DOI: 10.4274/anatoljmed.2026.67044
Anatol J Gen Med Res 2026;36(2):200-11

CLINICAL RESEARCH / KLİNİK ARAŞTIRMA

ANATOLIAN JOURNAL of GENERAL MEDICAL RESEARCH

A Human-centered Evaluation of Artificial Intelligence and a Single Pulmonologist's Responses in Pulmonary Medicine Communication

Göğüs Hastalıkları İletişiminde Yapay Zeka ve Tek Bir Göğüs Hastalıkları Hekiminin Yanıtlarının İnsan Odaklı Değerlendirmesi

Damla Serçe Unat[1,4], Ömer Selim Unat[2,5], Mutlu Onur Güçsav[1,5], Berkay Can Karakoç[3], Aysu Ayrancı[1], Ahmet Emin Erbaycu[1]

[1]İzmir Bakırçay University Faculty of Medicine, Department of Pulmonology, İzmir, Türkiye
[2]University of Health Sciences Türkiye, İzmir Dr. Suat Seren Chest Diseases and Surgery Training and Research Hospital, İzmir, Türkiye
[3]Başkent University, Institute of Science and Technology, Department of Defense Electronics and Software, Ankara, Türkiye
[4]Ege University Faculty of Medicine, Department of Basic Oncology, İzmir, Türkiye
[5]Ege University Faculty of Medicine, Department of Translational Pulmonology, İzmir, Türkiye

Cite as: Serçe Unat D, Unat ÖS, Güçsav MO, Karakoç BC, Ayrancı A, Erbaycu AE. A human-centered evaluation of artificial intelligence and a single pulmonologist's responses in pulmonary medicine communication. Anatol J Gen Med Res. 2026;36(2):200-11

Abstract

Objective: Large language models (LLMs) are increasingly used by patients and the public to access medical information. In pulmonary medicine, where information needs span a wide range of chronic and potentially serious diseases, the quality of artificial intelligence-generated communication requires systematic evaluation.

Methods: This cross-sectional comparative study evaluated responses to 37 standardized pulmonary medicine questions covering major respiratory domains. Each question was answered by four LLMs-Gemini 2.5, GPT-4, Claude Sonnet 4, and DeepSeek v3.2-and a board-certified professor of pulmonology with more than 25 years of clinical experience. Six blinded raters, including three pulmonologists and three laypersons, assessed anonymized responses using a five-point Likert scale across conceptually aligned domains: accuracy/perceived reliability, clarity, and scientific adequacy/practical usefulness. Inter-rater reliability was assessed using intraclass correlation coefficients (ICC). Comparative analyses were performed using repeated-measures one way analysis of variance with Bonferroni-adjusted pairwise comparisons and Friedman tests.

Results: Inter-rater reliability indicated slight single-measure agreement (ICC 0.09-0.12) and fair-to-moderate average-measure agreement (ICC 0.36-0.45), which is consistent with the subjective nature of communication-focused assessment. Overall, LLM-generated responses received higher communication-quality ratings than physician-generated responses across evaluation domains. Gemini 2.5 and GPT-4 achieved the most favorable scores, particularly on accuracy/perceived reliability and clarity. Differences between LLMs and the physician were more pronounced among lay raters than among expert raters. Among expert raters, physician-generated responses were rated among the highest for accuracy and scientific adequacy, although this pattern was not observed for clarity.

Conclusion: Contemporary LLMs may provide clear, reliable, and practically useful information on pulmonary medicine, particularly as perceived by lay users, when evaluated from both expert and lay perspectives. These findings support their adjunctive use in patient education and health communication and emphasize that LLMs cannot replace clinical judgment, professional accountability, or physician oversight.

Keywords: Large language model, artificial intelligence, medical informatics, health communication, patient education, pulmonary medicine

Address for Correspondence/Yazışma Adresi: Damla Serçe Unat, MD, İzmir Bakırçay University Faculty of Medicine, Department of Pulmonology, İzmir, Türkiye
E-mail: damla.serceunat@bakircay.edu.tr
ORCID ID: orcid.org/0000-0003-4743-5469

Received/Geliş tarihi: 17.05.2026
Accepted/Kabul tarihi: 26.06.2026
Published date/Yayınlanma tarihi: 31.08.2026

Öz

Amaç: Büyük dil modelleri (LLM), hastalar ve toplum tarafından tıbbi bilgiye ulaşmak amacıyla giderek daha sık kullanılmaktadır. Göğüs hastalıkları alanında bilgi gereksinimi kronik ve potansiyel olarak ciddi hastalıkları kapsadığından, yapay zeka tarafından oluşturulan tıbbi iletişim içeriğinin kalitesinin sistematik olarak değerlendirilmesi önem taşımaktadır.

Yöntem: Bu kesitsel karşılaştırmalı çalışmada, temel solunum hastalıkları alanlarını kapsayan 37 standart göğüs hastalıkları sorusuna verilen yanıtlar değerlendirildi. Her soru dört LLM-Gemini 2.5, GPT-4, Claude Sonnet 4 ve DeepSeek v3.2-ile 25 yıldan fazla klinik deneyime sahip bir göğüs hastalıkları profesörü tarafından yanıtlandı. Üç göğüs hastalıkları uzmanı ve üç uzman olmayan değerlendiriciden oluşan altı kör değerlendirici, anonimleştirilmiş yanıtları beşli Likert ölçeği ile; doğruluk/algılanan güvenilirlik, açıklık ve bilimsel yeterlilik/pratik yararlılık alanlarında değerlendirdi. Değerlendiriciler arası güvenilirlik sınıf içi korelasyon katsayısı (ICC) ile analiz edildi. Karşılaştırmalı analizlerde tekrarlı ölçümler tek yönlü varyans analizi, Bonferroni düzeltmeli ikili karşılaştırmalar ve Friedman testi kullanıldı.

Bulgular: Değerlendiriciler arası güvenilirlik, tekil ölçümlerde zayıf (ICC 0,09-0,12), ortalama ölçümlerde ise düşük-orta ile orta düzeyde (ICC 0,36-0,45) bulundu ve bu durum iletişim odaklı değerlendirmenin öznel yapısı ile uyumluydu. Genel olarak, büyük dil modeli yanıtları tüm değerlendirme alanlarında hekim yanıtlarına göre daha yüksek iletişim kalitesi puanları aldı. Gemini 2.5 ve GPT-4 özellikle doğruluk/algılanan güvenilirlik ve açıklık alanlarında en olumlu skorları elde etti. Büyük dil modelleri ile hekim yanıtları arasındaki farklar, uzman değerlendiricilere kıyasla uzman olmayan değerlendiricilerde daha belirgindi. Uzman değerlendiriciler arasında ise hekim tarafından oluşturulan yanıtlar doğruluk ve bilimsel yeterlilik açısından en yüksek puan alan yanıtlar arasında yer alırken, bu örüntü açıklık alanında gözlenmedi.

Sonuç: Güncel büyük dil modelleri, hem uzman hem de uzman olmayan değerlendiriciler açısından açık, güvenilir ve pratik olarak yararlı göğüs hastalıkları bilgisi sunabilir. Bu bulgular, büyük dil modellerinin hasta eğitimi ve sağlık iletişiminde yardımcı araçlar olarak kullanılabileceğini desteklemektedir. Bununla birlikte, bu sistemler klinik yargının, mesleki sorumluluğun ve hekim gözetiminin yerini alamaz.

Anahtar Kelimeler: Büyük dil modelleri, yapay zeka, tıbbi bilişim, sağlık iletişimi, hasta eğitimi, göğüs hastalıkları

Introduction

Effective medical communication is a core component of high-quality healthcare delivery, since the accuracy, reliability, and clarity of the information shared with patients and caregivers directly shape clinical understanding and decision-making. Prior work has consistently demonstrated that clear and accurate responses to patient inquiries are associated with greater patient satisfaction, improved treatment adherence, and a stronger physician-patient relationship, making comprehensible and trustworthy communication central to patient-centered care[1,2]. This is particularly consequential in populations with limited health literacy, where access to accurate information from trustworthy sources is essential for both treatment success and patient safety[3].

In parallel, the integration of the internet and digital media into everyday life has reshaped how patients and their relatives seek health information, with online searches for symptoms and diagnoses giving rise to the so-called "Dr. Google" phenomenon[4]. Much of the health-related material available online, however, is inaccurate, incomplete, or contextually misleading[5], with the potential to provoke unnecessary anxiety, misdirected care-seeking, or distrust toward medical advice. From a public health perspective, the dissemination of scientifically accurate, comprehensible, and trustworthy health information through digital channels has therefore become both an individual and a collective responsibility[6].

Within this evolving information ecosystem, artificial intelligence (AI) has progressively moved from a peripheral analytic tool to a core component of contemporary medical informatics. Earlier generations of AI demonstrated value in image-based diagnostic support, bioinformatics pipelines, and the interpretation of large-scale data from genetic biobanks, where machine learning approaches enabled scalable analyses of complex, heterogeneous datasets that would be impractical to perform manually[7-9]. Building on this foundation, AI systems have since been embedded into clinical decision-support workflows, risk stratification models, and individualized treatment planning informed by genomic, proteomic, and radiomic data[10,11]. What distinguishes the current phase, however, is the migration of AI from structured data and pixel-level analysis toward the communication layer of medicine-the space in which patients, caregivers, and clinicians exchange information, weigh options, and arrive at decisions. This shift reframes AI not only as an analytic engine but also as an active participant in health information delivery and brings the quality, transparency, and safety of AI-generated communication into the foreground of medical informatics research.

Beyond diagnostics and data analytics, large language models (LLMs) are transforming the communication layer of medicine. Leveraging advanced natural language processing, these systems have shown promise in medical question answering, summarization of clinical dialogues, and analysis of unstructured text from electronic health records[11,12]. Their use in healthcare has expanded rapidly, with domain-specific implementations developed to improve the accuracy, clarity, and empathy of responses to medical queries[11]. At the same time, their deployment raises non-trivial concerns regarding factual reliability, contextual nuance, data security, algorithmic bias, and the risk of confidently incorrect outputs-issues that underscore the need for safe, transparent, and human-centered integration of AI technologies into clinical practice[7]. Systematic evaluation of LLM-generated responses, especially within clinically meaningful, domain-specific contexts such as pulmonology, is therefore essential before such tools can be responsibly recommended for patient-facing use[13].

These considerations are particularly salient in pulmonary medicine. Respiratory conditions-including asthma, chronic obstructive pulmonary disease (COPD), pneumonia, pulmonary thromboembolism, interstitial lung diseases, obstructive sleep apnea, tuberculosis, and lung cancer-account for a substantial share of the global disease burden and result in ongoing public information needs. Many of these conditions are chronic, require sustained self-management, and involve decisions regarding inhaler use, oxygen therapy, smoking cessation, vaccination, surgical options, and end-of-life care. The accessibility, accuracy, and clarity of the information delivered to patients and their relatives therefore have direct implications for adherence, safety, and informed decision-making. Despite the rapid uptake of LLMs in this space, the available evidence remains fragmented: many existing evaluations examine a single model, rely exclusively on expert reviewers, or assess technical accuracy without considering how lay audiences perceive the information. Studies comparing multiple state-of-the-art LLMs head-to-head with an experienced physician on the same set of publicly relevant questions-using blinded evaluations by both clinicians and laypersons across complementary criteria of accuracy, clarity, and scientific or practical adequacy-remain limited. This gap is consequential because expert and lay perspectives often diverge: clinicians tend to weigh factual completeness and technical precision, whereas patients and the broader public appraise responses through the lens of perceived reliability, comprehensibility, and real-world usefulness. Capturing both perspectives is therefore essential to characterize the realistic role of LLMs in patient-facing health communication and decision-support tasks.

The present study was designed to address this gap. We conducted a comparative, blinded evaluation in which four contemporary LLMs (Gemini 2.5, GPT-4, Claude Sonnet 4, and DeepSeek v3.2) and a board-certified pulmonology professor with more than 25 years of clinical experience responded to the same set of 37 standardized pulmonary medicine questions derived from publicly observable information-seeking behavior. Using conceptually aligned criteria (accuracy/perceived reliability, clarity, and scientific adequacy/practical usefulness), anonymized responses were independently rated by three expert pulmonologists and three laypersons. By examining inter-rater reliability and systematically comparing model and physician performance across audience types, the study provides an informatics-oriented assessment of how well current LLMs support communication about pulmonary medicine to non-specialist users, and aims to clarify their potential-and limitations-as complementary tools alongside experienced clinicians in health information delivery and clinical decision-support workflows.

Materials and Methods

This study was a cross-sectional, comparative evaluation designed to assess the quality and reliability of responses generated by four LLMs and a human pulmonology expert.

Design and Procedure

This study did not involve patients, patient data, identifiable personal information, biological samples, or any clinical intervention. The raters evaluated anonymized written responses; therefore, ethics committee approval was not required according to institutional and national regulations. Informed consent is not applicable, as the study did not involve patients or identifiable personal data.

Question Selection

A total of 37 standardized questions related to chest diseases were developed by a pulmonology specialist who had no conflicts of interest with either the expert or layperson raters. The questions were designed to represent a balanced distribution across major subspecialties within pulmonary medicine. Particular attention was paid to ensuring that the set encompassed diverse, clinically relevant domains: asthma (3), COPD (4), pneumonia (4), pulmonary thromboembolism

(4), interstitial lung diseases (3), obstructive sleep apnea (4), lung cancer (7), tuberculosis (2), and other respiratory topics (6). In addition to its clinical relevance, the study was designed to reflect real-world patterns of health information consumption, focusing on publicly accessible questions and communication formats that influence health-related understanding beyond clinical settings. The complete 37-item question set is provided as supplementary material; formal psychometric content validation (e.g., a content-validity index) was not performed.

The questions' content was informed by Google Trends data, social media discussions, and frequently asked questions directed to a pulmonologist's professional social media account, to ensure that the questions reflected publicly relevant, real-world, and digitally mediated health concerns. This approach provided a representative sampling of topics commonly encountered in both clinical practice and public health communication contexts.

Participants and Evaluation Procedure

Six independent raters participated in the evaluation: three expert pulmonologists (E1, E2, and E3, with 7, 8, and 10 years of clinical experience, respectively) and three laypersons (L1, L2, and L3, a 32-year-old university-educated teacher, a 53-year-old employee and former amateur footballer, and a 45-year-old homemaker with a high-school education). All raters were blinded to the identity of the content source (LLM-generated vs. professor-written).

Four LLMs (Gemini 2.5, GPT-4, Claude Sonnet 4, and DeepSeek v3.2) were selected because they represent the most advanced publicly available generative models at the time of the study (prompts submitted in August 2025), encompassing diverse architectures, training paradigms, and user interfaces. Their inclusion enabled a balanced assessment across distinct LLM frameworks. A pulmonology professor with more than 25 years of clinical experience provided physician-generated responses. The physician is actively engaged in clinical practice, holds an administrative role, and is involved in faculty development and educator training activities. Because a single expert served as the human comparator, this benchmark reflects the performance of one highly experienced pulmonologist rather than that of physicians in general.

Each of the six raters evaluated the anonymized answers of the five respondents (four LLMs and one human professor). Ratings were assigned on a five-point Likert scale (1=lowest, 5=highest) based on the designated evaluation criteria for each group. None of the raters had any conflict of interest with either the professor or the LLM selection process. All evaluations were performed independently, and raters were blinded to the identity of each response source to ensure unbiased judgment based solely on content quality.

Conceptual Alignment of Evaluation Criteria

The three criteria used by layperson raters-perceived reliability, clarity, and practical usefulness-were designed to conceptually parallel the expert criteria of accuracy, clarity, and scientific adequacy, respectively. Perceived reliability corresponds to accuracy, as both reflect the extent to which information is regarded as correct and trustworthy, albeit from different perspectives: experts assess factual correctness, whereas laypersons judge credibility and the confidence it inspires. Clarity was retained consistently across groups, reflecting the mutual importance of linguistic simplicity and structural transparency. Finally, practical usefulness parallels scientific adequacy: experts evaluate the scientific rigor and completeness of the answer, whereas laypersons perceive the same quality in terms of real-world applicability and utility.

The criteria used for layperson evaluations were adapted from the methodological framework applied in the study by Charide et al.[14], which focused on public and user-centered assessment approaches in health communication research. In parallel, the expert evaluation criteria were determined following the methodological approach used by Akçay et al.[15] in their study evaluating the accuracy and reliability of ChatGPT 4's responses in thoracic surgery. This mapping ensures conceptual equivalence across evaluator groups while maintaining relevance to their respective levels of expertise. Because expert accuracy (factual correctness) and lay perceived reliability are related but non-identical constructs, inter-rater reliability is reported separately for experts and laypersons, and the pooled coefficient is interpreted with caution.

Procedural Neutrality and Data Collection

The questions were submitted to the LLMs by an independent engineer specializing in software development, who is one of the authors of this article. This engineer had no background in medicine and no prior acquaintance or conflict of interest with the participating pulmonology professor. Each LLM was reset to a new session before being questioned to prevent memory retention or contextual bias, ensuring that each

interaction represented a clean, independent exchange. All prompts were entered using the same computer, under identical network conditions and internet speed, to standardize response latency and interface performance. The engineer took no part in the writing, scoring, or evaluation stages and did not participate as either an expert or layperson rater, thereby ensuring full procedural neutrality and minimizing potential bias throughout data collection.

Statistical Analysis

Inter-rater reliability was assessed using intraclass correlation coefficients (ICC) to quantify agreement among raters. ICC values estimate the proportion of variance in ratings that is attributable to true score differences rather than to random error or rater subjectivity. Both single-measure and average-measure ICCs (two-way random effects model, absolute agreement type) were computed to capture individual and group-level reliability[16].

Interpretation of ICC values followed established benchmarks, with values below 0.20 indicating slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.00 almost perfect agreement[17].

Comparative analyses across the five response sources were performed using repeated-measures one way analysis of variance (ANOVA) because each rater provided scores for all sources, resulting in a within-subjects (dependent) data structure. This approach allowed comparison of mean ratings across conditions while accounting for the interdependence of repeated observations from the same raters. Bonferroni-adjusted pairwise comparisons were applied to control for type I error in multiple testing. To corroborate these findings and to address potential deviations from normality, Friedman tests were additionally conducted as non-parametric equivalents. The unit of analysis was the per-question mean across raters (n=37), with response source as the within-subject factor (degrees of freedom=4, 144). Sphericity was assessed (Greenhouse-Geisser), and a linear mixed-effects model with crossed random effects for questions and raters was acknowledged as a valid alternative.

To ensure robustness and to explore subgroup consistency, three complementary analytic models were applied: (1) all raters (E1-E3, L1-L3), (2) experts only (E1-E3), and (3) laypersons only (L1-L3). For each model, Bonferroni-adjusted pairwise comparisons were used to compare the five response sources across the three evaluation criteria. In these analyses, accuracy for expert raters was conceptually aligned with perceived reliability for layperson raters; clarity was shared by both groups; and scientific adequacy for expert raters was aligned with practical usefulness for layperson raters.

The primary analyses included all six raters, with subgroup analyses conducted separately for expert and layperson raters. Statistical analyses were conducted using IBM SPSS Statistics version 26 (IBM Corp., Armonk, NY, USA), with recomputation and verification performed using Python and the pingouin package.

Results

Inter-rater Reliability

Inter-rater reliability was assessed using a two-way random-effects model (absolute agreement). As summarized in Table 1, single-measure ICCs ranged from 0.09 to 0.12 and average-measure ICCs ranged from 0.36 to 0.45 across accuracy/perceived reliability, clarity, and scientific adequacy/practical usefulness (all $p<0.001$).

According to commonly applied interpretative frameworks, these values correspond to slight agreement for single measures and to fair-to-moderate agreement for average measures, indicating variability in individual ratings but improved consistency when scores were aggregated. Differences in scoring patterns across evaluators and domains were statistically significant ($p<0.001$), reflecting heterogeneity in perception rather than systematic measurement error.

Overall, these findings demonstrate that while individual-level agreement was limited, aggregated ratings showed greater consistency, underscoring the subjective nature of communication-based evaluations and the influence of evaluator background on inter-rater agreement metrics.

Summary of Evaluator Scores

Across evaluation domains, the relative ranking of mean scores remained stable, with LLM-generated responses generally receiving higher ratings than physician-generated responses. Differences in mean ratings across models and evaluator groups reflected variability in perceived quality rather than isolated or extreme scoring behavior. The observed score distributions indicate coherent, interpretable evaluation patterns across all six raters and support the internal consistency of the comparative assessment framework.

Table 1. Inter-rater reliability (intraclass correlation coefficients), by rater group

Criterion	Single ICC (95% CI)	Average ICC (95% CI)	Single	Average
All raters (n=6)				
Accuracy/perceived reliability	0.087 (0.04, 0.14)	0.364 (0.20, 0.50)	slight	fair
Clarity	0.118 (0.07, 0.18)	0.445 (0.31, 0.56)	slight	moderate
Scientific adequacy/practical usefulness	0.117 (0.07, 0.17)	0.442 (0.31, 0.56)	slight	moderate
Experts only (E1-E3)				
Accuracy/perceived reliability	0.194 (0.10, 0.29)	0.419 (0.25, 0.55)	slight	moderate
Clarity	0.161 (0.07, 0.25)	0.365 (0.20, 0.51)	slight	fair
Scientific adequacy/practical usefulness	0.207 (0.10, 0.32)	0.439 (0.25, 0.58)	slight	moderate
Laypersons only (L1-L3)				
Accuracy/perceived reliability	0.096 (0.00, 0.20)	0.243 (0.01, 0.43)	slight	fair
Clarity	0.148 (0.05, 0.25)	0.342 (0.14, 0.50)	slight	fair
Scientific adequacy/practical usefulness	0.192 (0.10, 0.29)	0.415 (0.24, 0.55)	slight	moderate

Accuracy (experts)/perceived reliability (laypersons), scientific adequacy (experts)/practical usefulness (laypersons) Two-way random-effects model, absolute agreement; 185 rated items. Bands per Landis and Koch[17], ICC: Intraclass correlation coefficients, CI: Confidence interval

When all six raters were included, all four LLMs received higher mean ratings across all evaluation criteria than did physician-generated responses. Among the LLMs, Gemini 2.5 and GPT-4 achieved the highest overall mean scores, followed by DeepSeek v3.2 and Claude Sonnet 4. Physician-generated responses consistently received lower average ratings, with the largest differences observed in the scientific adequacy domain. Measures of dispersion were modest across all models, indicating a relatively consistent pattern of scoring among evaluators (Table 2a).

Among expert raters, the single pulmonologist was rated among the highest, rather than below the models; the physician received an accuracy/perceived reliability mean (4.66), nearly identical to Gemini 2.5, which had the highest mean score (4.67), and a scientific adequacy mean (4.13) comparable to or higher than most LLMs. Clarity scores were similar across sources. Standard deviations were small, indicating consistent expert judgments (Table 2b).

Among layperson raters, differences in mean ratings across response sources were more pronounced. Across all evaluation criteria, LLM-generated responses received higher mean scores than physician-generated responses, with the largest differences observed in accuracy and clarity. Gemini 2.5 achieved the highest and most consistent mean ratings among the LLMs. In contrast, ratings assigned to physician-generated responses showed greater dispersion, reflected by higher standard deviations, indicating increased variability in evaluations by non-expert participants (Table 2c).

When mean ratings across all 37 questions were examined within each evaluation domain (accuracy and perceived reliability, clarity, and scientific adequacy or practical usefulness), LLM-generated responses generally showed higher mean scores, whereas physician-generated responses tended to cluster toward lower values. Overlapping data points indicate identical or highly similar scores assigned by different models to specific questions, reflecting comparable scoring patterns across models rather than isolated deviations. These distributions provide a visual summary of score alignment across questions and evaluation domains (Figure 1).

Comparative Analyses

A repeated-measures ANOVA demonstrated a significant main effect of response source across all rater configurations. When all six raters were included, statistically significant differences were observed among the five response sources for all three evaluation criteria ($p<0.001$). Differences were most pronounced for accuracy and perceived reliability, both of which showed the largest effect size (partial $\eta^2=0.26$). For all raters: accuracy/perceived reliability, $F(4.144)=12.84$, partial $\eta^2=0.263$; clarity, $F(4.144)=10.47$, partial $\eta^2=0.225$; scientific adequacy/practical usefulness, $F(4.144)=11.89$, partial $\eta^2=0.248$ (all $p<0.001$). Subgroup analyses for expert and layperson raters are presented in Supplementary Tables 1-3.

When analyses were restricted to expert pulmonologists, significant effects of response source remained for accuracy

Table 2a. Mean (± SD) scores by response source-all raters			
Response source	**A/PR (mean ± SD)**	**C (mean ± SD)**	**SA/PU (mean ± SD)**
Gemini 2.5	4.69±0.54	4.70±0.60	4.43±0.68
GPT-4	4.51±0.61	4.66±0.57	4.22±0.71
DeepSeek v3.2	4.38±0.87	4.48±0.75	4.18±0.80
Claude Sonnet 4	4.38±0.70	4.57±0.60	4.10±0.70
Professor (physician)	**4.24±0.97**	**4.26±0.97**	**3.93±0.92**
SD: Standard deviation, A/PR: Accuracy (experts)/perceived reliability (laypersons), C: Clarity, SA/PU: Scientific adequacy (experts)/practical usefulness (laypersons)			

Table 2b. Mean (± SD) scores by response source-experts (E1-E3)			
Response source	**A/PR (mean ± SD)**	**C (mean ± SD)**	**SA/PU (mean ± SD)**
Gemini 2.5	4.67±0.51	4.72±0.51	4.20±0.63
GPT-4	4.48±0.60	4.68±0.51	3.92±0.61
DeepSeek v3.2	4.52±0.55	4.61±0.51	4.00±0.69
Claude Sonnet 4	4.39±0.61	4.62±0.52	3.80±0.62
Professor (physician)	**4.66±0.51**	**4.59±0.58**	**4.13±0.65**
SD: Standard deviation, A/PR: Accuracy (experts)/perceived reliability (laypersons), C: Clarity, SA/PU: Scientific adequacy (experts)/practical usefulness (laypersons)			

Table 2c. Mean (± SD) scores by response source-laypersons (L1-L3)			
Response source	**A/PR (mean ± SD)**	**C (mean ± SD)**	**SA/PU (mean ± SD)**
Gemini 2.5	4.71±0.56	4.68±0.69	4.66±0.65
GPT-4	4.55±0.61	4.64±0.63	4.52±0.67
DeepSeek v3.2	4.23±1.08	4.35±0.92	4.35±0.87
Claude Sonnet 4	4.38±0.79	4.51±0.67	4.40±0.66
Professor (physician)	**3.83±1.13**	**3.93±1.16**	**3.74±1.09**
SD: Standard deviation, A/PR: Accuracy (experts)/perceived reliability (laypersons), C: Clarity, SA/PU: Scientific adequacy (experts)/practical usefulness (laypersons)			

(perceived reliability) and for scientific adequacy or practical usefulness, whereas differences in clarity were of smaller magnitude. In contrast, among layperson raters, significant differences among response sources were consistently observed for all evaluation criteria ($p<0.001$), with larger effect sizes indicating greater differentiation in mean ratings.

Across subgroup analyses, the direction and magnitude of differences varied by rater expertise, with LLM advantages most evident among lay raters. Detailed ANOVA results for each configuration are provided in Supplementary Tables 1-3.

Given the significant main effects observed, Bonferroni-adjusted post-hoc tests were performed to determine which specific model comparisons accounted for these differences. Detailed pairwise statistics, including mean differences, 95% confidence intervals, t-values, degrees of freedom, Cohen's dz, and Bonferroni-adjusted p-values, are provided in Supplementary Tables 4-6.

When all raters were included, Bonferroni-adjusted pairwise comparisons showed that Gemini 2.5 and GPT-4 were rated significantly higher than the physician on all three criteria ($p\leq0.015$), whereas DeepSeek v3.2 did not differ significantly from the physician on any criterion, and Claude Sonnet 4 differed only for clarity ($p=0.010$). Among the models, Gemini 2.5 scored significantly higher than DeepSeek v3.2 and Claude Sonnet 4 on accuracy/perceived reliability and scientific adequacy/practical usefulness, and higher than GPT-4 on those two criteria, with no significant difference in clarity (Table 3a).

Among expert raters, no LLM was rated significantly higher than the physician on any criterion after Bonferroni adjustment; however, the physician was rated significantly higher than Claude Sonnet 4 on accuracy/perceived reliability ($p=0.046$) and scientific adequacy ($p=0.004$). Experts rated all sources similarly for clarity. These findings indicate that, among professional evaluators, the experienced pulmonologist was judged to be at least as accurate and scientifically adequate as the strongest LLMs (Table 3b).

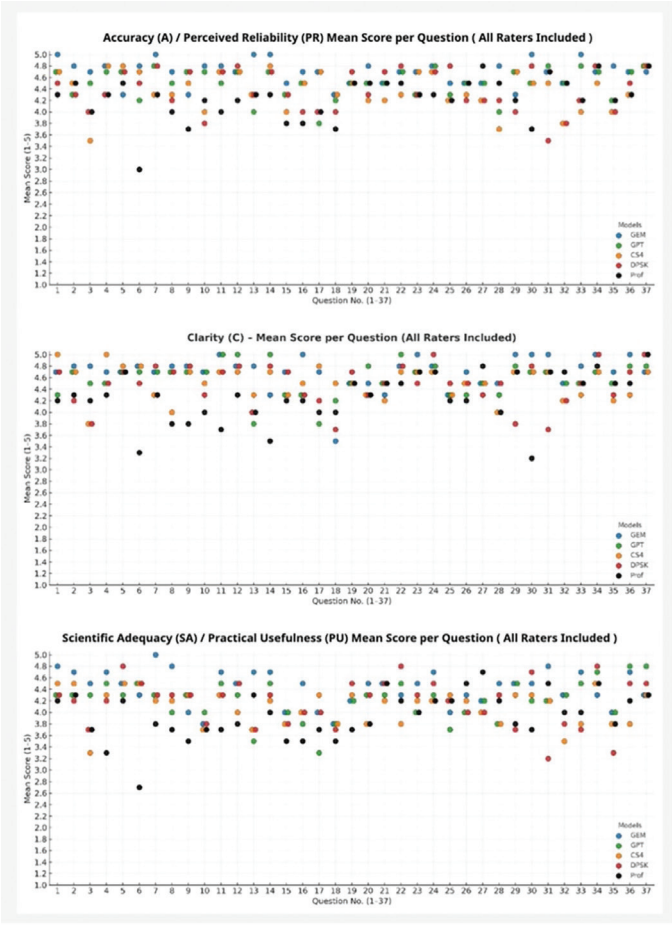

Figure 1. Distribution of mean scores per question across all raters

For each of the 37 questions, the mean Likert score (1-5, averaged across all six raters) is plotted separately for accuracy/perceived reliability, clarity, and scientific adequacy/practical usefulness. The x-axis represents the question number (1-37), and the y-axis represents the mean score. Color key: Gemini 2.5, GPT-4, Claude Sonnet 4, DeepSeek v3.2, and the physician. Overlapping markers indicate identical or closely similar means

Layperson raters' evaluations revealed statistically significant differences between response sources across all evaluation domains ($p<0.001$). Across measures of accuracy and perceived reliability, clarity, and scientific adequacy or practical usefulness, LLM-generated responses received higher mean ratings than physician-generated responses, with Gemini 2.5 and GPT-4 showing slightly higher mean ratings than Claude Sonnet 4 and DeepSeek v3.2. These differences were most pronounced for accuracy and for perceived reliability and clarity. Overall, lay evaluators consistently assigned higher ratings to LLM-generated content across domains, reflecting differences in perceived quality and usability rather than isolated rating effects (Table 3c).

Discussion

The present study demonstrated slight single-measure inter-rater reliability (ICC 0.09-0.12) and fair-to-moderate average-measure (ICC 0.36-0.45) inter-rater reliability, which is acceptable within the context of subjective, communication-focused assessments. Across evaluation domains, LLM responses received higher mean ratings than the physician did overall; however, this pattern was driven by lay raters. Expert raters rated the physician among the highest for accuracy and scientific adequacy, and no LLM significantly exceeded the physician's performance. Answer length was also positively associated with ratings across sources (Pearson $r\approx0.73$), identifying response length/format as a potential confounder. Mean response lengths differed across sources, with Gemini 2.5 producing the longest responses (approximately 33 words), followed by DeepSeek v3.2 (approximately 21 words), GPT-4 (approximately 20 words), the physician-generated responses (approximately 18 words), and Claude Sonnet 4 (approximately 14 words). Repeated-measures ANOVA confirmed a significant main effect of response source across all criteria, which was most pronounced for accuracy and perceived reliability. The divergence between lay and expert raters underscores the strong influence of evaluator background on communication-focused assessments.

These findings are broadly consistent with prior studies examining LLM performance in patient-oriented medical communication, while also highlighting important nuances. In a comparable study evaluating responses to rheumatology-related patient questions, both rheumatologists and patients assessed LLM-generated and clinician-generated answers. Although patient ratings did not differ significantly between the two approaches in terms of comprehensiveness or readability, rheumatologists rated LLM responses lower with respect to accuracy and completeness, and preference rates differed markedly between patients and experts[13]. In contrast, our study found higher ratings for LLM-generated responses, particularly among layperson raters, whereas differences within the expert subgroup were less pronounced. Notably, no statistically significant differences were found in clarity scores between LLM and physician responses.

Table 3a. Bonferroni-adjusted pairwise comparisons (p)-all raters			
Comparison	**A/PR (p)**	**C (p)**	**SA/PU (p)**
Claude Sonnet 4 vs. DeepSeek v3.2	1.000	1.000	1.000
Claude Sonnet 4 vs. GPT-4	0.349	1.000	0.752
Claude Sonnet 4 vs. Gemini 2.5	**<0.001**	0.365	**<0.001**
Claude Sonnet 4 vs. Professor	1.000	**0.010**	0.541
DeepSeek v3.2 vs. GPT-4	0.365	**0.025**	1.000
DeepSeek v3.2 vs. Gemini 2.5	**<0.001**	**0.021**	**0.021**
DeepSeek v3.2 vs. Professor	1.000	0.218	0.136
GPT-4 vs. Gemini 2.5	**0.009**	1.000	**0.011**
GPT-4 vs. Professor	**<0.001**	**<0.001**	**0.015**
Gemini 2.5 vs. Professor	**<0.001**	**<0.001**	**<0.001**
A/PR: Accuracy (experts)/perceived reliability (laypersons), C: Clarity, SA/PU: Scientific adequacy (experts)/practical usefulness (laypersons)			

Table 3b. Bonferroni-adjusted pairwise comparisons (p)-experts only			
Comparison	**A/PR (p)**	**C (p)**	**SA/PU (p)**
Claude Sonnet 4 vs. DeepSeek v3.2	0.789	1.000	0.130
Claude Sonnet 4 vs. GPT-4	1.000	1.000	1.000
Claude Sonnet 4 vs. Gemini 2.5	**0.010**	1.000	**<0.001**
Claude Sonnet 4 vs. Professor	**0.046**	1.000	**0.004**
DeepSeek v3.2 vs. GPT-4	1.000	1.000	1.000
DeepSeek v3.2 vs. Gemini 2.5	0.404	1.000	0.699
DeepSeek v3.2 vs. Professor	1.000	1.000	1.000
GPT-4 vs. Gemini 2.5	0.095	1.000	**0.026**
GPT-4 vs. Professor	0.132	1.000	0.284
Gemini 2.5 vs. Professor	1.000	1.000	1.000
A/PR: Accuracy (experts)/perceived reliability (laypersons), C: Clarity, SA/PU: Scientific adequacy (experts)/practical usefulness (laypersons)			

Table 3c. Bonferroni-adjusted pairwise comparisons (p)-laypersons only			
Comparison	**A/PR (p)**	**C (p)**	**SA/PU (p)**
Claude Sonnet 4 vs. DeepSeek v3.2	0.844	0.595	1.000
Claude Sonnet 4 vs. GPT-4	0.351	1.000	1.000
Claude Sonnet 4 vs. Gemini 2.5	**<0.001**	0.628	**0.036**
Claude Sonnet 4 vs. Professor	**<0.001**	**<0.001**	**<0.001**
DeepSeek v3.2 vs. GPT-4	**0.021**	**0.019**	0.814
DeepSeek v3.2 vs. Gemini 2.5	**<0.001**	**0.019**	**0.013**
DeepSeek v3.2 vs. Professor	**0.007**	**0.016**	**<0.001**
GPT-4 vs. Gemini 2.5	**0.040**	1.000	1.000
GPT-4 vs. Professor	**<0.001**	**<0.001**	**<0.001**
Gemini 2.5 vs. Professor	**<0.001**	**<0.001**	**<0.001**
Bold p-values indicate p<0.05, (p) computed on per-question means (n=37) with Bonferroni correction A/PR: Accuracy (experts)/perceived reliability (laypersons), C: Clarity, SA/PU: Scientific adequacy (experts)/practical usefulness (laypersons)			

Similar patterns have been reported in other clinical contexts. An emergency medicine study comparing LLM-generated patient education responses with those produced by orthopedic clinicians found more favorable ratings for AI-generated content across multiple domains, aligning with the communication-related advantages observed in our study[18]. Likewise, investigations focusing on patient questions related to lung cancer surgery reported high ratings for both accuracy and clarity of LLM responses, while emphasizing that such systems should be used as educational or supportive tools rather than as independent clinical decision-makers[15].

In contrast, studies assessing diagnostic and triage performance present a more cautious perspective. Evaluations comparing LLMs with physicians and laypersons have shown that while LLMs may outperform non-experts and provide information perceived as more useful than conventional search engine outputs, they continue to underperform physicians in diagnostic accuracy and complex clinical reasoning tasks[19]. Similarly, a recent systematic review reported that although LLMs can achieve accuracy comparable to specialists in selected imaging-related tasks, they remain limited in patient-centered communication and clinical reasoning, reinforcing the ongoing necessity of human clinical expertise[11]. Consistent with these observations, Liu et al.[20] demonstrated that LLMs performed below senior physicians in pulmonary diagnostic tasks, highlighting the distinction between theoretical knowledge synthesis and integrative, experience-based diagnostic reasoning.

Beyond comparisons between physicians and AI systems, our findings also point to differences among LLMs. Although Gemini 2.5 achieved higher mean ratings in several domains, these differences did not consistently reach statistical significance in assessments by expert evaluators. Among laypersons, however, Gemini 2.5 was rated more favorably than other models, suggesting that communicative fluency and interpretability may play a stronger role in public perception than domain-specific accuracy alone. This observation aligns with findings from intensive care medicine, where GPT-4-based models demonstrated high accuracy but were also associated with increased computational demands and concerns regarding confidently incorrect outputs, underscoring the importance of ongoing validation prior to broader clinical adoption[21].

Multiple evaluations of medical AI systems have further identified persistent challenges, including logical inconsistencies, contextual fragmentation, and limitations in deep diagnostic reasoning[22,23]. As Turner[24] has noted, trust remains an ethical concept that is difficult to ascribe to algorithmic systems lacking moral agency. In this context, the present findings support the view that LLMs should be regarded as adjunctive tools designed to augment- rather than replace-physicians. Their greatest potential lies in supporting medical communication and information delivery, while human clinicians remain essential for clinical judgment, accountability, empathy, and holistic decision-making at the core of medical practice.

Study Limitations

This study has several limitations. First, physician responses were obtained from a single pulmonologist, which may limit the generalizability of the findings to broader clinical practice. Second, the number of evaluators was limited, restricting the extent to which the results can be considered representative of larger expert and lay populations. Third, all primary outcome measures were based on subjective ratings of communication quality rather than on objective assessments of factual correctness or clinical validity. In addition, although prompts were predefined and AI-generated responses were archived at the time of evaluation, LLMs are continuously updated, which may affect reproducibility across different model versions. The study focused on perceived communication quality and the usefulness of information; therefore, the findings should not be interpreted as reflecting diagnostic accuracy, clinical reasoning performance, or decision-making capability. In addition, perceived accuracy was rated by humans without gold standard (e.g., guideline-based) adjudication of factual correctness; the human comparator was a single pulmonologist, limiting generalizability to physicians as a group; the question set did not undergo formal content validation; response length differed across sources and was positively associated with ratings, representing a potential format confounder; repeated-measures ANOVA, rather than mixed-effects modeling, was used for the primary comparisons. The linguistic and contextual characteristics of the questions and evaluators may also limit generalizability to other languages or health care settings.

Conclusion

This study suggests that current LLMs can generate pulmonology-related medical information that is perceived as clear and scientifically adequate by both expert and non-expert audiences. Although some models received higher

communication-related ratings, these results should be interpreted in light of the subjective nature of the evaluations and do not imply equivalence to clinical expertise. The advantage of LLMs was driven mainly by lay raters; among expert raters, the single pulmonologist was rated at least as highly as the strongest models for accuracy and scientific adequacy.

Importantly, the results underscore that LLMs are best viewed as supportive tools rather than as substitutes for physicians. Human clinicians remain essential for clinical reasoning, contextual judgment, empathy, and professional accountability. Accordingly, the integration of AI-assisted information delivery with physician oversight may enhance patient communication while preserving the central role of human expertise in medical practice.

Ethics

Ethics Committee Approval: This study did not involve patients, patient data, identifiable personal information, biological samples, or any clinical intervention. The raters evaluated anonymized written responses; therefore, ethics committee approval was not required according to institutional and national regulations.

Informed Consent: Not applicable, as the study did not involve patients or identifiable personal data.

Footnotes

Authorship Contributions

Surgical and Medical Practices: D.S.U., Ö.S.U., M.O.G., A.A., A.E.E., Concept: D.S.U., Ö.S.U., M.O.G., Design: D.S.U., Ö.S.U., Data Collection or Processing: D.S.U., Ö.S.U., B.C.K., A.A., Analysis or Interpretation: D.S.U., Ö.S.U., B.C.K., A.A., Literature Search: D.S.U., M.O.G., B.C.K., A.E.E., Writing: D.S.U., Ö.S.U.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

Supplementary Tables: https://d2v96fxpocvxx.cloudfront.net/cf9d60d6-523c-458a-a2e6-78728d3ffbb0/content-images/e99fe721-803e-4f4f-9249-125ccbd841b5.pdf

References

1. Lie HC, Juvet LK, Street RL Jr, et al. Effects of physicians' information giving on patient outcomes: a systematic review. J Gen Intern Med. 2022;37:651-63.
2. Zakaria M, Mazumder S, Faisal HM, et al. Physician communication behaviors on patient satisfaction in primary care medical settings in Bangladesh. J Prim Care Community Health. 2024;15:21501319241277396.
3. Sharkiya SH. Quality communication can improve patient-centred health outcomes among older patients: a rapid review. BMC Health Serv Res. 2023;23:886.
4. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel). 2021;9:1740.
5. Daraz L, Dogu C, Houde V, Bouseh S, Morshed KG. Assessing credibility: quality criteria for patients, caregivers, and the public in online health information-a qualitative study. J Patient Exp. 2024;11:23743735241259440.
6. Sonig A, Deeney C, Hurley ME, et al. What patients and caregivers want to know when consenting to the use of digital behavioral markers. NPP Digit Psychiatry Neurosci. 2024;2:19.
7. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics. 2017;37:505-15.
8. Neri E, Coppola F, Miele V, Bibbolino C, Grassi R. Artificial intelligence: who is responsible for the diagnosis? Radiol Med. 2020;125:517-21.
9. Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat Biotechnol. 2017;35:342-6.
10. Hosny A, Parmar C, Coroller TP, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. PLoS Med. 2018;15:e1002711.
11. Shen J, Zhang CJP, Jiang B, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR Med Inform. 2019;7:e10010.
12. Kusal S, Patil S, Choudrie J, Kotecha K, Mishra S, Abraham A. AI-based conversational agents: a scoping review from technologies to future directions. IEEE Access. 2022;10:92337-56.
13. Ye C, Zweck E, Ma Z, Smith J, Katz S. Doctor versus artificial intelligence: patient and physician evaluation of large language model responses to rheumatology patient questions in a cross-sectional study. Arthritis Rheumatol. 2024;76:479-84.
14. Charide R, Stallwood L, Munan M, et al. Knowledge mobilization activities to support decision-making by youth, parents, and adults using a systematic and living map of evidence and recommendations on COVID-19: protocol for three randomized controlled trials and qualitative user-experience studies. Trials. 2023;24:27.
15. Akçay O, Öztürk Ö, Acar T, Gürsoy S. Accuracy and reliability of ChatGPT in answering patient questions about lung cancer and its surgery: an expert panel evaluation by thoracic surgeons. J Cancer Educ. 2026;41:477-82.
16. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15:155-63.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74.
18. Liu J, Segal K, Daher M, et al. Artificial intelligence versus orthopedic surgeons as an orthopedic consultant in the emergency department. Injury. 2025;56:112297.

19. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. medRxiv [Preprint]. 2023:2023.01.30.23285067.
20. Liu X, Liu H, Yang G, et al. A generalist medical language model for disease diagnosis assistance. Nat Med. 2025;31:932-42.
21. Workum JD, Volkers BWS, van de Sande D, et al. Comparative evaluation and performance of large language models on expert level critical care questions: a benchmark study. Crit Care. 2025;29:72.
22. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30:2613-22.
23. Griot M, Hemptinne C, Vanderdonckt J, Yuksel D. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun. 2025;16:642.
24. Turner JH. Triangle of trust in cancer care? The physician, the patient, and artificial intelligence chatbot. Cancer Biother Radiopharm. 2023;38:581-4.