American Journal of Biomedical Science & Research

www.biomedgrid.com

ISSN: 2642-1747

Review Article

Problems of Implementing Large Language Models in Medicine

ON Andreeva[1], AV Domashev[2], EA Evlanova[3], AA Ryazanova[4] and A Yu Shcherbakov[5]*

[1]*Ph.D of Medical Sciences, doctor of telemedicine consultations, Head of the Gynecological Office Doctor 2000 LLC (Partner of the network of medical centers "MedSwiss"), Russia*

[2]*Head of the Engineering Center of the Academic Institute of Virtual and Augmented Reality of the Russian State Social University, Russia*

[3]*Analyst of the Scientific and Educational Center of Social Analytics of the Russian State Social University, Russia*

[4]*Leading Specialist of the Scientific and Educational Center of Social Analytics of the Russian State Social University, Russia*

[5]*Grand Ph. D, Professor of Technical Sciences, Head of the Department of Cognitive-Analytical and Neuro-Applied Technologies of the Russian State Social University, Leading Researcher of the State University of Management, Russia*

***Corresponding author:** A Yu Shcherbakov, Grand Ph.D, Professor of Technical Sciences, Head of the Department of Cognitive-Analytical and Neuro-Applied Technologies of the Russian State Social University, Leading Researcher of the State University of Management, Russia.

To Cite This article: *ON Andreeva, AV Domashev, EA Evlanova, AA Ryazanova and A Yu Shcherbakov*, Problems of Implementing Large Language Models in Medicine. Am J Biomed Sci & Res. 2026 30(2) AJBSR.MS.ID.003897,* **DOI:** 10.34297/AJBSR.2026.30.003897

Received: February 05, 2026; **Published:** February 19, 2026

Abstract

This article analyzes the readiness of medical artificial intelligence systems based on large multimodal language models for practical application, using international publications. Classic testing using standardized benchmarks proves insufficiently accurate, while the use of extended stress tests allows for the identification of diagnostic errors. It is shown that different test datasets actually test different qualities, and a methodology for their thematic stratification is presented. It is noted that average system ratings can create the illusion of progress and mislead about the models' readiness for real-world use. An idea is proposed for incorporating the patient's medical record or medical history to combat hallucinations in diagnosis. The main conclusion of the article is that medical benchmarks cannot be used as a direct indicator of AI readiness for clinical implementation.

Keywords: Diagnostics, Language model, Benchmark, Stress test, Medical artificial intelligence

Introduction

By 2030, the global AI market in healthcare is projected to exceed 50 billion dollars [1]. Since 2018, annual investments have reached 11.45 billion dollars, and by 2020, the market volume reached 2.2 billion dollars [2]. Interest in medical AI has grown significantly in recent years, with major developments in China including platforms like Ping An Good Doctor and other telemedicine systems serving over 60 million users [2]. However, the implementation of AI systems in clinical practice faces significant challenges. A critical question is whether current benchmarks accurately reflect the readiness of Large Language Models (LLMs) for real-world medical applications. This article examines this question through comprehensive stress testing of leading AI models.

Medical Benchmarks and Their Limitations

Benchmarks are standardized test datasets used to evaluate AI system performance. In medical AI, several key benchmarks have emerged:

a) NEJM (New England Journal of Medicine Image Challenge) – clinical case diagnostics [3]

b) JAMA (Journal of the American Medical Association Clinical Challenge) – complex clinical scenarios [4]

c) VQA-RAD (Radiology VQA dataset) – radiology image analysis [5]

d) PathVQA (Pathology VQA dataset) – pathology diagnostics

e) SLAKE (Structured medical VQA) – structured medical questions

f) OmniMedVQA – comprehensive multimodal medical questions [6]

According to ISO/IEC TR 19759:2015, benchmarks should provide objective assessment of system capabilities. However, recent research suggests that standard benchmarks may overestimate AI readiness for clinical deployment [7].

The Illusion of Progress: Analysis of Leading Models

Recent stress testing by Microsoft Research revealed significant gaps between benchmark performance and real-world capability [7]. Leading models including GPT-5 [8], Gemini-2.5 Pro [9], OpenAI-o3[10], OpenAI-o4-mini [10], GPT-4o [11], and DeepSeek-VL2 [12] were evaluated on both standard and challenging subsets of NEJM and JAMA datasets.

Performance on NEJM Dataset

Table 1: Performance comparison on NEJM easy vs. hard cases.

Model	Easy Cases (%)	Hard Cases (%)	Gap (%)
GPT-5	80.89	67.56	13.33
Gemini-2.5 Pro	79.95	65.01	14.94
OpenAI-o3	80.89	67.03	13.86
GPT-4o	66.90	37.28	29.62
OpenAI-o4-mini	75.91	66.49	9.42
DeepSeek-VL2	33.16	25.30	7.86

Performance on JAMA Dataset

Table 2: Performance comparison on JAMA easy vs. hard cases.

Model	Easy Cases (%)	Hard Cases (%)	Gap (%)
GPT-5	86.59	82.91	3.68
Gemini-2.5 Pro	84.84	74.93	9.91
OpenAI-o3	84.75	82.65	2.10
OpenAI-o4-mini	80.50	78.40	2.10
GPT-4o	69.90	68.90	1.00
DeepSeek-VL2	38.20	32.60	5.60

The data reveals that while models perform well on easy cases, performance drops significantly on challenging cases. Notably, NEJM shows larger performance gaps than JAMA, suggesting different complexity characteristics between datasets [7].

Stress Testing Methodology

To better understand model limitations, researchers conducted stratified testing on 175 NEJM cases, categorizing them by difficulty and specialty. The following subsections detail key findings.

Stress Test 1: Difficulty Stratification

Cases were divided into categories based on diagnostic complexity: (Table 3)

Table 3: Performance on standard vs. hard diagnostic cases from 175 NEJM samples.

Model	Standard Accuracy (%)	Hard Cases (%)
GPT-5	66.28	37.71
Gemini-2.5 Pro	67.42	37.14
OpenAI-o3	61.71	37.71
OpenAI-o4-mini	65.14	33.71
GPT-4o	45.71	3.40

GPT-4o showed catastrophic failure on hard cases (3.4% accuracy), while even top-performing models achieved only ~37% accuracy on challenging diagnostics [7].

Stress Test 2: Visual Dependency Analysis

Twenty cases requiring critical visual analysis were selected. Results showed:

a) GPT-4o achieved only 20% accuracy on visually-dependent cases

b) Models often provided confident but incorrect diagnoses

c) Visual reasoning remains a critical weakness

Stress Test 3: Multi-Image Reasoning

Cases requiring comparison of multiple images (e.g., before/after treatment, multiple views) revealed severe limitations: (Table 4)

Table 4: Performance comparison: single vs. multiple images.

Model	Single Image (%)	Multi-Image (%)
GPT-5	66.28	70.85
Gemini-2.5 Pro	67.42	70.28
OpenAI-o3	61.71	64.00

Interestingly, some models showed slight improvement with multiple images, suggesting potential for multi-view diagnostic enhancement [7].

Stress Test 4: Specialty-Specific Performance

Performance varied dramatically across medical specialties. GPT-5 results:

a) Category 1 specialties: 90.9% accuracy

b) Category 4 specialties: 20.0% accuracy

c) Overall standard cases: 66.3%

d) Category 4 hard cases: 5.25% accuracy

This reveals that averaged metrics mask critical specialty-specific failures [7].

Stress Test 5: Rare Disease Detection

Forty NEJM cases featuring rare diseases were analyzed: (Table 5)

Table 5: Common vs. rare disease diagnostic accuracy.

Model	Common Diseases (%)	Rare Diseases (%)	Gap (%)
GPT-5	83.3	51.7	31.6
Gemini-2.5 Pro	80.8	47.5	33.3
OpenAI-o3	76.7	52.5	24.2
OpenAI-o4-mini	71.7	37.5	34.2
GPT-4o	36.7	41.7	-5.0

All models except GPT-4o showed significant performance degradation on rare diseases, with gaps ranging from 24-34 percentage points [7].

Chain-of-Thought Prompting Evaluation

Three experiments tested Chain-of-Thought (CoT) prompting effectiveness [13]:

Experiment 1: 100 NEJM and VQA-RAD cases tested with CoT prompting showed minimal improvement over standard prompting [5].

Experiment 2: Visual reasoning tasks showed no significant benefit from CoT approaches.

Experiment 3: OmniMedVQA testing with OpenAI-o3 revealed that CoT increased computational cost without proportional accuracy gains [6].

OpenAI-o4-mini consumed 25+ tokens per response with CoT but showed only marginal improvements, suggesting diminishing returns for medical diagnostics [7].

Thematic Stratification of Benchmarks

Analysis of six major benchmarks revealed that each test fundamentally different capabilities: (Table 6)

Key finding: High performance on one benchmark does not predict performance on others. NEJM tests clinical reasoning depth, JAMA emphasizes differential diagnosis, while VQA-RAD focuses on visual interpretation [3,4,7].

Table 6: Thematic stratification of medical benchmarks.

Benchmark	Primary Focus
NEJM	Clinical reasoning with complex cases
JAMA	Differential diagnosis scenarios
VQA-RAD	Radiology image interpretation
PathVQA	Histopathology analysis
SLAKE	Structured medical knowledge
OmniMedVQA	Multimodal integration

This stratification reveals that:

a) Averaged benchmark scores mask critical capability gaps

b) Different benchmarks test orthogonal competencies

c) No single benchmark adequately predicts clinical readiness

Hallucination Mitigation: The Role of Patient History

A critical vulnerability of current LLMs is diagnostic hallucination – generating confident but incorrect diagnoses. Research suggests incorporating patient medical records and history as context could significantly reduce hallucinations [7,14].

Proposed Approach

1) Integrate complete patient history as input context

2) Include prior diagnostic results and treatment responses

3) Provide temporal disease progression data

4) Enable model access to patient-specific risk factors

Initial experiments suggest this approach could improve accuracy by 15-20% on complex cases, though comprehensive validation is needed [14].

Distribution Shifts and Real-World Deployment

Benchmark datasets may not reflect real clinical populations. Distribution shifts between test data and actual patient populations create additional challenges [15]:

a) Geographic variation in disease prevalence

b) Demographic differences in symptom presentation

c) Equipment and imaging protocol variations

d) Temporal changes in diagnostic criteria

Models trained on benchmark datasets may fail when encountering these real-world variations, even if benchmark performance is excellent [15].

Regulatory and Safety Considerations

Clinical AI deployment requires addressing multiple safety dimensions:

a) Diagnostic Safety: Minimizing false negatives in critical conditions

b) Transparency: Explainable reasoning for clinical decisions

c) Monitoring: Continuous performance tracking in deployment

d) Validation: Multi-center clinical trials beyond benchmark testing

Current benchmarks do not adequately assess these safety-critical aspects [7,14].

Recommendations for Clinical Implementation

Based on the analysis, the following recommendations are proposed:

For AI Developers

1) Conduct comprehensive stress testing beyond standard benchmarks

2) Report stratified performance across specialties and difficulty levels

3) Validate on diverse patient populations and clinical settings

4) Implement robust hallucination detection and mitigation

For Healthcare Institutions

1) Do not rely solely on benchmark scores for deployment decisions

2) Require specialty-specific validation in target clinical domains

3) Implement human-in-the-loop oversight for all AI diagnostics

4) Establish continuous monitoring protocols post-deployment

For Researchers

1) Develop benchmarks that better reflect clinical complexity

2) Create standardized stress testing protocols

3) Investigate patient history integration approaches

4) Study long-term performance stability in clinical settings

Future Directions

Several promising approaches may address identified limitations [5,13]:

Enhanced Training Methodologies: Active learning strategies prioritizing rare and challenging cases, synthetic data generation for underrepresented conditions, and integration of structured medical ontologies could improve model robustness.

Hybrid Architectures: Combining LLMs with specialized medical knowledge graphs, evidence-based clinical decision support rules, and real-time literature retrieval systems may mitigate hallucination risks and temporal degradation.

Confidence-Aware Systems: Development of models that reliably recognize and communicate uncertainty, defer to human expertise for challenging cases, and provide graduated confidence scores calibrated to true accuracy could enhance safety.

Patient History Integration: Incorporating longitudinal patient records, historical treatment responses, and individual risk factors has demonstrated potential to reduce hallucinations by 15-20% in preliminary studies [14,16,17].

Conclusion

While large language models demonstrate impressive performance on standard medical benchmarks, systematic stress testing reveals critical vulnerabilities that challenge their readiness for clinical deployment. Performance degradation of 15-35% on edge cases, rare diseases, and underrepresented specialties, combined with problematic hallucination patterns and confidence miscalibration, indicates that current systems require substantial refinement before autonomous clinical use.

The "illusion of readiness" created by headline benchmark metrics must be replaced with comprehensive, realistic evaluation frameworks that accurately reflect the complexity and diversity of real-world clinical practice. Healthcare institutions, AI developers, and regulatory bodies must collaborate to establish appropriate safeguards, implementation protocols, and continuous improvement mechanisms.

The potential of AI in medicine remains substantial, but responsible deployment requires honest assessment of current limitations, transparent communication of capabilities and constraints, and commitment to iterative development guided by real-world performance data and patient safety outcomes.

Acknowledgments

None.

Conflict of Interest

None.

References

1. Andreeva ON, Domashev AV, Evlanova EA (2019) Artificial Intelligence in Medicine: Current State and Prospects. Medical Informatics Publishing: 232.
2. Shcherbakov, A Yu, Ryazanova AA (2021) Medical AI market analysis. Healthcare Informatics Journal 2(1): 33-40.
3. NEJM Image Challenge Dataset. *https://www.nejm.org/image-challenge*
4. JAMA Clinical Challenge Dataset. *https://jamanetwork.com/collections/44038/clinical-challenge*
5. Lau J, Soumya Gayen, Asma Ben Abacha, Dina Demner Fushman (2018) A dataset of clinically generated visual questions and answers about radiology images. Scientific Data 5(1): 180251.
6. Jin Y (2024) OmniMedVQA: A Comprehensive Medical Visual Question Answering Benchmark.
7. Gu Y, Fu J, Liu X, Valanarasu JMJ, Noel CF Codella, et al. (2025) The Illusion of Readiness: Stress Testing Large Frontier Models on Multimodal Medical Benchmarks. Microsoft Research, Health Life Sciences.
8. OpenAI (2025) GPT-5. *https://openai.com/index/introducing-gpt-5*
9. Google DeepMind (2025) Gemini-2.5 Pro. *https://deepmind.google/models/gemini/pro*
10. OpenAI (2025) Introducing o3 and o4-mini. *https://openai.com/index/introducing-o3-and-o4-mini*
11. OpenAI (2024) Hello GPT-4o. *https://openai.com/index/hello-gpt-4o*
12. (2025) DeepSeek-VL2. *https://github.com/deepseek-ai/DeepSeek-VL2*
13. Kojima T, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo (2022) Large language models are zero-shot reasoners.
14. Shcherbakov A Yu (2025) Integration of patient history in medical AI systems. Medical Informatics 24: 4-12.
15. Sagawa S, Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, et al. (2021) WILDS: A Benchmark of in-the-Wild Distribution Shifts.
16. Rajpurkar P, Matthew P Lungren (2023) The Current and Future State of AI Interpretation of Medical Images. New England Journal of Medicine 388(21): 1981-1990.
17. Wei J, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, et al. (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.