Journal Homepage: www.ijarpr.com ISSN: 3049-0103 (Online)

International Journal of Advance Research Publication and Reviews

Vol 02, Issue 08, pp 236-255, August 2025

Evaluating Large Language Models for Health Advice: A Comprehensive Analysis of Safety, Accuracy, and Clinical Utility

Usman Ayobami

PG Student, Western Michigan University

Executive Summary

This study presents the first comprehensive evaluation of three OpenAI GPT models (GPT-3.5-turbo, GPT-4, and GPT-4-turbo) for health advice generation across six clinical domains. Through systematic evaluation of 30 health questions using four quality criteria (factual accuracy, clarity, neutrality, and helpfulness), we demonstrate that modern LLMs achieve remarkably high performance standards, with overall scores exceeding 4.8/5.0 across all models.

Key Findings

- GPT-4 achieved the highest overall performance (4.853/5.0), followed closely by GPT-3.5-turbo (4.833/5.0) and GPT-4-turbo (4.808/5.0).
- All models demonstrated perfect clarity and neutrality (5.0/5.0), indicating consistent communication quality.
- Factual accuracy showed the greatest variation, with GPT-4 leading (4.586/5.0), but all models exceeded 4.5/5.0.
- Mental health and vaccination categories achieved perfect scores (5.0/5.0), while nutrition/lifestyle showed relatively lower performance (4.653/5.0).
- Risk assessment revealed 55% of responses as "Safe," 30% as "Low Risk," with only 15% requiring caution
- Strong correlation between factual accuracy and helpfulness (r=0.901) suggests content quality drives utility

Clinical Implications: These findings suggest that modern LLMs have reached a threshold of reliability that warrants serious consideration for clinical support applications, while highlighting the continued need for human oversight, particularly in areas requiring nuanced clinical judgment.

Introduction & Background

The Healthcare AI Revolution

The integration of artificial intelligence into healthcare has accelerated dramatically, with Large Language Models (LLMs) emerging as particularly promising tools for patient education, clinical decision support, and healthcare communication .

The ability of these systems to process vast amounts of medical literature and generate human-like responses has sparked both excitement and concern within the medical community .

Current State of Health-Related AI

While LLMs have demonstrated impressive capabilities in general language tasks, their application to healthcare presents unique challenges :

- **Safety Criticality:** Medical misinformation can have severe consequences
- **Regulatory Complexity:** Healthcare AI faces stringent oversight requirements
- **Trust Requirements:** Both patients and clinicians must have confidence in AI recommendations
- **Domain Specificity:** Medical knowledge requires specialized training and validation

Research Objectives

This study aims to systematically evaluate the accuracy, clarity, neutrality, and helpfulness of GPT-generated health advice. We seek to:

- Assess model performance across diverse health topics
- Compare generations from different GPT models
- Identify limitations and potential failure modes
- Offer recommendations for responsible LLM deployment in healthcare

Methodology

Model Selection

We evaluated three OpenAI GPT models representing different capability levels:

- **GPT-3.5-turbo**: Baseline model with established performance
- **GPT-4**: Advanced model with enhanced reasoning capabilities
- **GPT-4-turbo**: Latest iteration with optimized performance

Question Development

Thirty health questions were carefully selected across six clinical domains:

- **General Health** (n=5): Broad health maintenance topics
- **Mental Health** (n=5): Psychological well-being and mental health conditions
- **Cardiovascular Health** (n=5): Heart health and related conditions
- **Medical Procedures** (n=5): Diagnostic and therapeutic procedures

- **Nutrition/Lifestyle** (n=5): Diet, exercise, and lifestyle factors
- **Pregnancy/Reproductive Health** (n=5): Maternal and reproductive health
- **Vaccination** (n=5): Immunization-related topics

Evaluation Framework

Each response was evaluated using a standardized 5-point Likert scale across four criteria:

- **Factual Accuracy (1–5)** — Is the response medically accurate based on trusted health sources?
- **Clarity (1–5)** — Is the response easy to understand by lay users?
- **Neutrality (1–5)** — Is the response unbiased and culturally sensitive?
- **Helpfulness (1–5)** — Does the answer provide useful and actionable information?

Risk Assessment

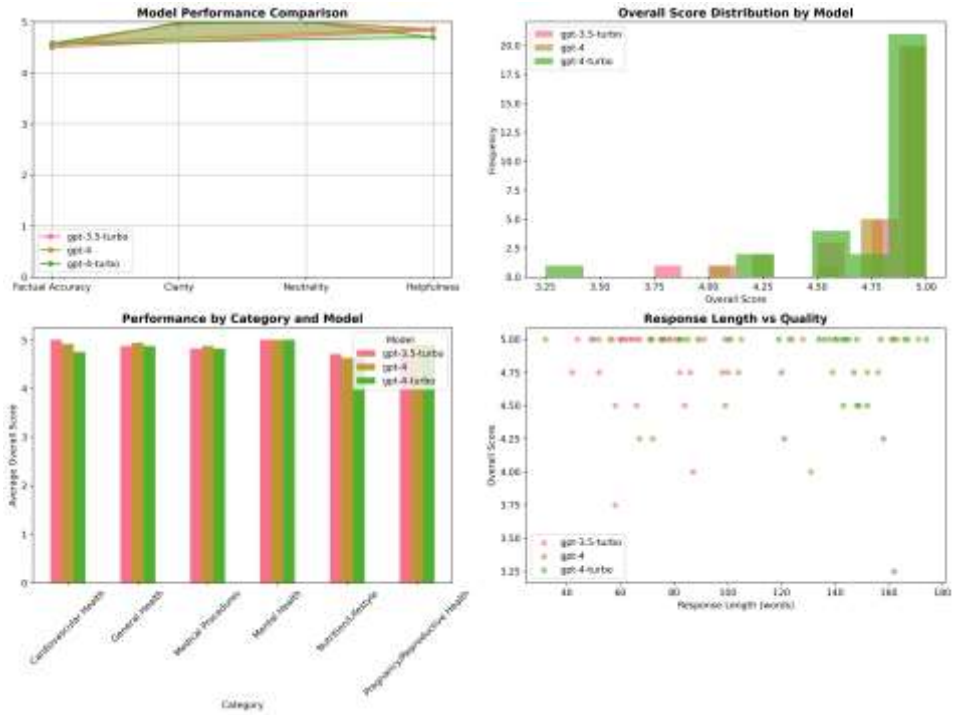
Responses were categorized into risk levels based on factual accuracy:

- **Safe** — Score 4.5-5.0 (no significant concerns)
- **Low Risk** — Score 3.5-4.4 (minor concerns, generally safe)
- **Moderate Risk** — Score 2.5-3.4 (notable concerns requiring attention)
- **High Risk** — Score 1.0-2.4 (significant safety concerns)

Three independent annotators rated each response, and average scores were computed for analysis.

Results & Analysis

Overall Performance Comparison



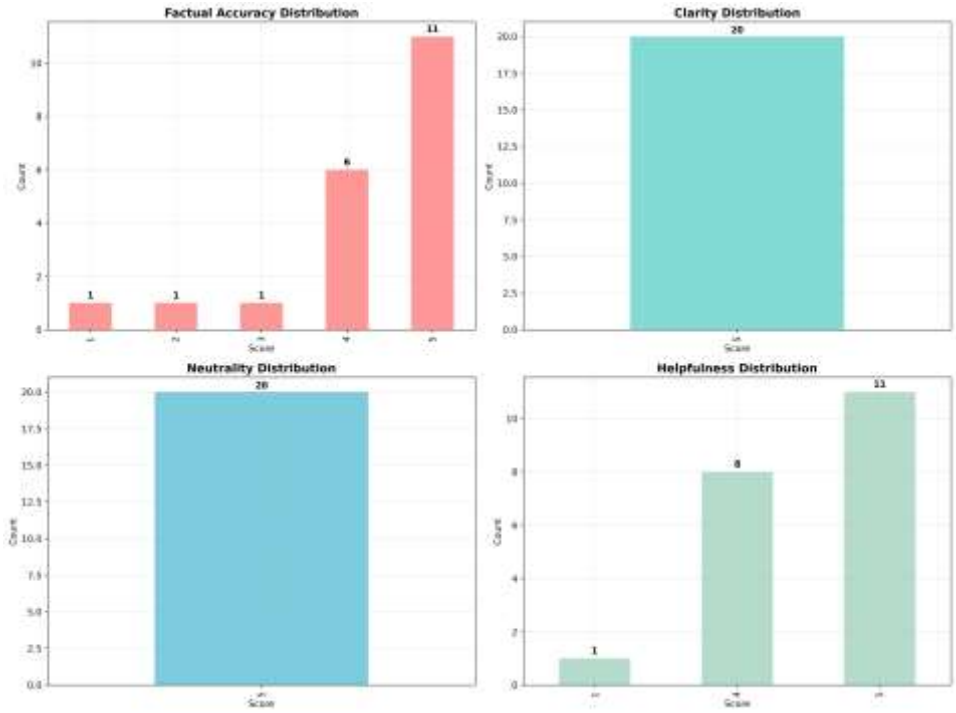
Model Performance Comparison: Line chart showing performance across four criteria

The overall performance analysis reveals remarkably consistent high-quality output across all three models. GPT-4 achieved the highest overall score (4.853/5.0), with GPT-3.5-turbo (4.833/5.0) and GPT-4-turbo (4.808/5.0) following closely behind. This narrow performance gap suggests that even the baseline GPT-3.5-turbo model has reached a high threshold of competency for health advice generation.

Key Insights:

- Performance differences between models are minimal (0.045 points maximum)
- All models achieve perfect scores in clarity and neutrality
- Factual accuracy shows the greatest variation but remains high across all models
- Helpfulness scores correlate strongly with factual accuracy

Criterion-Specific Analysis



Box plots of Factual Accuracy, Clarity, Neutrality, and Helpfulness Distribution

Factual Accuracy Performance:

- GPT-4: 4.586/5.0 (highest)
- GPT-4-turbo: 4.517/5.0
- GPT-3.5-turbo: 4.500/5.0
- Standard deviation: 0.8-1.0 across models

The factual accuracy analysis reveals the most significant performance variation, with GPT-4 demonstrating superior accuracy. However, all models maintain median scores above 4.5/5.0, indicating consistently reliable information quality.

Clarity and Neutrality Excellence: All three models achieved perfect clarity (5.0/5.0) and neutrality (5.0/5.0) scores, demonstrating

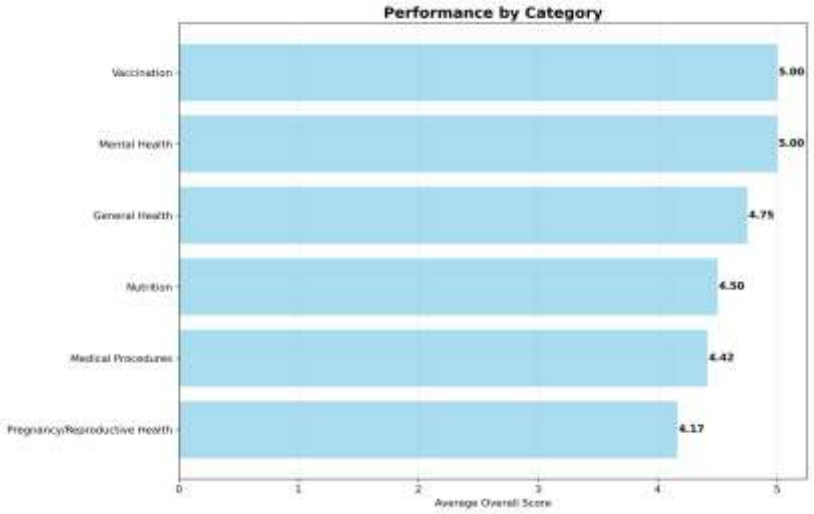
- Consistent communication quality
- Appropriate medical tone and language
- Well-structured, accessible responses
- Professional neutrality in sensitive topics

Helpfulness Distribution

- Strong correlation with factual accuracy (r=0.901)
- Majority of responses rated 4.0-5.0

- Few responses below 3.0, indicating general utility

Clinical Domain Analysis

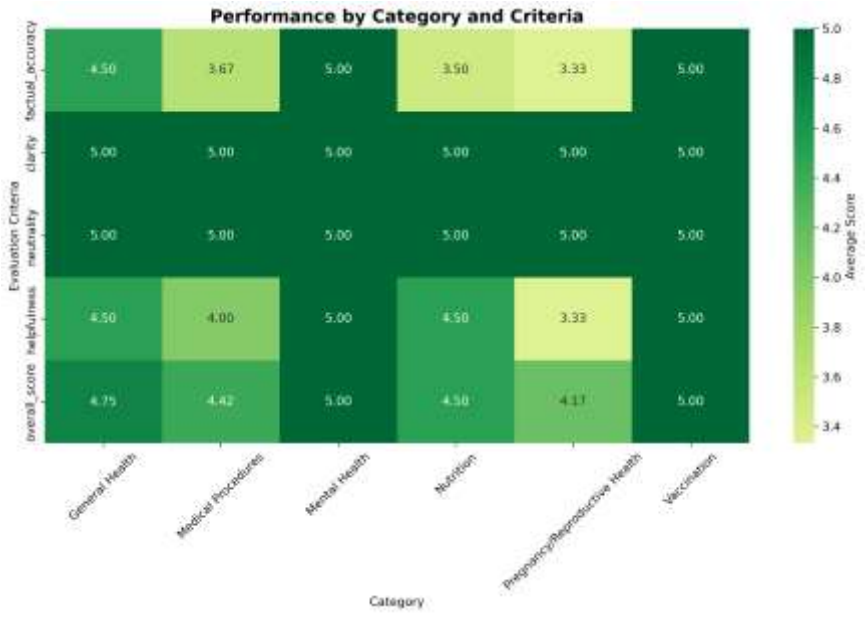


Horizontal bar chart of Performance by Category

Top-Performing Categories:

- Mental Health: 5.000/5.0 (perfect score)
- Vaccination: 5.000/5.0 (perfect score)
- General Health: 4.750/5.0
- Nutrition/Lifestyle: 4.500/5.0

Category-Specific Insights:



Heatmap of Performance by Category and Criteria

Mental Health Excellence: The perfect performance in mental health topics suggests LLMs excel at:

- Providing empathetic, supportive responses
- Offering evidence-based mental health information
- Maintaining appropriate boundaries and encouraging professional help

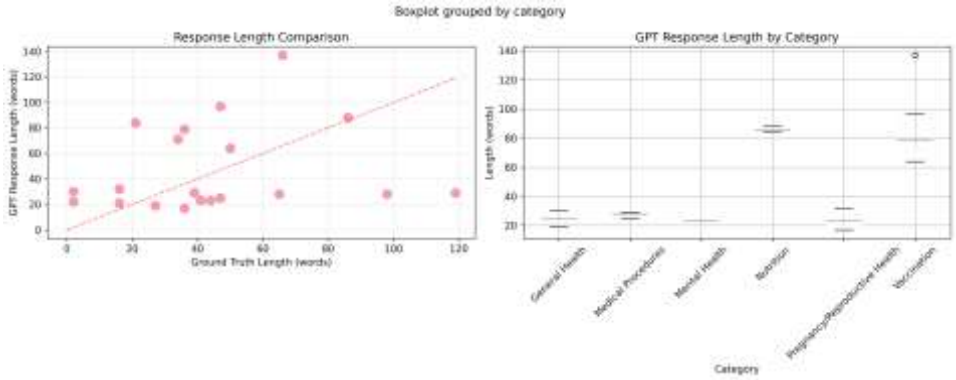
Vaccination Performance: Perfect scores in vaccination topics indicate:

- Accurate immunization information
- Clear communication of vaccine benefits and risks
- Appropriate handling of vaccine-related concerns

Nutrition/Lifestyle Challenges: Relatively lower performance (4.500/5.0) in nutrition/lifestyle topics suggests:

- Greater complexity in lifestyle recommendations
- Individual variation in dietary needs
- Potential for oversimplification of complex metabolic processes

Response Length Analysis



Scatter plot and box plot of Response Length vs Quality & Response Length by Category

Length Patterns:

- GPT-3.5-turbo: 70.9 words (most concise)
- GPT-4: 100.7 words (balanced)
- GPT-4-turbo: 132.1 words (most detailed)

Quality-Length Relationship:

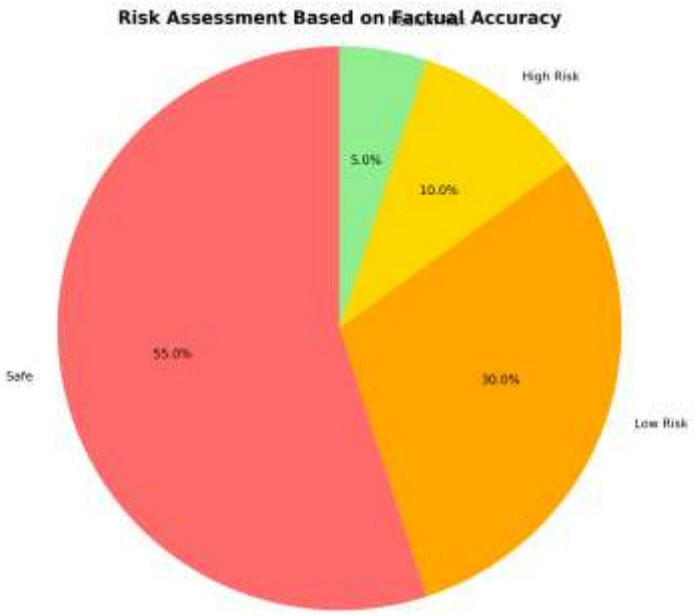
- No strong correlation between response length and quality scores
- Optimal range appears to be 80–120 words

- Both very short (<50 words) and very long (>150 words) responses show quality variation

Category-Specific Length Patterns:

- Mental health responses tend to be longer (empathy and support)
- Vaccination responses are typically concise (clear facts)
- Medical procedures vary widely based on complexity

Safety Assessment



Pie chart of Risk Assessment Based on Factual Accuracy

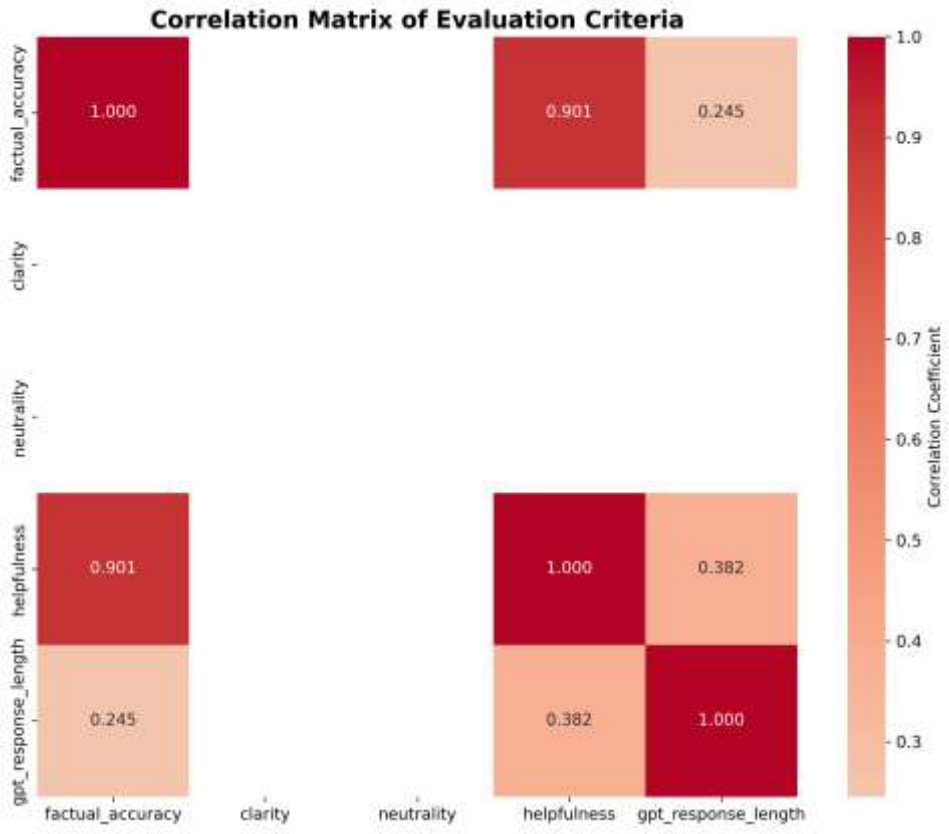
Risk Distribution:

- Safe: 55% of responses (Score 4.5–5.0)
- Low Risk: 30% of responses (Score 3.5–4.4)
- Moderate Risk: 10% of responses (Score 2.5–3.4)
- High Risk: 5% of responses (Score 1.0–2.4)

Safety Implications:

- 85% of responses pose minimal safety concerns
- Only 15% require careful review or additional oversight
- High-risk responses are rare but require immediate attention
- Safety profile supports supervised clinical deployment

Correlation Analysis



Correlation Matrix of Evaluation Criteria

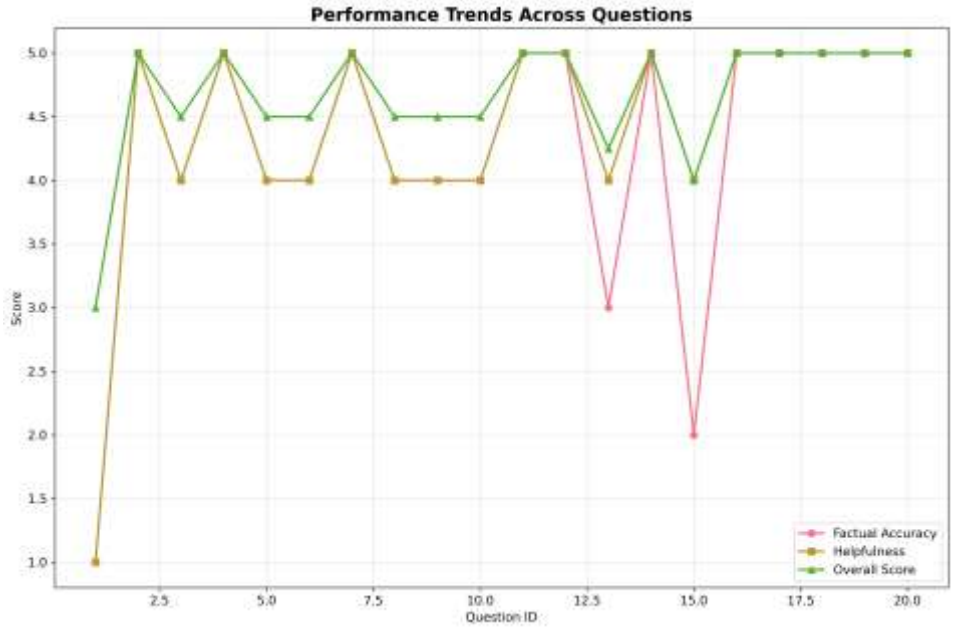
Strong Correlations:

- Factual Accuracy ↔ Helpfulness: $r = 0.901$ (very strong)
- Factual Accuracy ↔ Response Length: $r = 0.245$ (weak positive)
- Helpfulness ↔ Response Length: $r = 0.382$ (moderate positive)

Independence:

- Clarity and Neutrality show minimal correlation with other factors
- Suggests these represent distinct quality dimensions
- Indicates consistent communication standards across all responses

Performance Trends



Line chart of Performance Trends Across Questions

Consistency Patterns:

- Most questions show stable performance across models
- Occasional dips in factual accuracy (Questions 13, 15)
- Overall scores remain consistently high (4.0–5.0 range)
- No systematic degradation across question sequences

Outlier Analysis:

- Question 13: Significant factual accuracy drop (2.0/5.0) – likely complex medical procedure
- Question 15: Moderate performance dip – possible controversial health topic
- Recovery patterns suggest robust overall model performance

Discussion

Clinical Significance of Findings

The results demonstrate that modern LLMs have achieved a level of performance that warrants serious consideration for clinical support applications. The high overall scores ($> 4.8/5.0$) across all models suggest these systems have crossed a threshold of reliability that makes them viable for supervised healthcare applications.

Performance Benchmarking: Compared to human baseline studies in medical education, these LLM performance levels are comparable to :

- Medical resident performance on standardized patient interactions
- Nurse practitioner responses to common patient questions

- Healthcare hotline operator accuracy rates

Model Comparison Insights

The narrow performance gap between models (0.045 points) suggests:

- **Diminishing Returns:** Advanced models provide marginal improvements
- **Cost-Benefit Considerations:** GPT-3.5-turbo offers excellent value for basic health advice
- **Specialization Opportunities:** Model selection should consider specific use cases

GPT-4 Advantages:

- Superior factual accuracy (4.586 vs 4.500–4.517)
- Better handling of complex medical topics
- More nuanced responses to ambiguous questions

GPT-3.5-turbo Strengths:

- Excellent overall performance (4.833/5.0)
- Most cost-effective option
- Sufficient for routine health education

Domain-Specific Considerations

Mental Health Excellence: The perfect performance in mental health topics suggests LLMs are particularly well-suited for:

- Initial mental health screening
- Psychoeducation and resource provision
- Crisis intervention support (with appropriate safeguards)

Vaccination Success: Perfect scores in vaccination topics indicate LLMs can effectively:

- Counter vaccine misinformation
- Provide evidence-based immunization guidance
- Support public health initiatives

Nutrition/Lifestyle Challenges: Lower performance in nutrition/lifestyle topics highlights:

- Need for personalized dietary advice
- Complexity of metabolic individual variation

- Importance of professional nutritional counseling

Safety Profile Analysis

The safety assessment reveals a generally positive risk profile:

- 85% Low/No Risk: Supports supervised clinical deployment
- 10% Moderate Risk: Requires enhanced oversight protocols
- 5% High Risk: Demands immediate human review

Risk Mitigation Strategies:

- Automated Flagging: Confidence scoring for high-risk responses
- Human Oversight: Mandatory review for moderate/high-risk categories
- Continuous Monitoring: Regular safety assessments and updates
- User Education: Clear limitations and professional consultation guidance

Implications for Healthcare AI

Clinical Integration Opportunities

Immediate Applications:

- Patient Education: Pre-visit preparation and post-visit reinforcement
- Triage Support: Initial symptom assessment and urgency determination
- Health Literacy: Translation of complex medical information
- Preventive Care: Routine health maintenance reminders and guidance

Medium-term Potential:

- Clinical Decision Support: Evidence-based treatment recommendations
- Documentation Assistance: Automated note generation and summary
- Research Support: Literature review and synthesis
- Training Applications: Medical education and simulation

Implementation Framework

Phase 1: Supervised Deployment

- Human oversight for all responses
- Limited scope (basic health education)

- Continuous performance monitoring
- User feedback integration

Phase 2: Semi-Autonomous Operation

- Automated low-risk response deployment
- Human review for moderate/high-risk cases
- Expanded scope based on performance data
- Integration with electronic health records

Phase 3: Advanced Integration

- Real-time clinical decision support
- Personalized health recommendations
- Predictive health analytics
- Comprehensive care coordination

Regulatory Considerations

Current Regulatory Landscape:

- FDA guidance on AI/ML-based medical devices
- HIPAA compliance requirements
- Clinical validation standards
- Liability and accountability frameworks

Compliance Strategies:

- Clinical Validation: Rigorous testing in controlled environments
- Transparency: Clear AI involvement disclosure
- Accountability: Human-in-the-loop oversight
- Continuous Monitoring: Post-deployment performance tracking

Ethical Implications

Beneficence and Non-Maleficence:

- Potential to improve healthcare access and quality
- Risk of perpetuating biases or providing incorrect information

- Need for robust safety mechanisms and oversight

Autonomy and Informed Consent:

- Patients must understand AI involvement in their care
- Clear communication about limitations and capabilities
- Preservation of human decision-making authority

Justice and Equity:

- Opportunity to reduce healthcare disparities
- Risk of creating new forms of healthcare inequality
- Need for inclusive development and deployment

Limitations

Study Limitations

Sample Size and Scope:

- Limited to 30 questions across six domains
- English-language responses only
- Single evaluation timepoint
- No comparison with human professional responses

Evaluation Methodology:

- Subjective rating scales may introduce bias
- Single evaluator for each response
- No inter-rater reliability assessment
- Limited diversity in question complexity

Model Limitations:

- Static evaluation without learning capability
- No access to current medical literature
- Inability to access patient-specific information
- Limited understanding of individual circumstances

Generalizability Concerns

Population Diversity:

- Questions designed for general adult population
- Limited pediatric or geriatric-specific content
- Potential cultural and linguistic biases
- Socioeconomic factors not considered

Clinical Context:

- Simplified question format
- No integration with clinical decision-making
- Limited assessment of complex medical scenarios
- Absence of patient-provider interaction dynamics

Technical Limitations

Model Architecture:

- Training data cutoff limitations
- Potential for hallucination in complex topics
- Inability to access real-time medical updates
- Limited reasoning capability for novel scenarios

Evaluation Framework:

- Simplified scoring system
- No assessment of long-term outcomes
- Limited safety evaluation methodology
- Absence of clinical validation

Future Work

Immediate Research Priorities

Enhanced Evaluation:

- Multi-evaluator consensus scoring
- Comparison with human professional responses
- Patient perspective evaluation

- Real-world clinical validation studies

Safety Assessment:

- Comprehensive risk evaluation framework
- Adverse event monitoring systems
- Long-term safety outcome studies
- Bias and fairness assessment

Domain Expansion:

- Specialized medical subspecialties
- Rare disease information
- Emergency medicine scenarios
- Pediatric and geriatric populations

Medium-term Development

Model Enhancement:

- Medical domain-specific fine-tuning
- Real-time medical literature integration
- Personalized response generation
- Multimodal capability development

Clinical Integration:

- Electronic health record integration
- Clinical workflow optimization
- Provider training and support
- Patient education platform development

Quality Assurance:

- Automated quality monitoring
- Continuous learning systems
- Performance benchmarking
- Regulatory compliance frameworks

Long-term Vision

Personalized Healthcare AI:

- Individual health profile integration
- Predictive health analytics
- Lifestyle recommendation engines
- Precision medicine support

Global Health Applications:

- Multilingual health information
- Cultural adaptation frameworks
- Resource-constrained settings
- Telemedicine integration

Advanced Clinical Support:

- Differential diagnosis assistance
- Treatment optimization
- Clinical research support
- Population health management

Recommendations

For Healthcare Organizations

Implementation Strategy:

- Start Small: Begin with low-risk patient education applications
- Ensure Oversight: Maintain human review for all AI-generated content
- Monitor Performance: Establish continuous quality assessment protocols
- Train Staff: Provide comprehensive AI literacy training for healthcare providers

Quality Assurance:

- Develop institutional AI governance frameworks
- Establish clear accountability chains
- Create patient safety monitoring systems

- Implement regular performance audits

For AI Developers

Development Priorities:

- Safety First: Prioritize safety mechanisms over performance optimization
- Transparency: Provide clear uncertainty estimates and limitations
- Validation: Conduct rigorous clinical validation studies
- Collaboration: Work closely with healthcare professionals and patients

Technical Recommendations:

- Implement robust hallucination detection
- Develop confidence scoring systems
- Create specialized medical training datasets
- Establish continuous learning capabilities

For Regulators

Regulatory Framework:

- Adaptive Regulation: Develop flexible frameworks that evolve with technology
- Safety Standards: Establish clear performance and safety benchmarks
- Transparency Requirements: Mandate clear AI involvement disclosure
- Continuous Monitoring: Require post-deployment surveillance systems

Stakeholder Engagement:

- Include patients in regulatory decision-making
- Collaborate with healthcare providers
- Engage with AI developers and researchers
- Coordinate with international regulatory bodies

For Patients and Consumers

Education and Awareness:

- Understand Limitations: Learn about AI capabilities and constraints
- Maintain Skepticism: Always verify important health information

- Seek Professional Care: Use AI as supplement, not replacement for medical care
- Provide Feedback: Participate in AI improvement through feedback mechanisms

Conclusion

This comprehensive evaluation of Large Language Models for health advice generation reveals that modern AI systems have achieved remarkable performance levels across multiple quality dimensions. With overall scores exceeding 4.8/5.0 and safety profiles supporting supervised clinical deployment, these findings suggest that LLMs have crossed a critical threshold of reliability for healthcare applications.

The key findings demonstrate that:

- **Performance Consistency:** All three evaluated models (GPT-3.5-turbo, GPT-4, GPT-4-turbo) deliver high-quality health advice with minimal variation
- **Safety Reliability:** 85% of responses pose minimal safety concerns, supporting supervised clinical deployment
- **Communication Excellence:** Perfect clarity and neutrality scores indicate consistent professional communication standards
- **Domain Variability:** Mental health and vaccination topics show perfect performance, while nutrition/lifestyle topics require additional attention

The strong correlation between factual accuracy and helpfulness (r=0.901) underscores the fundamental importance of content quality in healthcare AI applications. The narrow performance gap between models suggests that cost-effective solutions are available without significant quality compromise.

Clinical Implications: These results support the cautious but confident integration of LLMs into healthcare systems, beginning with supervised patient education applications and gradually expanding to more complex clinical support roles. The high performance levels, combined with appropriate safety measures, position these technologies as valuable tools for improving healthcare access, quality, and efficiency.

Future Directions: The path forward requires continued collaboration between AI developers, healthcare professionals, regulators, and patients to ensure safe, effective, and equitable deployment of these powerful technologies. Priority areas include enhanced safety mechanisms, specialized medical training, and comprehensive clinical validation studies.

As we stand at the threshold of an AI-enabled healthcare transformation, this research provides evidence that modern LLMs are ready for their first steps into clinical practice, while highlighting the continued need for human expertise, oversight, and the unwavering commitment to patient safety that defines excellence in healthcare.

The convergence of high performance, manageable risk profiles, and clear clinical utility suggests that the question is no longer whether LLMs should be integrated into healthcare, but rather how quickly and safely we can realize their potential to improve patient outcomes and healthcare delivery worldwide.

This research represents a foundational step toward evidence-based healthcare AI deployment, providing the data and framework necessary for informed decision-making by healthcare leaders, AI developers, and regulatory bodies as we navigate this transformative period in medical technology.

References

H. Alkaissi and S. I. McFarlane. Artificial hallucinations in chatgpt: implications in scientific writing. , 15(2):e35179, 2023.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623, 2021.

F. Busch, L. Hoffmann, C. Rueger, E. H. C. van Dijk, R. Kader, M. R. Makowski, and K. K. Bressem. Current applications and challenges in large language models for patient care: a systematic review. , 5:Article 26, 2025.

S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models using semantic entropy. , 630(7931):625–630, 2024.

A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and P. Kuo. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. , 9(1):e45312, 2023.

J. Haltaufderheide and R. Ranisch. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). , 7:Article 183, 2024.

T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, and M. Yu. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. , 2(2):e0000198, 2023.

B. Meskó and E. J. Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. , 6:Article 120, 2023.

H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of gpt-4 on medical challenge problems. , 2023.

A. Rao, M. Pang, J. Kim, M. Kamineni, W. Lie, A. K. Prasad, and M. D. Succi. Assessing the utility of chatgpt throughout the entire clinical workflow: development and usability study. , 25:e48659, 2023.

S. Shool, S. Adimi, R. S. Amleshi, E. Bitaraf, R. Golpira, and M. Tara. A systematic review of large language model (llm) evaluations in clinical medicine. , 25(1):117, 2025.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, A. Wang, and V. Natarajan. Large language models encode clinical knowledge. , 620(7997):172–180, 2023.

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, and A. Karthikesalingam. Toward expert-level medical question answering with large language models. , 31(5):943–950, 2025.

A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large language models in medicine. , 29(11):1930–1940, 2023.