohen E, Shachar E, Somer J, Fink A, Morse E, et al. GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. NEJM AI. 2024 Apr 25;1(5):AIdbp2300192.
9. Omar M, Nassar S, Hijaze K, Glicksberg BS, Nadkarni GN, Klang E. Generating Credible Referenced Medical Research: A Comparative Study of Openai's Gpt-4 and Google's Gemini [Internet]. Rochester, NY; 2024 [cited 2024 Apr 22]. Available from: https://papers.ssrn.com/abstract=4780940
10. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care. 2023 Mar 21;27:120.
11. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Int J Inf Fusion. 2022 Jan;77:29–52.
12. Soroush A, Glicksberg BS, Zimlichman E, Barash Y, Freeman R, Charney AW, et al. Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. NEJM AI. 2024 Apr 25;1(5):AIdbp2300040.
13. Schwartz IS, Link KE, Daneshjou R, Cortés-Penfield N. Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation. Clin Infect Dis. 2023 Nov 16;
14. Poon AIF, Sung JJY. Opening the black box of AI-Medicine. J Gastroenterol Hepatol. 2021 Mar;36(3):581–4.
15. Xiong M, Hu Z, Lu X, Li Y, Fu J, He J, et al. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs [Internet]. arXiv; 2024 [cited 2024 Aug 8]. Available from: http://arxiv.org/abs/2306.13063
16. Townsend CM, Beauchamp RD, Evers BM, Mattox KL. Sabiston Textbook of Surgery: The Biological Basis of Modern Surgical Practice. Elsevier Health Sciences; 2016. 2191 p.
17. Loscalzo J. Harrison's principles of internal medicine. No Title [Internet].

[cited 2024 Jul 2]; Available from: https://cir.nii.ac.jp/crid/1130573781693502243
18. Kliegman RM, Behrman RE, Jenson HB, Stanton BMD. Nelson Textbook of Pediatrics E-Book. Elsevier Health Sciences; 2007. 3200 p.
19. Association AP. Diagnostic and Statistical Manual of Mental Disorders. Text Revis [Internet]. 2000 [cited 2024 Jul 2]; Available from: https://cir.nii.ac.jp/crid/1573950399819987840
20. Gabbe SG, Niebyl JR, Simpson JL, Landon MB, Galan HL, Jauniaux ERM, et al. Obstetrics: Normal and Problem Pregnancies E-Book. Elsevier Health Sciences; 2016. 1426 p.
21. Soni S, Roberts K. Paraphrasing to improve the performance of Electronic Health Records Question Answering. AMIA Summits Transl Sci Proc. 2020 May 30;2020:626–35.
22. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report [Internet]. arXiv; 2024 [cited 2024 Aug 10]. Available from: http://arxiv.org/abs/2303.08774
23. Longhurst CA, Singh K, Chopra A, Atreja A, Brownstein JS. A Call for Artificial Intelligence Implementation Science Centers to Evaluate Clinical Effectiveness. NEJM AI. 2024 Jul 25;1(8):AIp2400223.
24. Sezgin E. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. Digit Health. 2023 Jul 2;9:20552076231186520.
25. Straw I. The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future. Artif Intell Med. 2020 Nov;110:101965.
26. Čartolovni A, Malešević A, Poslon L. Critical analysis of the AI impact on the patient–physician relationship: A multi-stakeholder qualitative study. Digit Health. 2023 Dec 19;9:20552076231220833.
27. Yang H, Li M, Zhou H, Xiao Y, Fang Q, Zhang R. One LLM is not Enough: Harnessing the Power of Ensemble Learning for Medical Question Answering. medRxiv. 2023 Dec 24;2023.12.21.23300380.
28. Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble Learning for Disease Prediction: A Review. Healthcare. 2023 Jun 20;11(12):1808.
29. Edeh MO, Dalal S, Dhaou IB, Agubosim CC, Umoke CC, Richard-Nnabu NE, et al. Artificial Intelligence-Based Ensemble Learning Model for Prediction of Hepatitis C Disease. Front Public Health. 2022 Apr 27;10:892371.
30. Glicksberg BS, Timsina P, Patel D, Sawant A, Vaid A, Raut G, et al. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. J Am Med Inform Assoc JAMIA. 2024 May 21;ocae103.

Table 1: Confidence means of LLMs between correct and incorrect answers.

Model	*Confidence When Incorrect (mean, SD)*	*Confidence When Correct (mean, SD)*	*p-value*
GPT-4o	58.99 ± 14.31	64.38 ± 16.11	<0.01
Llama-70B	53.59 ± 22.38	59.50 ± 23.54	<0.01
Claude 3.5 Sonnet	67.37 ± 9.08	70.52 ± 11.07	<0.01
Gemini	85.55 ± 16.23	87.17 ± 16.58	0.35
Claude 3 Opus	67.32 ± 13.06	68.90 ± 15.65	0.61
GPT-4	83.34 ± 23.30	84.52 ± 22.43	0.07
Qwen2-72B	56.49 ± 18.55	58.59 ± 20.03	<0.01
Qwen2-7B	76.37 ± 17.11	74.45 ± 20.30	0.01
Mixtral 8x7B	82.99 ± 16.52	85.49 ± 14.62	0.04
Llama-8B	80.25 ± 17.40	79.67 ± 21.59	0.31
Llama OpenBio	78.14 ± 27.59	77.73 ± 28.78	0.83
GPT-3.5	82.85 ± 27.17	81.63 ± 28.66	0.81

Table 2: Individual larger\newer models benchmarking results.

Specialty	*Model*	*GPT-4.0*	*Llama3-70b*	*Claude 3.5 Sonnet*	*Gemini 1.5 Pro*	*Claude 3 Opus*	*GPT-4*
Surgery	Overall	70.92	61.47	70.45	60.76	67.14	69.50
	Original	70.92	55.32	66.67	58.87	65.96	68.79
	Rephrase 1	68.79	65.25	68.79	59.57	67.38	68.79
	Rephrase 2	73.05	63.83	75.89	63.83	68.09	70.92
Internal Medicine	Overall	75.13	58.47	73.54	58.99	70.45	64.02
	Original	73.81	57.14	71.43	56.35	66.67	61.11
	Rephrase 1	74.60	59.52	73.81	63.49	68.79	66.67
	Rephrase 2	76.98	58.73	75.40	57.14	75.89	64.29

Obstetrics and Gynecology	Overall	65.71	55.64	70.98	58.75	65.47	53.96
	Original	69.06	53.96	72.66	57.55	68.35	56.12
	Rephrase 1	62.59	56.12	71.22	59.71	68.35	53.96
	Rephrase 2	65.47	56.83	69.06	58.99	59.71	51.80
Pediatrics	Overall	71.72	68.35	71.38	57.58	68.01	67.00
	Original	68.69	64.65	70.71	54.55	67.68	64.65
	Rephrase 1	72.73	70.71	70.71	58.59	71.72	66.67
	Rephrase 2	73.74	69.70	72.73	59.60	64.65	69.70
Psychiatry	Overall	84.44	73.56	82.44	NA	79.78	77.33
	Original	77.33	67.33	83.33	NA	77.33	78.00
	Rephrase 1	88.00	76.00	81.33	NA	81.33	77.33
	Rephrase 2	88.00	77.33	82.67	NA	80.67	76.67

Table 3: Individual smaller\older models benchmarking results.

Specialty	***Model***	***Qwen-2-72b***	***Qwen-7b***	***Mistral***	***Llama 8b***	***Llama Bio***	***GPT-3.5***
Surgery	Overall	57.45	45.63	46.81	47.52	56.03	53.43
	Original	53.90	46.10	45.39	46.81	58.87	52.48
	Rephrase 1	61.70	47.52	46.10	48.23	54.61	53.90
	Rephrase 2	56.74	43.26	48.94	47.52	54.61	53.90
Internal Medicine	Overall	47.88	43.65	44.97	40.48	53.70	45.77
	Original	46.03	45.24	46.83	40.48	55.56	46.83
	Rephrase 1	49.21	42.86	43.65	40.48	52.38	43.65
	Rephrase 2	48.41	42.86	44.44	40.48	53.17	46.83
Obstetrics and Gynecology	Overall	55.16	43.65	44.36	43.17	53.72	42.93
	Original	56.83	43.17	43.88	44.60	56.83	46.04
	Rephrase 1	55.40	43.88	46.04	39.57	50.36	41.01
	Rephrase 2	53.24	43.88	43.17	45.32	53.96	41.73

Pediatrics	Overall	59.60	39.73	52.19	48.48	61.28	49.16
	Original	52.53	36.36	47.47	44.44	60.61	49.49
	Rephrase 1	61.62	45.45	53.54	51.52	62.63	46.46
	Rephrase 2	64.65	37.37	55.56	49.49	60.61	51.52
Psychiatry	Overall	67.33	54.22	63.33	60.89	70.44	57.56
	Original	67.33	54.00	62.00	60.00	67.33	62.67
	Rephrase 1	65.33	53.33	62.67	60.00	72.00	51.33
	Rephrase 2	69.33	55.33	65.33	62.67	72.00	58.67