Recent Progress in Sciences

Science Publication Inc.

Original Article Open Access

Bilingual performance of ChatGPT, Gemini, and DeepSeek in asthma, allergy, and respiratory infection queries

Mohammed Sallam [1,2,3,4,5], Adrian Stanley [4,6], Johan Snygg [4,7,8], Hasanain Al-Shakerchi [4,9], Omar Al Atragchi [10], Rania Abusamra [4,11], Malik Sallam [12,13,*]

[1] Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates
[2] Department of Management, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates
[3] Department of Management, School of Business, International American University, Los Angeles, CA 90010, United States of America
[4] College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences (MBRU), Dubai P.O. Box 505055, United Arab Emirates
[5] Department of Pharmacology and Therapeutics, College of Medicine and Health Sciences, United Arab Emirates University (UAEU), Al Ain P.O. Box 17666, United Arab Emirates
[6] Department of Clinical Management, Mediclinic Middle East, Dubai P.O. Box 123812, United Arab Emirates
[7] Department of Management, Mediclinic City Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates
[8] Department of Anesthesia and Intensive Care, University of Gothenburg, Sahlgrenska Academy, 41345 Gothenburg, Sweden
[9] Department of Internal Medicine, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates
[10] Department of Family Medicine, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates
[11] Department of Pediatric Pulmonology, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai P.O. Box 505004, United Arab Emirates
[12] Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Amman 11942, Jordan
[13] Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Amman 11942, Jordan
[*] Correspondence: Malik Sallam. email: malik.sallam@ju.edu.jo

Received: September 15, 2025; Revised: December 10, 2025; Accepted: January 16, 2026; Published: January 23, 2026

Abstract

Generative artificial intelligence (genAI) models are rapidly being adopted for health information delivery. Nevertheless, systematic cross-linguistic evaluations of their clinical reliability—particularly in high-burden conditions such as asthma, allergy, and respiratory tract infections (RTIs)—remain limited. The aim of this study was to compare the English and Arabic performance of ChatGPT-4o, Gemini, and DeepSeek in responding to common asthma, allergy, and RTI queries using a validated clinical assessment framework. A bilingual evaluation was conducted using 30 frequently asked questions (FAQs) related to asthma, allergy, and RTIs. Each question was submitted in English and Arabic to ChatGPT-4o, Gemini, and DeepSeek. Responses were evaluated independently by three bilingual clinical experts using the CLEAR framework for Completeness, Accuracy, and Relevance of the generated content. Inter-rater reliability was assessed using intraclass correlation coefficients (ICCs). Language and model comparisons were analyzed using non-parametric Kruskal-Wallis and Mann-Whitney U tests. The study followed the METRICS reporting guideline for genAI in healthcare. ChatGPT-4o consistently outperformed Gemini and DeepSeek across all CLEAR dimensions and the two languages. In English, the mean CLEAR scores were: ChatGPT-4o: 3.90, Gemini: 2.50, DeepSeek: 2.09. In Arabic, ChatGPT-4o again scored highest (3.63), followed by Gemini (2.38) and DeepSeek (1.84). All inter-model differences were statistically significant ($p < 0.001$). Inter-rater reliability was excellent across dimensions: ICC for completeness = 0.858, accuracy = 0.917, relevance = 0.950 (all $p < 0.001$), confirming strong consistency and validity in scoring. Within each genAI model, English outputs significantly outperformed Arabic in completeness,

doi:10.70462/rps.2026.3.001

accuracy, relevance, and the overall CLEAR score. Domain-wise, asthma queries achieved the highest performance across models and languages, while allergy queries showed the lowest accuracy. ChatGPT-4o demonstrated superior bilingual performance, while Gemini and DeepSeek exhibited significant limitations, particularly in Arabic. These findings highlight persistent language-based disparities in genAI health outputs. Rigorous cross-linguistic evaluation and domain-specific fine-tuning are essential to ensure safe and equitable deployment of genAI tools in global health communication.

Keywords Hypersensitivity, natural language processing, health communication, multilingualism, health literacy

1. Introduction

Nearly three decades ago, a quiet revolution began in the way patients accessed health information—one that gradually shifted the center of gravity from clinician-mediated encounters to patient-driven digital exploration [1, 2]. As internet access expanded and health websites proliferated, individuals increasingly turned to online platforms for answers to their health concerns [3-5]. Medscape, launched in 1996, marked the digitization of medical knowledge for clinicians while WebMD, founded in 1998 and expanded to become the first major platform to provide medical information directly to consumers [6]. Similarly, the National Health Service (NHS) launched a website in 1999, laying the foundation for national online health advice being the UK's most visited health website, with over 50 million monthly users [7]. The early 2000s ushered in a new era of participatory health information-seeking, marked by the rise of user-generated forums, such as "PatientsLikeMe", and the growing influence of search engines like Google, which quickly became the default starting point for health-related queries [8-10]. Landmark surveys by the Pew Research Center revealed that, as early as the mid-2000s, more than 70% of American internet users routinely searched for health information online—a finding that signaled a profound and sustained shift in how patients engage with medical knowledge [11].

The following decade saw the convergence of mobile technology and artificial intelligence (AI). Smartphones made health applications, symptom checkers, and telehealth services both portable and pervasive [12-14]. This digital revolution was further catalyzed by the coronavirus disease 2019 (COVID-19) pandemic, which dramatically accelerated reliance on virtual health platforms, as millions avoided clinical settings and instead consulted digital tools and healthcare chatbots [15-17]. Since 2022, generative AI (genAI) tools such as ChatGPT have emerged as a new frontier in health information seeking [18]. With their ability to generate fluent, conversational responses to complex medical queries across multiple languages—albeit with variable accuracy—genAI models are rapidly redefining the patient's first point of contact with the healthcare system without stepping into a clinic or consulting a professional [18-23].

The digital age has thus already redefined how individuals engage with their health, with search engines, online forums, and symptom-checker applications becoming routine fixtures in patient self-management [24]. Yet the emergence of genAI marks a key inflection point in digital health. Unlike traditional tools, genAI models—such as ChatGPT, DeepSeek, Gemini, Grok, Meta AI, and Claude, among others—can now deliver personalized, multilingual, and context-aware responses to complex medical queries [18, 25, 26]. These genAI models provide conversational medical guidance directly to users; nevertheless, their widespread adoption necessitates critical reflection on issues of reliability, accountability, and patient safety [27, 28]. For many patients, particularly those managing chronic but episodic conditions such as asthma and allergic diseases, these tools offer a tempting alternative to formal clinical consultation, especially in moments when symptoms feel familiar, non-urgent, or routine [29-31]. This shift in health-seeking behavior is far from trivial. Asthma, allergic disorders, and respiratory tract infections (RTIs) are among the most prevalent non-communicable and communicable conditions globally, transcending demographic and economic boundaries [32-35].

Allergies and asthma are characterized by fluctuating symptoms and signs, and the ever-present risk of sudden exacerbations [36-38]. Patients often navigate between periods of control and acute flare-ups, making decisions about when—and whether—to engage healthcare services based on subjective symptom appraisal, previous experiences, and access to care [39]. In this context, a genAI model that provides timely, linguistically accessible guidance offers a form of immediacy and personalization unmatched by static websites or conventional health brochures. The potential implications for both patient autonomy and clinical oversight would be considerable.

The appeal of genAI applications in patient self-management lies in its immediacy, convenience, and adaptability [40-42]. The genAI models can synthesize vast medical knowledge into tailored, comprehensible responses within seconds. For individuals with asthma, genAI can offer guidance on proper inhaler technique, identify environmental triggers, or deliver stepwise action plans aligned with current guidelines [43]. For those with allergic rhinitis or mild food allergies, genAI tools may provide quick reference points for allergen avoidance, interpretation of diagnostic results, or over-the-counter medication use [44, 45]. The conversational nature of genAI allows for interactive follow-up questions, approximating—albeit imperfectly—the dynamics of a clinical consultation [46]. Accessibility is another critical advantage: these tools can operate on smartphones and low-bandwidth connections, broadening reach to populations with limited access to healthcare professionals [47, 48]. For non-English-speaking patients, multilingual capabilities offer a bridge over linguistic barriers that often constrain comprehension of conventional medical resources [49]. In principle, these attributes of genAI may help narrow disparities in healthcare access, support patient autonomy, and promote adherence to evidence-based self-care [50].

Yet the promise of genAI is inextricably linked to the potential for harm [18]. The very qualities that make these tools appealing—linguistic fluency, authoritative tone, and interactive engagement—can also lend false confidence to inaccurate or incomplete content [51]. In the context of asthma care, an erroneous suggestion about medication titration, inhaler usage, or the need for emergent evaluation could result in poor disease control, avoidable exacerbations, or even life-threatening outcomes [52]. Similarly, in allergy management, misleading advice about allergen avoidance or risk severity could expose individuals to preventable harm or dangerous reassurance [53]. These risks are amplified by a structural vulnerability since most genAI models are not intrinsically capable of distinguishing high-quality, evidence-based guidance from outdated, biased, or unverified sources [51, 54]. Unless rigorously curated or clinically constrained, genAI outputs may mirror the noise of the internet more than the signal of peer-reviewed medical evidence. For patients without clinical training, the polished language of genAI can mask these deficiencies, blurring the line between sound guidance and subtle misinformation [55, 56].

Across the healthcare landscape, genAI is rapidly moving from conceptual promise to practical application [57]. Emerging evidence suggests that large language models (LLMs) can rival, and at times surpass, students and physicians in standardized knowledge assessments [58-60]. Yet their role in direct patient engagement—particularly in self-management contexts—remains the subject of ongoing debate [18, 41, 61]. While some health systems are cautiously integrating AI-driven chatbots into patient portals, others warn against premature, unsupervised use due to unresolved concerns about privacy, reliability, accountability, and patient safety [62-64]. One underappreciated limitation of genAI is the potential for language bias [65]. A majority of Western-based LLMs are trained predominantly on English-language data, with variable representation of other languages, including Arabic [66-68]. Consequently, performance in non-English languages may be less accurate, less nuanced, or more prone to omission of culturally relevant context as shown in multiple recent studies in the context of healthcare [69-72]. This is particularly relevant in Arab countries, where asthma and allergic diseases are prevalent and health engagement often occurs in Arabic.

The SNAPSHOT epidemiological program, conducted across Egypt, Turkey, and a Gulf cluster (Kuwait, Saudi Arabia, and the United Arab Emirates (UAE)), found that asthma prevalence ranged from 4.4% to 7.6%, while the prevalence for allergic rhinitis was 3.6% in Egypt, and 6.4% in the Gulf cluster [73, 74]. These findings highlight the urgent need for culturally sensitive, linguistically accurate tools—particularly bilingual genAI—to support chronic disease self-management in Arabic-speaking populations. Recent studies have shown that various genAI models frequently underperformed in Arabic compared to English across a range clinical domains [22, 75-77], while a recent study showed that the performance was excellent in ophthalmology [78]. Discrepancies in genAI performance across languages risk amplifying existing inequities in access to accurate medical information—undermining its promise as a tool for health equity. Despite growing concern, few studies have systematically benchmarked the accuracy of AI-generated medical content across Arabic and English within the same clinical context [22, 75-78]. This gap is particularly concerning in asthma and allergy care, where clear, evidence-based guidance is essential. Inconsistent or imprecise responses in a patient's native language may not only compromise clinical safety but also erode trust in emerging digital health tools.

As public adoption of genAI accelerates, rigorous evaluation of these tools for patient use is urgently needed. Assessments must be clinically relevant, methodologically transparent, and attuned to the linguistic and cultural contexts in which they are deployed. Without such evidence, the healthcare community risks either endorsing potentially harmful technologies or unduly withholding tools that could enhance patient autonomy and outcomes. To address this gap, the current study aimed to evaluate the performance of three widely accessible and used genAI models—ChatGPT, Gemini, and DeepSeek—when responding to frequently asked questions (FAQs) about asthma, allergy, and RTIs. The primary objective was to assess the completeness, accuracy, and relevance of model responses across the two languages. Secondary aims included identifying nuanced language-related performance gaps and to inform clinicians, policymakers, and AI developers on the safe, equitable deployment of genAI for patient self-management.

2. Methods

2.1Study design

This descriptive cross-sectional study adhered to the METRICS checklist to ensure transparency and rigor in AI-based health research [79]. We evaluated three prevalent clinical conditions: (1) asthma, (2) general allergies, and (3) RTIs which are commonly subject to online patient self-search when symptoms are perceived as non-urgent. Using lay-language prompts, we queried three genAI models (ChatGPT-4o, DeepSeek-V3, and Google Gemini Flash 2.5) in both English and Arabic to assess their capacity to deliver linguistically consistent and clinically relevant content. The queries spanned symptom recognition, triggers, prevention, and management. Outputs were evaluated by bilingual three clinical experts for completeness, accuracy, and relevance based on the validated CLEAR tool for assessing the quality of AI-generated content in healthcare [80]. No ethical permission was necessary, as the study relied solely on publicly available platforms and involved no patient data.

A total of 30 matched bilingual queries were selected to enable paired comparison of genAI responses in English and Arabic. Based on standard sample size calculations for paired means, a sample of 28 pairs achieves 80% power to detect a mean difference of 0.5 with a standard deviation of 0.9 at a two-sided alpha of 0.05. This effect size reflects a moderate and clinically meaningful difference in clarity and accuracy scores on a 5-point scale. The chosen sample size of 30 thus allowed sufficient power while ensuring balanced representation across the three clinical domains: asthma, allergy, and respiratory infections. Calculations for the minimum sample size of queries were performed using Statulator Sample Size Calculator for Comparing Paired Differences [81].

2.2 Query development and selection

A total of 30 layperson-oriented queries were developed to evaluate genAI performance across three prevalent, self-managed conditions: asthma, general allergies, and RTIs. Each category included 10 queries, selected based on four criteria: frequency in online health information–seeking behavior, clinical relevance for non-urgent scenarios, clarity of language, and alignment with educational and preventive themes rather than acute care. Specifically, the queries were intentionally framed to reflect non-urgent, informational, and educational patient questions, rather than high-stakes diagnostic, therapeutic, or clinical decision-making scenarios. Accordingly, the study was designed to evaluate the quality of genAI outputs as tools for health education and self-management support, not as substitutes for professional medical care or clinical decision support systems.

The queries were developed through consensus by the first and senior authors (MoS and MaS)—both bilingual Arabic-English health professionals with complementary expertise. The first author is a pharmacist with over two decades of experience in patient-centered health communication, and the senior author is a consultant in clinical microbiology and immunology. The initial queries were formulated in Arabic to reflect the primary language of health engagement in the target region, then translated into English by the senior author. This was followed by a reverse translation process and final review by both authors to ensure conceptual and linguistic consistency across both languages. The queries were designed to maximize conceptual, linguistic, and functional diversity rather than surface lexical variation alone. Additionally, the queries were intentionally distributed across multiple dimensions of patient inquiry, including disease definition and etiology, symptom recognition, risk perception, triggers and prevention, diagnosis, treatment principles, prognosis, and guidance on when to seek medical care. This approach ensured coverage of distinct cognitive and explanatory demands typically posed by patients, ranging from factual recall to interpretive and decision-support–oriented questions.

To anchor the queries in real-world relevance, content was informed by publicly available health information from trusted sources. Asthma-related questions were derived from the Global Initiative for Asthma (GINA) FAQs and the American Academy of Allergy, Asthma & Immunology (AAAAI) [82, 83]; allergy queries reflected common layperson concerns published by WebMD and the AAAAI [84, 85]; and RTIs queries were adapted from the World Health Organization (WHO) (covering coronavirus disease 2019 (COVID-19) and influenza) and NHS Borders (focusing on respiratory syncytial virus (RSV)) [86-88]. The final set of queries are shown in Table 1.

2.3 GenAI models evaluated and prompting protocol

Three widely used genAI models were evaluated: ChatGPT-4o (OpenAI, subscription tier, default settings), DeepSeek-V3 (public release, default settings), and Gemini 2.5 Flash (Google, default settings) [89-91]. All interactions were conducted through the publicly available chat-based user interfaces rather than via application programming interfaces (APIs), reflecting the most common mode of access for layperson users seeking health information. All models were accessed on the same day (15 July 2025) to minimize variability due to model updates. Default system parameters were used for all genAI models, with no modification of temperature, sampling strategies, or other generation settings, consistent with typical end-user behavior. Each of the 30 patient-centered queries was submitted in both English and Arabic, for a total of 180 unique outputs (30 queries × 2 languages × 3 genAI models).

To avoid bias from prior context or prompt history, a new chat session was initiated for each query. Specifically, each query was submitted in a fresh, single-turn session with no prior conversational context, intentionally simulating a first-contact patient interaction (e.g., a user entering a single health question into a chatbot or search interface). This design optimized comparability across genAI models and languages by eliminating contextual carryover effects. The order of language presentation (English vs. Arabic) was randomized for each model to control for order effects. Queries were entered verbatim in each language, without any prompt engineering or manual refinement, to reflect typical layperson interaction. No pre-processing or post-processing was applied to the genAI responses. Full, unedited model responses were saved and are available in the (Appendix).

Table 1 The final set of layperson-formulated frequently asked questions (FAQs) on asthma, allergy, and respiratory tract infections used for generative AI (genAI) evaluation.

Asthma	Allergy	RTIs
What is asthma and what causes it	What are allergies and why do they happen	What is COVID-19 and how does it spread
Can asthma go away on its own	How can I tell if I have allergies or a cold	What is influenza and how is it different from a cold
What are common symptoms of asthma	What are the most common allergy symptoms	What is RSV and who is most at risk
How is asthma diagnosed	What foods most often cause allergic reactions	How can I protect myself from getting COVID-19
What triggers an asthma attack	Can allergies be cured	How can I prevent getting the flu
Can exercise make asthma worse	Are allergies inherited from parents	How can I prevent RSV in babies and older adults
How should I use my inhaler correctly	What treatments are available for allergies	What are the common symptoms of COVID-19
Can asthma be controlled without medicine	Can allergies cause asthma	What are the common symptoms of influenza
Is asthma dangerous or life-threatening	How can I prevent allergic reactions	What are the symptoms of RSV infection
Can children grow out of asthma	Can someone develop allergies later in life	When should I see a doctor for a respiratory infection

RTIs: Respiratory tract infections; COVID-19: Coronavirus disease 2019; RSV: Respiratory syncytial virus.

2.4 Evaluation of genAI responses by expert raters

Each genAI-generated response was independently evaluated by three bilingual (Arabic and English) clinical experts using the CLEAR assessment tool, a validated instrument that scores content quality across three domains: completeness, accuracy, and relevance [80]. Completeness evaluated whether the response addressed the query comprehensively; accuracy assessed the factual correctness and alignment with current evidence-based guidelines; and relevance measured clarity, conciseness, and absence of misleading or extraneous information [80]. Items were scored on a 5-point Likert scale (1 = poor to 5 = excellent), and domain scores were averaged across raters. Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC). Prior to scoring, all raters received standardized instructions on use of the CLEAR framework. A reference answer key, agreed upon in advance, ensured consistency in expert judgments across languages and topics.

The CLEAR framework was selected because it is a validated instrument specifically designed to evaluate AI-generated health information, with established content validity and inter-rater reliability across diverse clinical topics [78, 92-94]. Its focus on completeness, accuracy, and relevance aligns directly with the objectives of evaluating patient-facing genAI outputs rather than clinician-level decision support.

The evaluation panel included three bilingual consultant clinicians with substantial and complementary expertise in asthma, allergy, and RTIs across adult and pediatric populations. HA, a UK-trained consultant in internal and acute medicine, has over 15 years of experience managing adult respiratory conditions across inpatient and ambulatory care. OAA, a UK-certified family medicine consultant with 19 years of clinical practice, holds accredited training in chronic disease management and has extensive experience with asthma and viral respiratory illnesses in primary care. RA, a pediatric pulmonologist with 17 years of clinical experience, trained at Cambridge and previously served at King's College Hospital, where she led services in non-invasive ventilation and complex pediatric respiratory care. This expert panel was well-positioned to evaluate the completeness, accuracy, and relevance of genAI-generated responses across age groups and clinical contexts. The raters were aware of the generating genAI model; however, evaluations were conducted independently, without inter-rater discussion. No time limits were imposed.

2.5 Statistical analysis

All statistical analyses were conducted using IBM SPSS Statistics version 26 for Windows (IBM Corp, Armonk, NY). Descriptive statistics, including means, standard deviations (SDs), and 95% confidence intervals for the mean for the error bars, were calculated for each of the CLEAR components—completeness, accuracy, relevance, and the aggregated overall CLEAR score—across language (English, Arabic), model (ChatGPT-4o, Gemini, DeepSeek), and clinical domain (asthma, allergy, RTIs). Given the ordinal nature of the Likert-scale ratings and non-normal distribution of scores (confirmed via Shapiro-Wilk tests), non-parametric tests were employed. Kruskal-Wallis H (K-W) tests were used to detect significant differences in CLEAR component scores across genAI models and across clinical domains within each language. Where significant, post-hoc comparisons were planned using Mann-Whiteny *U* (M-W) test with Bonferroni adjustment (0.017 as the cut-off for statistical significance considering the multiple comparisons). To assess language-based performance discrepancies within each model, M-W tests were conducted to compare English and Arabic responses for each model across all four CLEAR domains. Inter-rater reliability of expert assessments was evaluated using a two-way mixed-effects intraclass correlation coefficient (ICC). ICC values were interpreted according to conventional thresholds (values ≥ 0.75 indicating good agreement; ≥ 0.90, excellent agreement) [95]. Error bars reflecting the 95% CIs for each mean were plotted for visual comparison. A *p* value < 0.050 was considered statistically significant unless stated otherwise.

3. Results

3.1 Validity of assessment by the three expert raters

The inter-rater reliability across the three expert evaluators were assessed separately for each of the three CLEAR dimensions: completeness, accuracy, and relevance. Each dimension demonstrated excellent psychometric properties, supporting the reliability of the scoring framework applied to AI-generated outputs in both English and Arabic. For completeness, for average measures was also 0.858 (95% CI: 0.818 to 0.891), indicating high agreement. The corresponding single-measure ICC was 0.669 (95% CI: 0.600 to 0.731), and the F test showed statistical significance (F = 7.056; *p* < 0.001), confirming that the level of agreement between raters was substantially greater than expected by chance. For the accuracy dimension, the average-measures ICC was 0.917 (95% CI: 0.894 to 0.936), with a single-measure ICC of 0.787 (95% CI: 0.738 to 0.831). Again, the F test demonstrated significant reliability (F = 12.109, *p* < 0.001). Finally, the relevance dimension showed the highest degree of inter-rater agreement with the average-measures ICC was likewise 0.950 (95% CI: 0.936 to 0.962). The single-measure ICC was 0.864 (95% CI: 0.830 to 0.893), and the F test again confirmed the strength of this agreement (F = 20.110, *p* < 0.001, Table 2).

3.2 Performance of the GenAI models stratified per language and CLEAR component-level analysis (completeness, accuracy, and relevance)

Significant differences were observed across the three genAI models—ChatGPT-4o, DeepSeek, and Gemini—in all evaluated domains of performance (completeness, accuracy, relevance, and overall CLEAR score), both in English and Arabic responses. In English, ChatGPT-4o consistently received the highest scores across all assessment dimensions, with an overall CLEAR score mean of 3.90 ± 0.11, compared with 2.50 ± 0.18) for Gemini and 2.09 ± 0.21 for DeepSeek. Specifically, for completeness, ChatGPT-4o achieved a mean of 3.84 ± 0.24, significantly higher than Gemini 2.58 ± 0.28 and DeepSeek (2.18 ± 0.24, *p* < 0.001). Accuracy followed a similar trend (ChatGPT-4o: 3.88 ± 0.16 vs. Gemini: 2.26 ± 0.48 vs. DeepSeek: 2.12 ± 0.54, *p* < 0.001), as did relevance (ChatGPT-4o: 3.99 ± 0.06 vs. Gemini: 2.67 ± 0 vs. DeepSeek: 1.96 ± 0.29, *p* < 0.001). Post-hoc with M-W showed that ChatGPT-4o better than Gemini (*p* < 0.001), DeepSeek (*p* < 0.001), and Gemini is better than DeepSeek (*p* < 0.001, Figure 1).

In Arabic, ChatGPT-4o again demonstrated superior performance with an overall CLEAR score mean of 3.63 ± 0.22), followed by Gemini (2.38 ± 0.14) and DeepSeek (1.84 ± 0.19, *p* < 0.001). Completeness was markedly higher for ChatGPT-4o (3.57 ± 0.32) compared to Gemini (2.38 ± 0.12) and DeepSeek (1.83 ± 0.24, *p* < 0.001). Accuracy scores were 3.46 ± 0.31 for ChatGPT-4o, 2.09 ± 0.38 for Gemini, and 1.68 ± 0.42 for DeepSeek (*p* < 0.001). For relevance, ChatGPT-4o achieved a mean score of 3.88 ± 0.20, whereas Gemini and DeepSeek lagged at 2.67 and 2.00, respectively (*p* < 0.001). Post-hoc with M-W showed that ChatGPT-4o was better than Gemini (*p* < 0.001), DeepSeek (*p* < 0.001), and Gemini is better than DeepSeek (*p* < 0.001, Figure 1).

3.3 Language discrepancy within each GenAI model

Substantial discrepancies emerged when comparing AI-generated responses between English and Arabic within each genAI model. For all three tools, performance in English consistently surpassed that in Arabic across most CLEAR components. ChatGPT-4o demonstrated the most pronounced language divergence. English outputs significantly outperformed Arabic for completeness (U = 226.0, *p* < 0.001), accuracy (U = 122.0, *p* < 0.001), relevance (U = 329.0, *p* = 0.006), and overall CLEAR score (U = 128.5, *p* < 0.001, Figure 2). DeepSeek exhibited consistently lower performance in Arabic, particularly for completeness (U = 156.0, *p* < 0.001), accuracy (U = 236.5, *p* = 0.001), and overall score (U = 164.0, *p* < 0.001, Figure 2). Interestingly, relevance scores were statistically different in the opposite direction (U = 345.0, *p* = 0.030), with Arabic outputs achieving higher ranks than English (mean ranks: 34.0 vs. 27.0, Figure 2). Gemini, although showing overall lower scores than ChatGPT-4o, still displayed statistically significant discrepancies for completeness (U = 260.0, *p* = 0.001) and overall CLEAR score (U = 278.5, *p* = 0.007), with English outperforming Arabic. However, accuracy differences were not statistically significant (U = 367.5, *p* = 0.165), and relevance scores were identical (U = 450.0, *p* = 1.000, Figure 2).

Table 2 Inter-rater reliability and internal consistency for the three CLEAR dimensions.

Dimension	ICC (Average Measures)	95% CI (Lower–Upper)	ICC (Single Measures)	F value	*p* value
Completeness	0.858	0.818–0.891	0.669	7.056	<0.001
Accuracy	0.917	0.894–0.936	0.787	12.109	<0.001
Relevance	0.950	0.936–0.962	0.864	20.110	<0.001

ICC: Intraclass Correlation Coefficient. All values calculated using a two-way mixed-effects model. Each dimension was scored independently by three raters using a 5-point Likert scale.

3.4 Domain-specific insights (asthma, allergy, RTIs)

Performance across clinical domains revealed the following patterns in genAI response quality. Asthma-related queries consistently received the highest ratings across all three models and both languages. ChatGPT-4o demonstrated superior performance in asthma content (mean overall CLEAR: 3.90 in English, 3.68 in Arabic), while even lower-performing models such as DeepSeek and Gemini reached their highest scores in the asthma category. Allergy-related queries were the most challenging domain for genAI models, especially in terms of accuracy with Gemini scoring as low as 2.00 in Arabic accuracy, and DeepSeek also dropped to 2.07 in English. RTIs occupied a middle ground in performance. ChatGPT-4o maintained strong and consistent scores across languages and domains. However, Gemini and DeepSeek demonstrated variability in both accuracy and completeness, particularly in Arabic. Statistical testing using K-W comparisons showed significant differences in CLEAR scores between clinical domains, especially for models with lower baseline performance (e.g., Gemini and DeepSeek) as shown in Table 3.

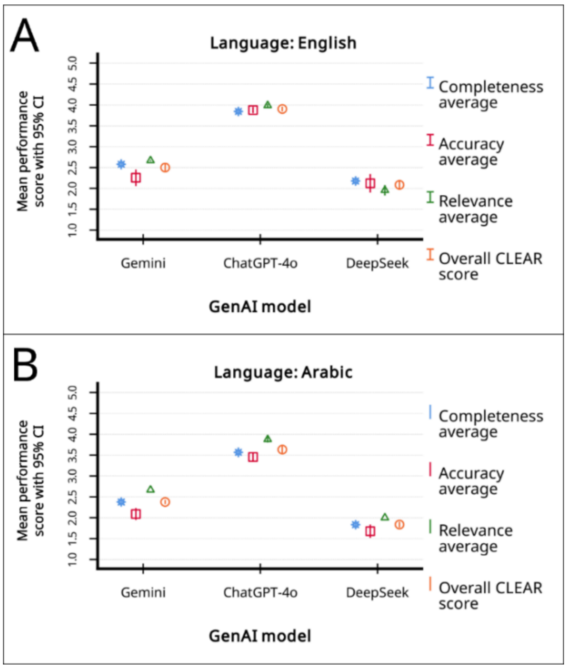

Figure 1 Comparative performance of generative AI (genAI) models stratified by language. (A) and (B) illustrate the mean performance scores with 95% confidence intervals (CIs) for three genAI models—Gemini, ChatGPT-4o, and DeepSeek—across four CLEAR dimensions: Completeness (blue 8-pointed star), Accuracy (red square), Relevance (green triangle), and the Overall CLEAR Score (orange circle), stratified by language. Panel (A) presents results for responses generated in English. Panel (B) presents the same analysis for Arabic responses.

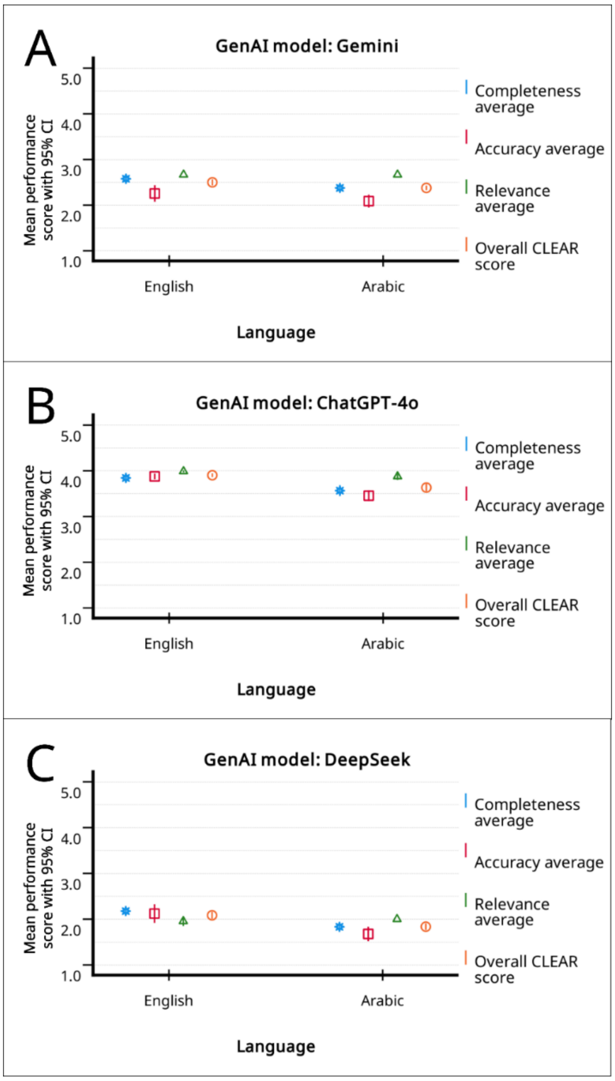

Figure 2 Language-based performance differences across genAI models using the CLEAR evaluation framework. Mean scores with 95% confidence intervals (CIs) are shown for each genAI model: (A) Gemini, (B) ChatGPT-4o, and (C) DeepSeek—stratified by language (English vs. Arabic) and evaluated across the three CLEAR components: Completeness (blue 8-pointed star), Accuracy (red square), Relevance (green triangle), and the Overall CLEAR Score (orange circle).

Table 3 CLEAR component scores by clinical domain, genAI model, and language.

Language	GenAI model	Query category	Completeness average	*p* value	Accuracy average	*p* value	Relevance average	*p* value	Overall CLEAR score	*p* value
			Mean ± SD		Mean ± SD		Mean ± SD		Mean ± SD	
English	*Gemini*	*Asthma*	2.63 ± 0.19	0.421	2.70 ± 0.46	0.002	2.67 ± 0	1.000	2.67 ± 0.17	0.002
		Allergy	2.57 ± 0.35		2.07 ± 0.47		2.67 ± 0		2.43 ± 0.16	
		RTIs	2.53 ± 0.28		2.00 ± 0		2.67 ± 0		2.40 ± 0.09	
	ChatGPT-4o	*Asthma*	3.83 ± 0.18	0.611	3.87 ± 0.17	0.870	4.00 ± 0	0.368	3.90 ± 0.08	0.385
		Allergy	3.87 ± 0.36		3.90 ± 0.16		4.00 ± 0		3.92 ± 0.14	
		RTIs	3.83 ± 0.18		3.87 ± 0.17		3.97 ± 0.11		3.89 ± 0.10	
	DeepSeek	*Asthma*	2.23 ± 0.27	0.646	2.37 ± 0.81	0.281	1.67 ± 0	<0.001	2.09 ± 0.29	0.971
		Allergy	2.13 ± 0.23		2.07 ± 0.38		2.07 ± 0.21		2.09 ± 0.18	
		RTIs	2.17 ± 0.24		1.93 ± 0.21		2.13 ± 0.28		2.08 ± 0.16	
Arabic	*Gemini*	*Asthma*	2.43 ± 0.16	0.142	2.00 ± 0.54	0.195	2.67 ± 0	1.000	2.37 ± 0.20	0.397
		Allergy	2.37 ± 0.11		2.00 ± 0		2.67 ± 0		2.34 ± 0.04	
		RTIs	2.33 ± 0		2.27 ± 0.34		2.67 ± 0		2.42 ± 0.11	
	ChatGPT-4o	*Asthma*	3.60 ± 0.31	0.588	3.53 ± 0.36	0.613	3.90 ± 0.16	0.987	3.68 ± 0.19	0.695
		Allergy	3.63 ± 0.29		3.43 ± 0.32		3.87 ± 0.23		3.64 ± 0.22	
		RTIs	3.47 ± 0.36		3.40 ± 0.26		3.87 ± 0.23		3.58 ± 0.24	
	DeepSeek	*Asthma*	1.97 ± 0.29	0.155	1.73 ± 0.56	0.625	2.00 ± 0	1.000	1.90 ± 0.25	0.608
		Allergy	1.73 ± 0.21		1.60 ± 0.44		2.00 ± 0		1.78 ± 0.17	
		RTIs	1.80 ± 0.17		1.70 ± 0.25		2.00 ± 0		1.83 ± 0.09	

RTI: Respiratory tract infection; SD: standard deviation; *p* values were calculated using the Kruskal-Wallis tests.

4. Discussion

The results of this bilingual evaluation study reflect a paradoxical juncture in the evolution of genAI in medicine—a moment defined as much by its notable advancement as by the persistent fault lines it revealed. On one hand, the trajectory of genAI progress is unambiguous. Recent comparative studies have underlined the dramatic leap in medical proficiency between ChatGPT-3.5 and the latest successors, ChatGPT-4o and ChatGPT-5, with particularly impressive gains observed in disciplines such as microbiology [96], ophthalmology [78], and infectious diseases [22]. These findings echo the broader narrative of genAI models maturing beyond mere linguistic sophistication to encompass increasingly domain-specific competency [97]. Although ChatGPT outperformed its counterparts across all tested domains, previous studies showed that even top AI models fall short of expert-level comprehensiveness in areas like colorectal cancer as recently highlighted by Peng *et al*. [98]. Moreover, AI-generated content often lacks source transparency—raising concerns about misinformation and emphasizing the imperative for clinician-guided refinement of these tools [99, 100].

As with all epistemic revolutions, the disruption brought by genAI lies in what it subtly implies. In this study, Arabic-language outputs often displayed an elegant surface manifested in grammatically correct, syntactically fluent, and stylistically persuasive content. Nevertheless, beneath this fluency in Arabic, our evaluation revealed a quiet deficit. Critical omissions, inaccuracies, and a lack of clinical completeness were recurrent, particularly in allergy responses. This is not a mere matter of flawed translation; it is symptomatic of deeper structural inequities in how LLMs are built and trained as highlighted by Guo *et al*. [101]. Most LLMs rely on tokenization schemes and training corpora developed primarily for English and other high-resource languages [102]. As a result, semantic density, domain-specific medical terminology, and clinical details tend to be more robustly encoded in English than in Arabic which is a morphologically rich language with high contextual dependence and substantial dialectal variation [103-105]. This structural bias places Arabic—and similarly underrepresented languages—at a disadvantage [106]. Reinforcement learning methods used to fine-tune genAI models often prioritize the production of text that is fluent, coherent, and stylistically engaging [107]. In under-represented languages, this optimization may favor surface-level coherence over evidentiary depth, allowing linguistically plausible outputs to obscure subtle gaps in clinical content [108]. However, this emphasis on natural language delivery does not always guarantee factual accuracy—particularly in languages that are underrepresented in training data, such as Arabic. In our study, this manifested as responses that, while linguistically polished, were occasionally incomplete or clinically imprecise. Notably, this decoupling between communicative fluency and clinical reliability was reflected in instances where relevance scores remained relatively high despite reduced accuracy, underlining how persuasive language may mask informational deficiencies. This raises a subtle yet important concern: the ease with which users (patients in this case) may conflate verbal fluency with medical reliability. Such misalignment is not merely theoretical. When AI outputs are used to inform patients, especially in settings where language equity is lacking, these inconsistencies may inadvertently mislead, reduce trust in digital health tools, or reinforce existing disparities [109, 110]. In multilingual healthcare environments, this dynamic risks delivering systematically less reliable guidance to non-English-speaking users, thereby amplifying rather than mitigating inequities in health communication [111, 112]. Rather than a technical flaw alone, these findings serve as a reminder of the ethical responsibility to ensure that genAI tools are culturally and linguistically grounded—not only at the level of translation, but also in terms of domain-specific knowledge representation and clinical completeness—especially when applied to sensitive domains like patient education and clinical communication.

In this study, the CLEAR framework—based on completeness, accuracy, and relevance—served as a rating tool and a structured method to assess the clinical reliability of AI-generated responses. Its utility has been demonstrated across diverse medical domains [22, 60, 77, 78, 92, 113], and our findings further support its adaptability in a bilingual, domain-specific context. Notably, the exceptionally high inter-rater agreement achieved across Arabic and English evaluations signals more than just methodological soundness; it suggests that CLEAR can serve as a durable, language-agnostic standard for appraising medical genAI outputs. Here, the convergence of expert assessments across languages and clinical subdomains reinforces the internal validity of our findings and the external validity of CLEAR instrument itself. Within the comparative landscape of genAI models evaluated in this study, ChatGPT-4o consistently delivered the highest scores for completeness, accuracy, and relevance across both English and Arabic content. This superior performance likely reflects the advantages of its architecture, which integrates reinforcement learning from human feedback, improved multilingual tokenization, and targeted fine-tuning [114, 115]. However, the observed decline in ChatGPT-4o's performance when responding in Arabic—particularly in clinical accuracy and completeness—serves as an important warning. This concern is the unequal development and benefit of genAI across languages. Similar disparities were noted in earlier studies evaluating previous versions of ChatGPT, reinforcing the persistent nature of this challenge [22, 75-77]. For example, Samaan *et al*. evaluated ChatGPT's responses to 91 cirrhosis-related questions and found that while nearly three-quarters of Arabic answers were broadly correct, one-third were less accurate than their English counterparts and more than 13% were completely incorrect [76]. Similarly, in a recent CLEAR-based evaluation of infectious disease queries, English responses outperformed Arabic across most quality dimensions, with disparities observed in completeness, accuracy, and relevance across multiple genAI models [22].

DeepSeek, in contrast, demonstrated the most pronounced performance limitations across all domains evaluated in this study. This was unexpected, given prior benchmarks that emphasized its potential in multilingual applications [26, 78]. Its performance in Arabic was particularly concerning, with accuracy scores rendering many responses indistinguishable from superficially plausible yet misleading content. Gemini positioned itself in the intermediate tier—outperforming DeepSeek but consistently trailing ChatGPT-4o, especially in Arabic. While its responses often aligned topically with the queries, accuracy remained a persistent weakness, suggesting ongoing challenges in maintaining factual consistency across languages. A particularly revealing anomaly emerged in DeepSeek's Arabic relevance scores, which paradoxically surpassed those in English. This counterintuitive finding may suggest that high linguistic fluency or contextual alignment does not necessarily equate to clinical accuracy. It is plausible that the Arabic outputs, although structurally aligned with user queries, masked significant gaps in factual content. Such outputs risk creating an illusion of informativeness—an especially dangerous outcome in healthcare contexts, where confidently stated but inaccurate information may misguide patient care more insidiously than overtly incorrect statements [116]. This emphasizes the critical importance of benchmarking genAI not only for its surface-level coherence but also for its clinical substance [117].

Beyond model-to-model comparisons, our analysis revealed notable domain-specific disparities in genAI performance. These differences are perhaps unsurprising, as the quality of AI output often reflects the structure, consistency, and prevalence of the underlying medical knowledge in its training data. However, these results were mostly non-significant, except for Gemini, while ChatGPT demonstrated a minor non-significant advantage for allergy. These differences likely reflect not only variability in model capabilities but also inherent differences in conceptual structure and clinical complexity across clinical domains. Asthma emerged as the most robust domain across all models, likely due to the widespread availability of guideline-driven content such as that from the GINA and the AAAAI [82, 83]. The clear, protocolized nature of asthma management appears well-suited to the structured learning paradigms of LLMs. In contrast, allergy content consistently scored lowest—particularly in accuracy. This may reflect the inherently heterogeneous and subjective nature of allergic diseases, which span a wide spectrum of symptoms, etiologies, and patient experiences [118, 119]. Moreover, allergy information is frequently diluted by misinformation online, including unproven diagnostic tools and questionable dietary or environmental interventions, which may contaminate public-facing datasets used for model training [120-122]. On the other hand, RTIs fell between these two extremes, demonstrating moderate performance. The variability in RTI content—driven by evolving pathogens, seasonal patterns, and rapidly changing public health recommendations—may challenge consistency in AI-generated responses. Collectively, these observations suggest that genAI performance varies not only by language and model architecture but also by clinical domain-level conceptual complexity and decision criticality. Future studies could strengthen domain-specific analyses by independently rating query complexity or clinical criticality, enabling stratified evaluation within domains. Such approaches would help disentangle model limitations from inherent differences in medical knowledge structures, while preserving the patient-facing focus of the current dataset.

Among the most notable findings in our study is the persistent performance gap between English and Arabic—a disparity that casts serious doubt on the global readiness of genAI for equitable use in healthcare. This linguistic divide underlines a deeper structural vulnerability in current AI development pipelines, which continue to privilege high-resource languages in both model training and evaluation. The consequences are far-reaching. In a world increasingly reliant on digital health tools and global health communication, multilingual fidelity is essential [123, 124]. Without it, we risk amplifying existing disparities and inadvertently introducing new ones [125]. AI models that deliver fluent but inaccurate content in underrepresented languages may mislead users who lack access to alternative sources or the health literacy to cross-check information. The illusion of reliability can be more dangerous than obvious error. In such contexts, the failure to support linguistically diverse populations becomes more than an oversight—it becomes an ethical lapse [126]. As genAI moves closer to integration into public health education, clinical decision support, and patient engagement, its ability to operate safely and reliably across languages must be held to the same standards as its performance in English [57, 127].

It is important to distinguish between knowledge-oriented and reasoning-oriented clinical tasks when interpreting the present findings. The FAQ set evaluated in this study primarily probes patient-facing, knowledge-oriented functions of genAI, including explanation of disease concepts, symptoms, triggers, prevention strategies, and general management principles. These tasks emphasize accurate synthesis, clarity, and faithful communication of established medical knowledge rather than multistep clinical reasoning. In contrast, high-risk clinical scenarios often require complex reasoning processes—such as differential diagnosis, risk stratification, individualized treatment selection, and contextual decision-making—which were deliberately excluded from the current evaluation. This boundary was intentional, reflecting the study's focus on health education and self-management support rather than clinical decision support. Viewed in this light, the present work provides a rigorous baseline assessment of bilingual genAI performance in patient-facing content, while highlighting the need for complementary research that directly evaluates reasoning-intensive and safety-critical clinical tasks.

Several areas for future research emerge from this study findings, particularly around enhancing the reliability and equity of genAI in clinical practice. One immediate priority is a systematic audit of the training data that underpins these AI models. Our findings suggest that the underperformance of Arabic content—especially in allergy-related queries—may stem from insufficient representation of Arabic-language clinical corpora during training. Without addressing these imbalances, genAI models will continue to reinforce existing disparities in patient education and clinical support. Another concern lies in the temporal stability of AI-generated content. Medicine evolves rapidly, particularly during periods of infectious disease outbreaks, as seen with RSV and COVID-19. It is essential to examine whether genAI responses remain consistent, timely, and evidence-based as new guidelines emerge. The current study was conducted at a single time point; however, future work should assess how reliably these models keep pace with the evolving medical literature and public health priorities [128]. Language structure also appears to influence output quality. In morphologically rich languages like Arabic, subtle changes in phrasing may alter the model's interpretation and response accuracy. Understanding how prompt structure interacts with model output—commonly referred to as prompt engineering—could improve response quality in these languages [129]. Such optimization would allow clinicians and patients to interact more effectively with AI tools in their native language. Finally, future evaluations must involve real-world users—patients with varying levels of health literacy—to assess whether high technical scores translate into real understanding and safe decision-making. As these tools increasingly enter patient-facing environments, ensuring that outputs are not only accurate but also comprehensible and actionable becomes critical. In this context, clinical validation must be more than a checklist—it must be a commitment to responsible innovation. Finally, the present evaluation represented a static, single-turn assessment and did not examine genAI models' capacity for adaptive refinement following expert feedback. In clinical practice, emerging "copilot" paradigms increasingly rely on iterative, multi-turn interactions in which clinicians identify omissions, challenge inaccuracies, or request clarification. Future research could extend the current framework by implementing structured, multi-turn dialogues in which expert feedback is provided and models are prompted to revise their responses accordingly. Quantifying changes in performance such as pre- and post-feedback differences in CLEAR scores would enable evaluation of genAI responsiveness, learning behavior, and error correction capacity, thereby complementing static benchmarking with interaction-aware assessment.

This study has several limitations that warrant consideration upon interpreting the results as follows. First, it focused exclusively on asthma, allergy, and RTIs—common conditions, but not representative of the full spectrum of medical practice. GenAI performance may differ in specialties that involve more complex decision-making or less standardized education. Additionally, although the query set was purposefully designed to achieve broad conceptual and functional coverage within each domain, the number of queries (30 in total) was not intended to support exhaustive exploration of linguistic variation or fine-grained subgroup analyses. While this approach aligns with real-world patient-facing use cases and was sufficient for the study's primary comparative objectives, future research should examine larger and more heterogeneous query sets to capture additional dimensions of linguistic richness, topic complexity, and semantic variability. Second, while expert agreement was strong, the assessment relied solely on clinician judgment. Without input from patients or caregivers, particularly regarding relevance, the real-world applicability of the findings may be limited. Third, all prompts were delivered in Modern Standard Arabic, which, while widely taught, does not reflect the regional dialects commonly used in patient interactions [75]. This may limit generalizability within Arabic-speaking populations since performance may differ when genAI models are queried using dialectal Arabic or colloquial phrasing, representing an important area for future investigation. Fourth, the evaluation reflects a single time point in a rapidly evolving genAI landscape. Updates to models like ChatGPT-4o and Gemini could significantly alter their outputs over time. In addition, the single-turn, zero-context interaction design does not capture the dynamics of multi-turn dialogue, such as clarification requests, follow-up questions, or adaptive personalization that may occur during extended user–AI interactions. While this approach was intentionally chosen to enhance standardization and cross-model comparability, it may underestimate or overestimate genAI performance in real-world conversational settings where iterative interaction can refine or correct responses. Fifth, because these models are proprietary, we had no access to their training data or fine-tuning processes, which restricts our ability to understand or explain observed differences, especially across languages. Sixth, although our sample of 30 questions was sufficient for primary comparisons, it was not powered to explore more nuanced patterns, such as performance variation by topic complexity. Seventh, we focused on completeness, accuracy, and relevance but did not formally assess hallucination rates or user trust—factors that are crucial in clinical settings. Lastly, we evaluated only English and Arabic. Other low-resource or structurally distinct languages may present even greater challenges that were not captured here. These limitations highlight the need for broader, more inclusive, and multilingual evaluations, ideally involving both clinicians and patients, to support safe and equitable integration of genAI into healthcare.

5. Conclusions

This bilingual evaluation demonstrated that ChatGPT-4o currently offers the highest-quality patient-facing content in asthma, allergy, and RTIs, particularly in English. The CLEAR framework showed excellent inter-rater reliability, supporting its utility in multilingual model evaluation. However, significant performance gaps—most notably in Arabic outputs and in clinical domains such as allergy—highlight ongoing limitations in generalizability and reliability across languages and content areas. These findings highlight the need for domain-specific and language-sensitive validation of genAI tools before clinical integration. Without equitable performance across linguistic contexts, such tools risk perpetuating disparities in health communication. Future efforts should prioritize transparency, multilingual optimization, and clinical oversight to ensure safe and inclusive deployment of genAI in healthcare.

Abbreviations

AAAAI	The American Academy of Allergy, Asthma & Immunology
AI	Artificial intelligence
API	Application programming interface
CI	Confidence interval
CLEAR	Completeness, Lack of false information, Evidence, Appropriateness, and Relevance
COVID-19	Coronavirus disease 2019
FAQs	Frequently asked questions
genAI	Generative artificial intelligence
GINA	Global Initiative for Asthma
ICC	Intraclass correlation coefficient
K-W	Kruskal-Wallis H test
LLM	Large language model
METRICS	Model, Evaluation, Timing, Range/Randomization, Individual factors, Count, and Specificity of prompts and language
M-W	Mann-Whitney *U* test
NHS	The National Health Service
RSV	Respiratory syncytial virus
RTI	Respiratory tract infection
SD	Standard deviations
UAE	United Arab Emirates
WHO	World Health Organization

Acknowledgments

The authors declare that Generative AI was used in the creation of this manuscript. ChatGPT-5 was employed for language refinement (improving grammar, sentence structure, and readability of the manuscript) and technical writing assistance (providing suggestions for structuring complex technical descriptions more effectively). We confirm that all AI-assisted processes were critically reviewed by the authors to ensure the

integrity and reliability of the results. The final decisions and interpretations presented in this article were solely made by the authors.

Author Contributions

MoS: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing. AS: Data curation; Investigation; Methodology; Validation; Writing – review & editing. JS: Data curation; Investigation; Methodology; Validation; Writing – review & editing. HA: Data curation; Investigation; Methodology; Validation; Writing – review & editing. OAA: Data curation; Investigation; Methodology; Validation; Writing – review & editing. RA: Data curation; Investigation; Methodology; Validation; Writing – review & editing. MaS: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Competing Interests

The authors declare that they have no conflicts of interest.

Availability of Data and Materials

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

Additional Materials

Additional materials have been added to this paper's page, including:

1. Appendix.

References

1. Cline RJ, Haynes KM. Consumer health information seeking on the Internet: The state of the art. Health education research. 2001;16(6):671-692.
2. Lemire M, Paré G, Sicotte C, Harvey C. Determinants of Internet use as a preferred source of information on personal health. International journal of medical informatics. 2008;77(11):723-734.
3. Tonsaker T, Bartlett G, Trpkov C. Health information on the Internet: gold mine or minefield? Canadian Family Physician. 2014;60(5):407-408.
4. Andreassen HK, Bujnowska-Fedak MM, Chronaki CE, Dumitru RC, Pudule I, Santana S, et al. European citizens' use of E-health services: a study of seven countries. BMC public health. 2007;7(1):53.
5. Kummervold P, Chronaki C, Lausen B, Prokosch H-U, Rasmussen J, Santana S, et al. eHealth trends in Europe 2005-2007: a population-based survey. Journal of Medical Internet Research. 2008;10(4):e42.
6. Frishauf P. Medscape–The first 5 years. Medscape General Medicine. 2005;7(2):5.
7. National Health Service (NHS). NHS website for England: Find information and services to help you manage your health [Internet]. Leeds, England: National Health Service; 2025. Available from: https://www.nhs.uk/.
8. Smith CA, Wicks PJ. PatientsLikeMe: Consumer health vocabulary as a folksonomy. AMIA annual symposium proceedings. 2008;2008:682.
9. Arruda RMd, Ayoub IA, Nunes R, Azevedo Neto RSd, Nunes MdPT. Consulting “Dr. Google”: how the digital search for internet health information influences doctor-patient relationship. Cadernos de Saúde Pública. 2025;41:e00153623.
10. Cacciamani GE, Dell’Oglio P, Cocci A, Russo GI, Abreu ADC, Gill IS, et al. Asking “Dr. Google” for a second opinion: the devil is in the details. European urology focus. 2021;7(2):479-481.
11. Fox S. Pew Research Center. Online Health Search 2006: Part 1. 113 Million Internet Users Seek Health Information Online [Internet]. Washington, DC: Pew Research Center; 2006. Available from: https://www.pewresearch.org/internet/2006/10/29/part-1-113-million-internet-users-seek-health-information-online/.
12. Wac K. Smartphone as a personal, pervasive health informatics services platform: literature review. Yearbook of medical informatics. 2012;21(01):83-93.
13. Zawati MnH, Lang M. Does an app a day keep the doctor away? AI symptom checker applications, entrenched bias, and professional responsibility. Journal of Medical Internet Research. 2024;26:e50344.
14. Ozdalga E, Ozdalga A, Ahuja N. The smartphone in medicine: a review of current and potential use among physicians and students. Journal of Medical Internet Research. 2012;14(5):e128.
15. Shen Y-T, Chen L, Yue W-W, Xu H-X. Digital technology-based telemedicine for the COVID-19 pandemic. Frontiers in medicine. 2021;8:646506.
16. Rouvinen H, Turunen H, Lindfors P, Kinnunen JM, Rimpelä A, Koivusilta L, et al. Online health information-seeking behaviour and mental well-being among Finnish higher education students during COVID-19. Health promotion international. 2023;38(6):daad143.
17. Almalki M, Azeez F. Health chatbots for fighting COVID-19: a scoping review. Acta Informatica Medica. 2020;28(4):241-247.
18. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare. 2023;11(6):887.
19. Alanzi TM. Impact of ChatGPT on teleconsultants in healthcare: perceptions of healthcare experts in Saudi Arabia. Journal of multidisciplinary healthcare. 2023:2309-2321.
20. Alanezi F. Factors influencing patients’ engagement with ChatGPT for accessing health-related information. Critical Public Health. 2024;34(1):1-20.
21. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of medical information provided by ChatGPT: assessment against clinical

guidelines and patient information quality instrument. Journal of Medical Internet Research. 2023;25:e47479.

22. Sallam M, Al-Mahzoum K, Alshuaib O, Alhajri H, Alotaibi F, Alkhurainej D, et al. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. BMC Infectious Diseases. 2024;24(1):799.
23. Rebitschek FG, Carella A, Kohlrausch-Pazin S, Zitzmann M, Steckelberg A, Wilhelm C. Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information. npj Digital Medicine. 2025;8(1):343.
24. Madanian S, Nakarada-Kordic I, Reay S, Chetty Th. Patients' perspectives on digital health tools. PEC innovation. 2023;2:100171.
25. Esmaeilzadeh P, Maddah M, Mirzaei T. Using AI chatbots (eg, CHATGPT) in seeking health-related information online: The case of a common ailment. Computers in Human Behavior: Artificial Humans. 2025;3:100127.
26. Sallam M, Al-Mahzoum K, Sallam M, Mijwil MM. DeepSeek: Is it the end of generative AI monopoly or the mark of the impending doomsday? Mesopotamian Journal of Big Data. 2025;2025:26-34.
27. Chow JC, Li K. Large language models in medical chatbots: opportunities, challenges, and the need to address AI risks. Information. 2025;16(7):549.
28. Xu R, Wang Z. Generative artificial intelligence in healthcare from the perspective of digital media: Applications, opportunities and challenges. Heliyon. 2024;10(12):e32364.
29. Goktas P, Damadoglu E. Future of allergy and immunology: is artificial intelligence the key in the digital era? Annals of Allergy, Asthma & Immunology. 2025;134(4):396-407.e392.
30. González-Díaz SN, Morais-Almeida M, Ansotegui IJ, Macouzet-Sánchez C, Ordóñez-Azuara YG, Camarena-Galván J, et al. Artificial intelligence in allergy practice: Digital transformation and the future of clinical care. World Allergy Organization Journal. 2025;18(8):101078.
31. Tan LD, Nguyen N, Lopez E, Peverini D, Shedd M, Alismail A, et al. Artificial intelligence in the management of asthma: a review of a new frontier in patient care. Journal of Asthma and Allergy. 2025:1179-1191.
32. Pawankar R. Allergic diseases and asthma: a global public health concern and a call to action. World Allergy Organization Journal. 2014;7(1):12.
33. Oh J, Kim S, Kim MS, Abate YH, Abd ElHafeez S, Abdelkader A, et al. Global, regional, and national burden of asthma and atopic dermatitis, 1990–2021, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. The lancet respiratory medicine. 2025;13(5):425-446.
34. Sirota SB, Doxey MC, Dominguez R-MV, Bender RG, Vongpradith A, Albertson SB, et al. Global, regional, and national burden of upper respiratory infections and otitis media, 1990–2021: a systematic analysis from the Global Burden of Disease Study 2021. The Lancet Infectious Diseases. 2025;25(1):36-51.
35. Bender RG, Sirota SB, Swetschinski LR, Dominguez R-MV, Novotney A, Wool EE, et al. Global, regional, and national incidence and mortality burden of non-COVID-19 lower respiratory infections and aetiologies, 1990–2021: a systematic analysis from the Global Burden of Disease Study 2021. The Lancet Infectious Diseases. 2024;24(9):974-1002.
36. Gohal G, Moni SS, Bakkari MA, Elmobark ME. A review on asthma and allergy: current understanding on molecular perspectives. Journal of Clinical Medicine. 2024;13(19):5775.
37. Monk AS, Worden CP, Benaim EH, Klatt-Cromwell C, Thorp BD, Ebert Jr CS, et al. The impact of occupational exposures on chronic rhinosinusitis: a scoping review. Exploration of Asthma & Allergy. 2024;2(4):301-308.
38. Ibekwe PU, Ekop E, Otu T, Bassi P, Ukonu BA. Atopic dermatitis in adults: prevalence, clinical pattern, and contact sensitization. Exploration of Asthma & Allergy. 2024;2(5):450-460.
39. Skolnik N, Yawn BP, Correia de Sousa J, Vázquez MMM, Barnard A, Wright WL, et al. Best practice advice for asthma exacerbation prevention and management in primary care: an international expert consensus. NPJ Primary Care Respiratory Medicine. 2024;34(1):39.
40. Chamola V. Generative AI for Transformative Healthcare: A Comprehensive Study of Emerging Models, Applications, Case Studies, and Limitations. Ieee Access. 2024;12:31078-31106.
41. Maleki Varnosfaderani S, Forouzanfar M. The role of AI in hospitals and clinics: transforming healthcare in the 21st century. Bioengineering. 2024;11(4):337.
42. Babel A, Taneja R, Mondello Malvestiti F, Monaco A, Donde S. Artificial intelligence solutions to increase medication adherence in patients with non-communicable diseases. Frontiers in Digital Health. 2021;3:669869.
43. Drummond D, Adejumo I, Hansen K, Poberezhets V, Slabaugh G, Hui CY. Artificial intelligence in respiratory care: perspectives on critical opportunities and challenges. Breathe. 2024;20(3):230189.
44. Gori A, Zicari A, Barreto M, Della Giustina A, Sfika I, Pattini S, et al. Artificial Intelligence-Driven Innovations in Allergy. Italian Journal of Pediatric Allergy and Immunology. 2025;39(1):22-25.
45. van Breugel M, Fehrmann RS, Bügel M, Rezwan FI, Holloway JW, Nawijn MC, et al. Current state and prospects of artificial intelligence in allergy. Allergy. 2023;78(10):2623-2643.
46. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025:442-450.
47. Fu B, Hadid A, Damer N. Generative AI in the context of assistive technologies: Trends, limitations and future directions. Image and Vision Computing. 2025;154:105347.
48. Deniz-Garcia A, Fabelo H, Rodriguez-Almeida AJ, Zamora-Zamorano G, Castro-Fernandez M, Alberiche Ruano MdP, et al. Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: current scenario and challenges. Journal of Medical Internet Research. 2023;25:e44030.

49. Mohamed YA, Khanan A, Bashir M, Mohamed AHH, Adiel MA, Elsadig MA. The impact of artificial intelligence on language translation: a review. Ieee Access. 2024;12:25553-25579.

50. Witkowski K, Dougherty RB, Neely SR. Public perceptions of artificial intelligence in healthcare: ethical concerns and opportunities for patient-centered care. BMC Medical Ethics. 2024;25(1):74.

51. Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications. 2024;11(1):1278.

52. Braido F. Failure in asthma control: reasons and consequences. Scientifica. 2013;2013(1):549252.

53. Pouessel G, Morisset M, Schoder G, Santos C, Villard-Truc F, Just J, et al. Update on the emergency action plan for allergic reactions in children and adolescents. Position of the "Allergy at school" and "Food allergy" working groups of the French Allergology Society. Revue Française d'Allergologie. 2020;60(2):83-89.

54. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. PLOS Digital Health. 2024;3(11):e0000651.

55. Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. arXiv. 2024. DOI: 10.48550/arXiv.2408.15266.

56. Bélisle-Pipon J-C. Why we need to be careful with LLMs in medicine. Frontiers in medicine. 2024;11:1495582.

57. Yim D, Khuntia J, Parameswaran V, Meyers A. Preliminary evidence of the use of generative AI in health care clinical services: systematic narrative review. JMIR Medical Informatics. 2024;12(1):e52073.

58. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nature Medicine. 2025;31(3):943-950.

59. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023;2(2):e0000198.

60. Sallam M, Al-Salahat K, Eid H, Egger J, Puladi B. Human versus artificial intelligence: ChatGPT-4 outperforming Bing, bard, ChatGPT-3.5 and humans in clinical chemistry Multiple-Choice questions. Advances in Medical Education and Practice. 2024:857-871.

61. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. Heliyon. 2024;10(4):e26297.

62. Li J. Security implications of AI chatbots in health care. Journal of Medical Internet Research. 2023;25:e47551.

63. Sallam M, Al-Mahzoum K, Sallam M. Generative Artificial Intelligence and Cybersecurity Risks: Implications for Healthcare Security Based on Real-life Incidents. Mesopotamian Journal of Artificial Intelligence in Healthcare. 2024;2024:184-203.

64. Habli I, Lawton T, Porter Z. Artificial intelligence in health care: accountability and safety. Bulletin of the World Health Organization. 2020;98(4):251-256.

65. Wei X, Kumar N, Zhang H. Addressing Bias in Generative Ai: Challenges and Research Opportunities in Information Management. Information & Management. 2025;62(2):104103.

66. Schut L, Gal Y, Farquhar S. Do Multilingual LLMs Think In English? arXiv. 2025. DOI: 10.48550/arXiv.2502.15603.

67. Myung J, Lee N, Zhou Y, Jin J, Putri R, Antypas D, et al. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. arXiv. 2024. DOI: 10.48550/arXiv.2406.09948.

68. Nacar O, Sibaee ST, Ahmed S, Atitallah SB, Ammar A, Alhabashi Y, et al., editors. Towards inclusive Arabic LLMs: A culturally aligned benchmark in Arabic large language model evaluation. Proceedings of the First Workshop on Language Models for Low-Resource Languages; 2025 January 20; Abu Dhabi, United Arab Emirates. Stroudsburg, PA: Association for Computational Linguistics.

69. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: evaluation study. JMIR nursing. 2023;6:e47305.

70. Guigue PA, Meyer R, Thivolle-Lioux G, Brezinov Y, Levin G. Performance of ChatGPT in French language Parcours d'Accès Spécifique Santé test and in OBGYN. International Journal of Gynecology & Obstetrics. 2024;164(3):959-963.

71. Morreel S, Mathysen D, Verhoeven V. Aye, AI! ChatGPT passes multiple-choice family medicine exam. Medical Teacher. 2023;45(6):665-666.

72. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. Journal of Medical Internet Research. 2024;26:e51926.

73. Tarraf H, Aydin O, Mungan D, Albader M, Mahboub B, Doble A, et al. Prevalence of asthma among the adult general population of five Middle Eastern countries: results of the SNAPSHOT program. BMC pulmonary medicine. 2018;18(1):68.

74. Al-Digheari A, Mahboub B, Tarraf H, Yucel T, Annesi-Maesano I, Doble A, et al. The clinical burden of allergic rhinitis in five Middle Eastern countries: results of the SNAPSHOT program. Allergy, Asthma & Clinical Immunology. 2018;14(1):63.

75. Sallam M, Mousa D. Evaluating ChatGPT performance in Arabic dialects: A comparative study showing defects in responding to Jordanian and Tunisian general health prompts. Mesopotamian Journal of Artificial Intelligence in Healthcare. 2024;2024:1-7.

76. Samaan JS, Yeo YH, Ng WH, Ting P-S, Trivedi H, Vipani A, et al. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. Arab Journal of Gastroenterology. 2023;24(3):145-148.

77. Sallam M, Al-Mahzoum K, Almutawaa RA, Alhashash JA, Dashti RA, AlSafy DR, et al. The performance of OpenAI ChatGPT-4 and Google Gemini in virology multiple-

choice questions: a comparative analysis of English and Arabic responses. BMC Research Notes. 2024;17(1):247.

78. Sallam M, Alasfoor IM, Khalid SW, Al-Mulla RI, Al-Farajat A, Mijwil MM, et al. Chinese generative AI models (DeepSeek and Qwen) rival ChatGPT-4 in ophthalmology queries with excellent performance in Arabic and English. Narra J. 2025;5(1):e2371.
79. Sallam M, Barakat M, Sallam M. A Preliminary Checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence–based models in health care education and practice: development study involving a literature review. Interactive journal of medical research. 2024;13(1):e54704.
80. Sallam M, Barakat M, Sallam M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by artificial intelligence-based models. Cureus. 2023;15(11):e49373.
81. Dhand NK, MS K. Statulator: An online statistical calculator. Sample Size Calculator for Comparing Two Paired Means [Internet]. Statulator; 2025. Available from: http://statulator.com/SampleSize/ss2PM.html.
82. Global Initiative for Asthma – GINA. FAQs: Answers to frequently asked questions about asthma [Internet]. Fontana, WI: Global Initiative for Asthma – GINA; 2025. Available from: https://ginasthma.org/about-us/faqs/.
83. American Academy of Allergy AI. Overview of asthma symptoms, asthma diagnosis, asthma treatment and asthma [Internet]. Milwaukee, WI: American Academy of Allergy AI; 2025. Available from: https://www.aaaai.org/conditions-treatments/asthma.
84. WebMD Editorial Contributors. Allergies: Your Top Questions Answered. WebMD [Internet]. WebMD Editorial Contributors; 2025. Available from: https://www.webmd.com/allergies/allergies-faq.
85. American Academy of Allergy AI. Learn about symptoms, diagnosis, treatment and management for these allergies [Internet]. Milwaukee, WI: American Academy of Allergy AI; 2025. Available from: https://www.aaaai.org/conditions-treatments/allergies.
86. WHO. Q&As on COVID-19 and related health topics [Internet]. Geneva: WHO; 2025. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub.
87. NHS Borders. RSV - Frequently Asked Questions [Internet]. Scottish: NHS Borders; 2025. Available from: https://www.nhsborders.scot.nhs.uk/patients-and-visitors/respiratory-syncytial-virus-(rsv)/frequently-asked-questions/.
88. WHO. Regional Office for the Eastern Mediterranean. Influenza (seasonal): Influenza Q&As [Internet]. WHO2025. Available from: https://www.emro.who.int/health-topics/influenza/questions-and-answers.html#influenza.
89. OpenAI. ChatGPT-4o [Internets]. OpenAI; 2025. Available from: https://chatgpt.com/?model=gpt-4o.
90. DeepSeek. DeepSeek-V3 [Internet]. DeepSeek; 2025. Available from: https://chat.deepseek.com/.
91. Google. Gemini Flash 2.5 (version 2.5) [Internet]. Google; 2025. Available from: https://gemini.google.com/app.
92. Muluk SY, Olcucu N. The role of artificial intelligence in the primary prevention of common musculoskeletal diseases. Cureus. 2024;16(7):e65372.
93. Alnsour MM, Alenezi R, Barakat M, Al-Omiri MK. Assessing ChatGPT's suitability in responding to the public's inquires on the effects of smoking on oral health. BMC Oral Health. 2025;25(1):1207.
94. Incerti Parenti S, Bartolucci ML, Biondi E, Maglioni A, Corazza G, Gracco A, et al. Online patient education in obstructive sleep apnea: ChatGPT versus Google Search. Healthcare. 2024;12(17):1781.
95. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine. 2016;15(2):155-163.
96. Sallam M, Irshaid A, Snygg J, Albadri R, Sallam M. Rapid Evolution of Large Language Models in Medical Education: Comparative Performance of ChatGPT-3.5, ChatGPT-5, and DeepSeek on Medical Microbiology MCQs. Contemporary Education and Teaching Research. 2025;6:295-309.
97. Yuan M, Bao P, Yuan J, Shen Y, Chen Z, Xie Y, et al. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. Medicine Plus. 2024;1(2):100030.
98. Peng W, Feng Y, Yao C, Zhang S, Zhuo H, Qiu T, et al. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions. Scientific reports. 2024;14(1):2840.
99. Checcucci E, Rodler S, Piazza P, Porpiglia F, Cacciamani GE. Transitioning from "Dr. Google" to "Dr. ChatGPT": the advent of artificial intelligence chatbots. Translational Andrology and Urology. 2024;13(6):1067-1070.
100. Cazzamatta R, Sarısakaloğlu A. AI-generated misinformation: A case study on emerging trends in fact-checking practices across Brazil, Germany, and the United Kingdom. Emerging Media. 2025;3:214-251.
101. Guo Y, Guo M, Su J, Yang Z, Zhu M, Li H, et al. Bias in large language models: Origin, evaluation, and mitigation. arXiv. 2024. DOI: 10.48550/arXiv.2411.10915.
102. Seo J, Kim J, Byun S, Shin H. How does a Language-Specific Tokenizer affect LLMs? arXiv. 2025. DOI: 10.48550/arXiv.2502.12560.
103. Saadi N, Raha T, Christophe C, Pimentel MA, Rajan R, Kanithi PK. Bridging Language Barriers in Healthcare: A Study on Arabic LLMs. arXiv. 2025. DOI: 10.48550/arXiv.2501.09825.
104. Mohammad R, Alkhnbashi OS, Hammoudeh M. Optimizing Large Language Models for Arabic Healthcare Communication: A Focus on Patient-Centered NLP Applications. Big Data and Cognitive Computing. 2024;8(11):157.
105. Ibrahim A, Hosseini A, Helmy H, Lakhdhar W, Serag A, editors. Bridging Dialectal Gaps in Arabic Medical LLMs through Model Merging. Proceedings of The Third Arabic Natural Language Processing Conference; 2025 November 8-9; Suzhou, China. Stroudsburg, PA: Association for Computational Linguistics.

106. Qin L, Chen Q, Zhou Y, Chen Z, Li Y, Liao L, et al. A survey of multilingual large language models. Patterns. 2025;6(1):101118.
107. Wang S, Zhang S, Zhang J, Hu R, Li X, Zhang T, et al. Reinforcement learning enhanced llms: A survey. arXiv. 2024. DOI: 10.48550/arXiv.2412.10400.
108. Alshehhi M, Sharshar A, Guizani M. Towards Inclusive NLP: Assessing Compressed Multilingual Transformers Across Diverse Language Benchmarks. arXiv. 2025. DOI: 10.48550/arXiv.2507.19699.
109. Dankwa-Mullan I. Health equity and ethical considerations in using artificial intelligence in public health and medicine. Preventing chronic disease. 2024;21:E64.
110. Weiner EB, Dankwa-Mullan I, Nelson WA, Hassanpour S. Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. PLOS Digital Health. 2025;4(4):e0000810.
111. Schlicht IB, Sayin B, Zhao Z, Labonté FM, Barbera C, Viviani M, et al. Disparities in Multilingual LLM-Based Healthcare Q&A. arXiv. 2025. DOI: 10.48550/arXiv.2510.17476.
112. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of language barriers for healthcare: a systematic review. Oman medical journal. 2020;35(2):e122.
113. Aljamani S, Hassona Y, Fansa HA, Saadeh HM, Jamani KD. Evaluating Large Language Models in Addressing Patient Questions on Endodontic Pain: A Comparative Analysis of Accessible Chatbots. Journal of Endodontics. 2025;51:1617-1624.
114. Islam R, Moushi OM. Gpt-4o: The cutting-edge advancement in multimodal llm. TechRxiv. 2025. DOI: 10.36227/techrxiv.171986596.65533294/v1.
115. ŞAHİN EG. Comparative performance of ChatGPT, Gemini, and DeepSeek on endodontic exam questions in Turkish and English. 2025. DOI: 10.21203/rs.3.rs-6738945/v1.
116. Palmer A, Gorman S. Misinformation, Trust, and Health: The Case for Information Environment as a Major Independent Social Determinant of Health. Social Science & Medicine. 2025:118272.
117. Sallam M, Khalil R, Sallam M. Benchmarking generative AI: A call for establishing a comprehensive framework and a generative AIQ test. Mesopotamian Journal of Artificial Intelligence in Healthcare. 2024;2024:69-75.
118. Bonini S, Rasi G, Torre A, D'Amato M, Matricardi PM. The heterogeneity of allergic phenotypes: genetic. Ann Allergy Asthma Immunol. 2001;87:48-51.
119. Wang J, Zhou Y, Zhang H, Hu L, Liu J, Wang L, et al. Pathogenesis of allergic diseases and implications for therapeutic interventions. Signal transduction and targeted therapy. 2023;8(1):138.
120. Stukus DR. Tackling medical misinformation in allergy and immunology practice. Expert Review of Clinical Immunology. 2022;18(10):995-996.
121. Verdi M, Candido D, Madan J, Bernstein JA, Bukstein D, Anagnostou A, et al. Addressing Anxiety and Depression in the Allergy Clinic Through Motivational Interviewing, Brief Cognitive Behavioral Therapy, and Curious Questions. The Journal of Allergy and Clinical Immunology: In Practice. 2025;13:1960-1969.e1962.
122. Anagnostou A. Addressing common misconceptions in food allergy: a review. Children. 2021;8(6):497.
123. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. Frontiers in Digital Health. 2023;5:1264780.
124. Khan R, Khan S, Almohaimeed HM, Almars AI, Pari B. Utilization, challenges, and training needs of digital health technologies: perspectives from healthcare professionals. International journal of medical informatics. 2025;197:105833.
125. Badr J, Motulsky A, Denis J-L. Digital health technologies and inequalities: a scoping review of potential impacts and policy recommendations. Health Policy. 2024;146:105122.
126. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. The Lancet Digital Health. 2024;6(11):e848-e856.
127. Reddy S. Generative AI in healthcare: an implementation science informed translational path on application, integration and governance. Implementation Science. 2024;19(1):27.
128. Sallam M. Bibliometric top ten healthcare-related ChatGPT publications in the first ChatGPT anniversary. Narra J. 2024;4(2):e917.
129. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns. 2025;6:101260.