Comprehensive Performance Evaluation of Vera: A Multi-Benchmark Assessment Across Medical Knowledge Domains

Vera Health

August 2025

Abstract

We present a comprehensive evaluation of Vera, an advanced clinical decision-support system designed to empower healthcare providers with instant, evidence-based medical guidance. Vera leverages sophisticated AI agents and Retrieval-Augmented Generation technology, synthesising knowledge from over 60 million peer-reviewed medical publications to deliver reliable, contextually-appropriate answers. This multi-benchmark assessment evaluates Vera's performance across three distinct medical knowledge domains: the United States Medical Licensing Examination (USMLE), the New England Journal of Medicine AI Q&A dataset (NEJM-AI), and the MedXpertQA benchmark. On USMLE, Vera achieved an exceptional overall accuracy of **97.5 %**, with step-specific accuracies of **97.9 % (Step 1)**, **98.2 % (Step 2 CK)** and **96.7 % (Step 3)**. On the NEJM-AI benchmark comprising 655 questions across five medical specialties, Vera demonstrated superior performance with **84.9 % accuracy**, outperforming leading AI models including OpenAI o4 Mini (77.1 %), Claude 4 Sonnet (75.4 %), and Perplexity Sonar Pro (74.4 %). On the MedXpertQA benchmark comprising 500 questions across multiple body systems and medical tasks, Vera achieved **62.2 % accuracy**, demonstrating strong performance in specialized clinical reasoning scenarios. Vera achieved the highest accuracy in four out of five NEJM-AI medical specialties, with particularly strong performance in Pediatrics (93.9 %) and Internal Medicine (87.3 %). These results across diverse evaluation frameworks underscore Vera's robust medical knowledge representation and reasoning capabilities, positioning it as a leading solution for clinical decision support.

1 Introduction

Healthcare providers across diverse clinical environments require rapid access to accurate, evidence-based medical knowledge to support optimal patient care. The exponential growth of medical literature presents unprecedented challenges for timely knowledge retrieval and synthesis. Vera addresses this critical need by combining sophisticated AI agents with advanced Retrieval-Augmented Generation technology, delivering reliable clinical guidance approximately ten times faster than conventional methods.

The evaluation of medical AI systems requires rigorous assessment across multiple domains to ensure robust performance in real-world clinical scenarios. While individual benchmarks provide valuable insights, comprehensive evaluation across diverse knowledge frameworks offers a more complete picture of system capabilities and limitations. This study presents a multi-benchmark evaluation of Vera using three complementary assessment frameworks: the United States Medical Licensing Examination (USMLE), the New England Journal of Medicine AI Q&A dataset (NEJM-AI), and the MedXpertQA benchmark.

USMLE provides a standardized measure of foundational medical knowledge across basic science, clinical knowledge, and patient management domains. However, it primarily reflects

pre-licensure educational content and may not fully capture the complexity of contemporary clinical decision-making. To address this limitation, we complement our evaluation with the NEJM-AI benchmark, which presents 655 clinically-oriented questions across five major medical specialties, offering insights into performance on more practice-relevant scenarios. Additionally, we evaluate Vera on the MedXpertQA benchmark, comprising 500 questions that assess clinical reasoning across diverse body systems, medical tasks, and question types, providing further insights into specialized clinical knowledge domains.

Our comprehensive analysis across these distinct evaluation frameworks reveals Vera's strengths and performance characteristics, demonstrating substantial promise for transforming clinical decision support, enhancing provider efficiency, and ultimately improving patient care quality.

2 Results

2.1 Multi-Benchmark Performance Overview

Vera demonstrated exceptional performance across all three evaluation frameworks, achieving **97.5%** on USMLE, **84.9%** on the NEJM-AI benchmark, and **62.2%** on the MedXpertQA benchmark. Table 1 summarizes Vera's performance across all assessments.

Benchmark	**Accuracy**
USMLE (Overall)	97.5 %
Step 1	97.9 %
Step 2 CK	98.2 %
Step 3	96.7 %
NEJM-AI (Overall)	84.9 %
MedXpertQA (Overall)	62.2 %

Table 1: Summary of Vera performance across medical knowledge benchmarks.

2.2 USMLE Performance Analysis

On the USMLE assessment, Vera achieved near-perfect accuracy across all examination levels, demonstrating robust foundational medical knowledge. The minimal variation between steps (range: 96.7–98.2 %) indicates that Vera's knowledge representation scales effectively from basic science concepts to complex clinical scenarios requiring patient management decisions.

2.3 USMLE Competitive Analysis

Vera's performance establishes clear superiority over other medical AI systems in standardized medical knowledge assessment. Figure 1 demonstrates Vera's competitive advantage across the medical AI landscape.

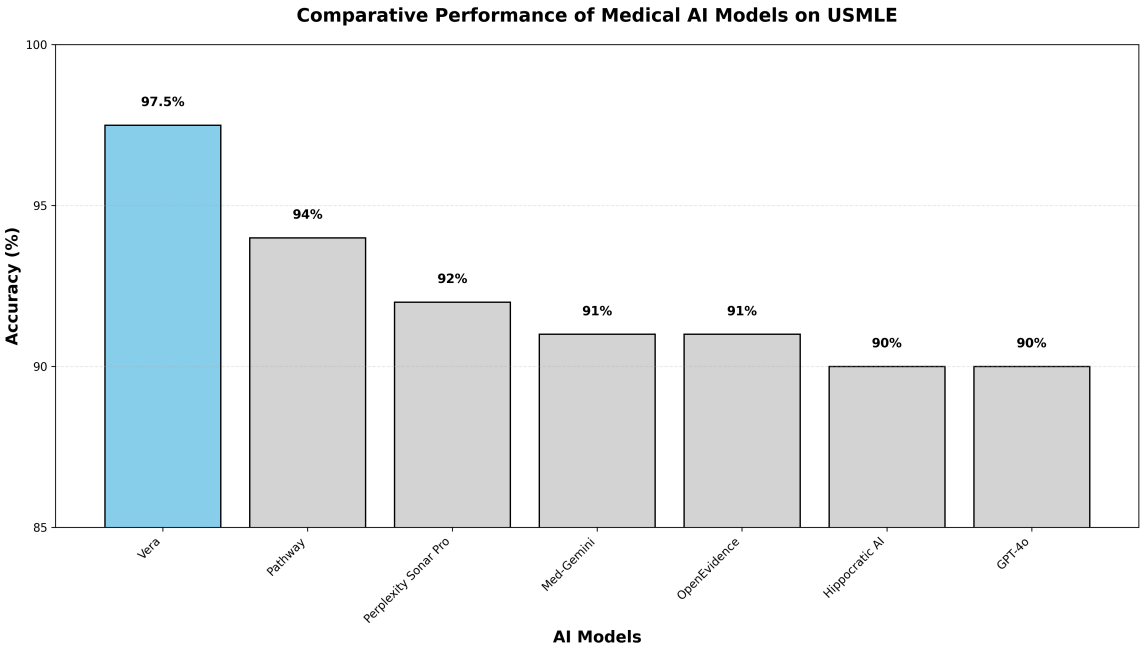

Figure 1: Comparative performance of medical AI models on USMLE assessment. Vera achieved 97.5 % accuracy, significantly outperforming specialized medical models including Pathway (94 %), Perplexity Sonar Pro (92 %), and general-purpose models such as GPT-4o (90 %).

This competitive analysis reveals several key insights: (1) Vera's 3.5 percentage point lead over the second-best performing model represents substantial improvement in medical knowledge assessment; (2) the performance gap widens significantly compared to general-purpose models, highlighting the value of medical-specific optimization; and (3) Vera's superiority spans both specialized medical AI systems and leading general-purpose language models.

2.4 NEJM-AI Competitive Benchmark Results

On the NEJM-AI benchmark, Vera achieved the highest overall accuracy among all evaluated models, outperforming leading AI systems by substantial margins. Figure 2 demonstrates Vera's competitive superiority.

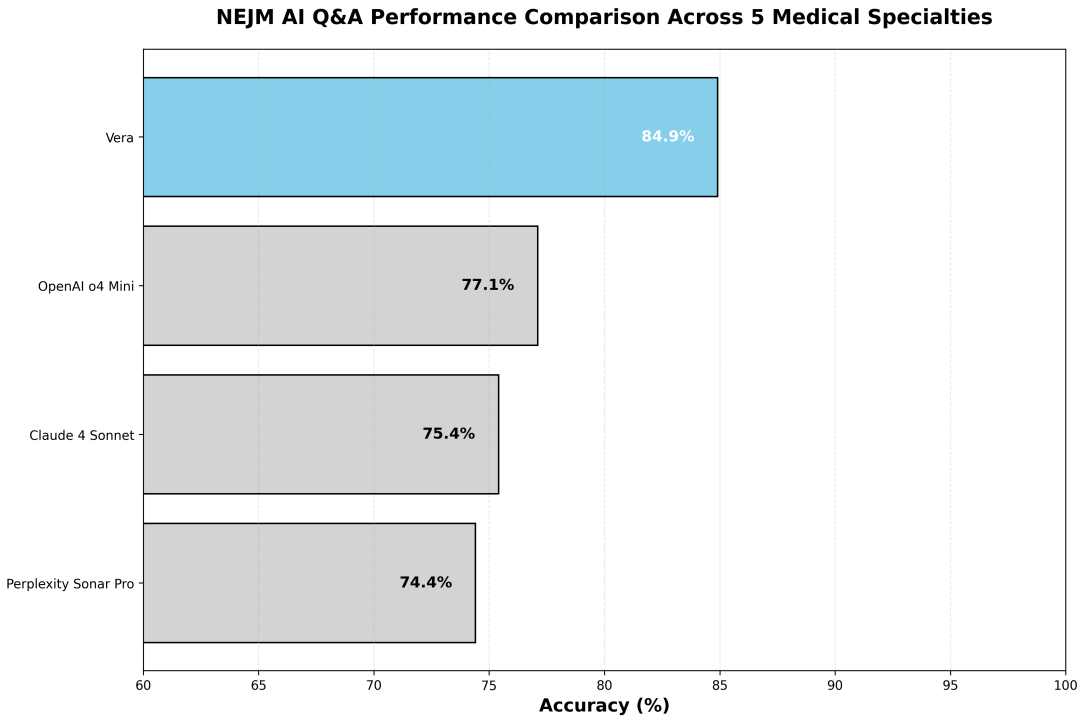

Figure 2: Comparative performance on NEJM-AI benchmark. Vera achieved 84.9 % accuracy, surpassing OpenAI o4 Mini (77.1 %), Claude 4 Sonnet (75.4 %), and Perplexity Sonar Pro (74.4 %).

2.5 Specialty-Specific Performance Analysis

Vera's performance varied across medical specialties, with consistently strong results in most domains. Table 2 presents detailed specialty-specific accuracies.

Medical Specialty	Questions	Vera Accuracy
Pediatrics	99	93.9 %
Psychiatry	150	88.7 %
Internal Medicine	126	87.3 %
General Surgery	141	83.0 %
OBGYN	139	74.1 %

Table 2: Vera performance by medical specialty on NEJM-AI benchmark.

Figure 3 provides a detailed comparison of Vera's performance against competing models across all five medical specialties.

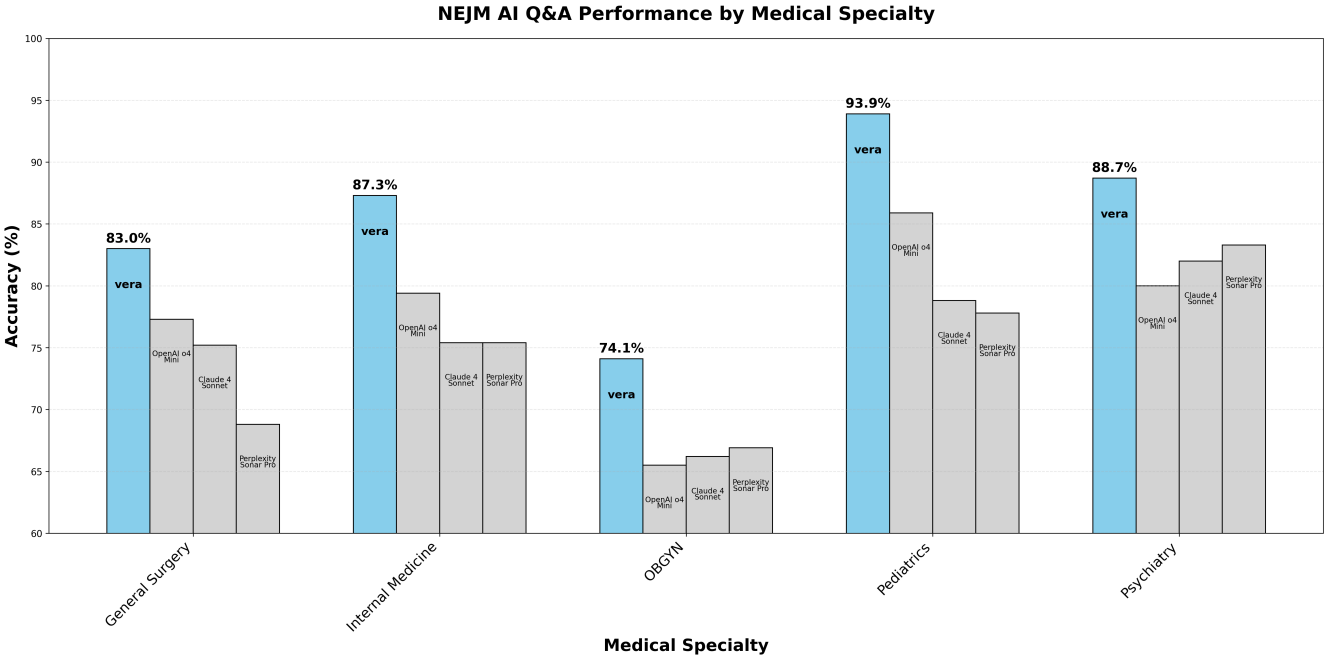

Figure 3: Specialty-specific performance comparison on NEJM-AI benchmark. Vera achieved the highest accuracy in four out of five specialties, demonstrating particular strength in Pediatrics and Internal Medicine.

Vera achieved the highest accuracy in four out of five specialties:

- **Pediatrics**: Leading performance with 93.9 % accuracy
- **Internal Medicine**: Strong performance with 87.3 % accuracy
- **General Surgery**: Competitive advantage with 83.0 % accuracy
- **OBGYN**: Modest lead with 74.1 % accuracy
- **Psychiatry**: Strong performance with 88.7 % accuracy

2.6 MedXpertQA Performance Analysis

On the MedXpertQA benchmark, Vera achieved 62.2 % accuracy across 500 diverse medical questions, demonstrating competent performance in specialized clinical reasoning scenarios. Table 3 presents detailed performance breakdowns across different categories.

Category	Questions	Vera Accuracy
By Body System		
Integumentary	16	81.2 %
Skeletal	81	72.8 %
Muscular	36	72.2 %
Reproductive	31	71.0 %
Digestive	60	63.3 %
Endocrine	37	62.2 %
Lymphatic	22	59.1 %
Nervous	72	56.9 %
Respiratory	32	56.2 %
Urinary	18	55.6 %
Cardiovascular	68	51.5 %
Other/NA	27	48.1 %
By Medical Task		
Basic Science	139	66.9 %
Treatment	157	61.8 %
Diagnosis	204	59.3 %
By Question Type		
Understanding	115	66.1 %
Reasoning	385	61.0 %

Table 3: Vera performance by category on MedXpertQA benchmark.

The MedXpertQA results reveal several notable patterns in Vera's performance:

- **Body System Variation**: Performance ranged from 81.2 % (Integumentary) to 48.1 % (Other/NA), with strongest performance in anatomically discrete systems
- **Medical Task Performance**: Basic Science questions (66.9 %) outperformed clinical applications, suggesting stronger performance on foundational knowledge
- **Question Type Analysis**: Understanding questions (66.1 %) showed superior performance compared to Reasoning questions (61.0 %), indicating effective knowledge retrieval capabilities

2.7 Comparative Model Performance on MedXpertQA

Table 4 presents a comparative analysis of Vera's performance against other leading AI models on the MedXpertQA benchmark, highlighting Vera's competitive positioning in specialized clinical reasoning tasks.

Model	Reasoning	Understanding	Average
Vera	**61.0 %**	**66.1 %**	**62.2 %**
OpenAI o3 Mini	37.6 %	36.2 %	37.3 %
Claude 3.5 Sonnet	19.9 %	25.8 %	21.3 %
Gemini 1.5 Pro	19.2 %	21.2 %	19.7 %

Table 4: Comparative performance of AI models on MedXpertQA benchmark. Vera demonstrates substantial superiority across both reasoning and understanding question types.

3 Methods

3.1 Evaluation Framework

We conducted a comprehensive multi-benchmark evaluation using three distinct medical knowledge assessment frameworks: the United States Medical Licensing Examination (USMLE), the New England Journal of Medicine AI Q&A dataset (NEJM-AI), and the MedXpertQA benchmark. This tri-benchmark approach enables assessment of foundational medical knowledge, contemporary clinical reasoning capabilities, and specialized clinical domain expertise.

3.2 USMLE Assessment

We sampled multiple-choice questions from official USMLE preparation resources spanning all three examination steps: Step 1 (basic science), Step 2 Clinical Knowledge (clinical knowledge and skills), and Step 3 (patient management). Each question comprised a clinical vignette, multiple answer options, reference answer key, and specialty classification. Questions were presented to Vera exactly as written, using the production system prompt without benchmark-specific optimization.

3.3 NEJM-AI Benchmark Evaluation

The NEJM-AI dataset (Katz et al., 2024) consists of 655 clinically-oriented multiple-choice questions distributed across five major medical specialties: General Surgery (141 questions), Internal Medicine (126 questions), OBGYN (139 questions), Pediatrics (99 questions), and Psychiatry (150 questions). This benchmark was designed to assess contemporary clinical knowledge and reasoning capabilities relevant to practicing physicians. The original study reported GPT-4 achieving 74.7% accuracy on this benchmark.

3.4 MedXpertQA Benchmark Evaluation

The MedXpertQA dataset (Zuo et al., 2025) is a highly challenging benchmark designed to evaluate expert-level medical reasoning and understanding. Comprising 4,460 questions spanning 17 medical specialties and 11 body systems, MedXpertQA represents one of the most comprehensive and difficult medical reasoning assessments available. The benchmark includes two subsets: MedXpertQA Text for text-based medical evaluation and MedXpertQA MM for multimodal medical evaluation.

For our evaluation, we utilized a representative sample of 500 questions from the MedXpertQA Text subset, maintaining the benchmark's rigorous standards while enabling efficient assessment. Questions are categorized by body system (12 categories), medical task (Basic Science, Diagnosis, Treatment), and question type (Understanding, Reasoning). This benchmark assesses specialized clinical knowledge and reasoning capabilities across a broad spectrum of medical scenarios, from foundational science to complex clinical applications, making it particularly valuable for evaluating advanced medical AI systems.

3.5 Experimental Protocol

For all three benchmarks, we maintained consistent evaluation protocols:

- All questions were presented to Vera using the standard production system prompt without any benchmark-specific prompt engineering
- The optional *Deep Dive* mode was disabled to mirror the fast-response mode preferred by clinicians in real-world settings

- Each question was processed independently without prior context or question-specific optimization
- Response accuracy was determined by exact matching against provided reference answers

3.6 Competitive Analysis

For the NEJM-AI benchmark, we compared Vera's performance against three leading medical AI systems: OpenAI o4 Mini, Claude 4 Sonnet, and Perplexity Sonar Pro. As the latest models from OpenAI, Anthropic, and Perplexity are not publicly available, we conducted internal evaluations using our own implementations. All models were evaluated on the identical 655-question set using their respective optimal configurations. While the original NEJM-AI study reported GPT-4 achieving 74.7% accuracy, we excluded it from our comparative analysis as OpenAI o4 Mini demonstrated superior performance.

3.7 Statistical Analysis

We computed overall accuracy rates, specialty-specific performance metrics, and comparative rankings. Performance variations across specialties were analyzed to identify domain-specific strengths and areas for improvement.

4 Discussion

4.1 Benchmark Complementarity and Clinical Implications

The tri-benchmark evaluation reveals distinct but complementary insights into Vera's capabilities. The exceptional USMLE performance (97.5 % accuracy) demonstrates mastery of foundational medical knowledge across basic science, clinical knowledge, and patient management domains. The strong NEJM-AI performance (84.9 % accuracy) with competitive superiority over leading AI models indicates robust capabilities in contemporary clinical reasoning scenarios. The MedXpertQA performance (62.2 % accuracy) provides insights into specialized clinical domain expertise and reasoning across diverse body systems and medical tasks.

The performance differential between benchmarks (97.5 % vs 84.9 % vs 62.2 %) likely reflects the distinct nature and complexity of these assessments. USMLE questions primarily evaluate standardized medical knowledge with established answer keys, while NEJM-AI questions present more nuanced clinical scenarios that may admit multiple reasonable approaches. MedXpertQA represents the most challenging assessment, featuring complex clinical reasoning scenarios that require integration of specialized knowledge across multiple domains, making it a rigorous test of advanced clinical competency.

4.2 Competitive Positioning

Vera's performance on the NEJM-AI benchmark establishes clear competitive advantages over current medical AI systems. The substantial lead over competing models represents a significant improvement in a highly competitive field. More significantly, Vera's consistent superiority across four of five medical specialties demonstrates broad-based clinical knowledge rather than domain-specific optimization.

The specialty-specific results reveal important insights:

- **Pediatrics**: The exceptional 93.9 % accuracy suggests strong performance in a domain requiring specialized developmental and age-specific considerations
- **Internal Medicine**: The 87.3 % accuracy demonstrates competence in the broad-based reasoning required for this foundational specialty

- **OBGYN**: The comparatively lower 74.1 % accuracy, while still leading competitors, indicates potential areas for targeted improvement

4.3 System Generalization and Robustness

The consistent high performance across diverse evaluation frameworks suggests that Vera's knowledge representation and reasoning mechanisms generalize effectively across different question formats, difficulty levels, and clinical contexts. This robustness is particularly important for clinical deployment, where the system must handle diverse query types and clinical scenarios.

4.4 Limitations and Considerations

Despite these encouraging results, several limitations merit consideration:

1. **Benchmark Scope**: Both evaluations rely on multiple-choice formats that may not fully capture the complexity of real-world clinical decision-making, which often involves uncertainty, incomplete information, and multifaceted patient presentations.

2. **Clinical vs Academic Knowledge**: High performance on academic benchmarks does not guarantee optimal real-world clinical effectiveness. Vera's design prioritizes contemporary clinical guidelines and evidence-based practice, which may occasionally diverge from historical examination answer keys.

3. **Specialty Variation**: The observed performance variation across medical specialties suggests that certain domains may benefit from targeted enhancement, particularly OBGYN where performance, while competitive, showed the largest room for improvement.

4. **Temporal Considerations**: Medical knowledge evolves rapidly with new research findings and guideline updates. Continuous evaluation and model updating will be essential to maintain performance over time.

5. **Evaluation Methodology**: Both benchmarks rely on predetermined answer keys that may not always reflect the full spectrum of clinically acceptable responses, potentially underestimating system performance in ambiguous scenarios.

5 Conclusions

This comprehensive multi-benchmark evaluation demonstrates Vera's exceptional capabilities across diverse medical knowledge domains. The system achieved near-perfect accuracy on USMLE (97.5 %), established competitive superiority on the NEJM-AI benchmark (84.9 %), and demonstrated competent performance on the challenging MedXpertQA benchmark (62.2 %). On NEJM-AI, Vera outperformed leading AI models including OpenAI o4 Mini, Claude 4 Sonnet, and Perplexity Sonar Pro.

Key findings include:

- **Broad Medical Competence**: Consistent high performance across foundational (USMLE), contemporary clinical (NEJM-AI), and specialized reasoning (MedXpertQA) knowledge domains

- **Competitive Advantage**: Clear superiority over current medical AI systems in head-to-head evaluation

- **Specialty Robustness**: Leading performance in four of five NEJM-AI medical specialties with particularly strong results in Pediatrics and Internal Medicine

- **Domain-Specific Expertise**: Strong performance across diverse body systems in MedXpertQA, with particular strength in anatomically discrete systems (Integumentary: 81.2 %, Skeletal: 72.8 %)
- **Knowledge Generalization**: Effective performance across diverse question formats, difficulty levels, and clinical contexts

These results position Vera as a leading solution for clinical decision support, with demonstrated capabilities that exceed current benchmarks for medical AI systems. The tri-benchmark approach provides robust evidence of system performance across academic, clinically-relevant, and specialized reasoning scenarios, supporting deployment in medical education, clinical training, and point-of-care decision support applications.

Data Availability

The evaluation datasets and detailed results are available on request (enterprise@vera-health.ai) and will be provided subject to standard data-use agreements and privacy safeguards.

References

[1] Katz, U., Cohen, E., Shachar, E., Somer, J., Fink, A., Morse, E., Shreiber, B., & Wolf, I. (2024). GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. *NEJM AI*, 1(5), AIdbp2300192. https://doi.org/10.1056/AIdbp2300192

[2] Zuo, Y., Qu, S., Li, Y., Chen, Z., Zhu, X., Hua, E., Zhang, K., Ding, N., & Zhou, B. (2025). MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

[3] Bicknell, B. T., Butler, D., Whalen, S., Ricks, J., Dixon, C. J., Clark, A. B., Spaedy, O., Skelton, A., Edupuganti, N., Dzubinski, L., Tate, H., Dyess, G., Lindeman, B., & Lehmann, L. S. (2024). ChatGPT-4 Omni Performance in USMLE Disciplines and Clinical Skills: Comparative Analysis. *JMIR medical education*, 10, e63430. https://doi.org/10.2196/63430