applied sciences

MDPI

Systematic Review

Quantifying Readability in Chatbot-Generated Medical Texts Using Classical Linguistic Indices: A Review

Robert Olszewski [1,2], Jakub Brzeziński [1,*], Klaudia Watros [1] and Jacek Rysz [3]

[1] Department of Gerontology and Public Health, National Institute of Geriatrics, Rheumatology and Rehabilitation, Spartańska 1 Street, 02-637 Warsaw, Poland; robert.olszewski@spartanska.pl (R.O.); klaudia.watros@spartanska.pl (K.W.)
[2] Department of Ultrasound, Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5B Street, 02-106 Warsaw, Poland
[3] Department of Nephrology, Hypertension and Family Medicine, Medical University of Lodz, Ul. Zeromskiego 113, 90-549 Lodz, Poland; jacek.rysz@umed.lodz.pl
* Correspondence: jakub.brzezinski@spartanska.pl; Tel: +48-22-670-9262

Abstract

The rapid development of large language models (LLMs), including ChatGPT, Gemini, and Copilot, has led to their increasing use in health communication and patient education. However, their growing popularity raises important concerns about whether the language they generate aligns with recommended readability standards and patient health literacy levels. This review synthesizes evidence on the readability of medical information generated by chatbots using established linguistic readability indices. A comprehensive search of PubMed, Scopus, Web of Science, and Cochrane Library identified 4209 records, from which 140 studies met the eligibility criteria. Across the included publications, 21 chatbots and 14 readability scales were examined, with the Flesch–Kincaid Grade Level and Flesch Reading Ease being the most frequently applied metrics. The results demonstrated substantial variability in readability across chatbot models; however, most texts corresponded to a secondary or early tertiary reading level, exceeding the commonly recommended 8th-grade level for patient-facing materials. ChatGPT-4, Gemini, and Copilot exhibited more consistent readability patterns, whereas ChatGPT-3.5 and Perplexity produced more linguistically complex content. Notably, DeepSeek-V3 and DeepSeek-R1 generated the most accessible responses. The findings suggest that, despite technological advances, AI-generated medical content remains insufficiently readable for general audiences, posing a potential barrier to equitable health communication. These results underscore the need for readability-aware AI design, standardized evaluation frameworks, and future research integrating quantitative readability metrics with patient-level comprehension outcomes.

Keywords: medical chatbots; readability; health communication; large language models; digital health; artificial intelligence; patient education

Academic Editors: Salvatore Gallo and Jing Jin

Received: 20 November 2025
Revised: 5 January 2026
Accepted: 27 January 2026
Published: 30 January 2026

1. Background

In recent years, there has been a marked increase in research on the readability of medical texts and educational materials intended for patients. Numerous studies have demonstrated that the comprehensibility of health information is a key factor in determining the effectiveness of communication between healthcare professionals and patients

[1–4]. A high level of linguistic complexity may limit patients' ability to interpret recommendations, make informed health decisions, and adhere to prescribed therapies [5–7].

In the scientific literature, quantitative readability indices such as the Flesch Reading Ease, Flesch–Kincaid Grade Level, Gunning Fog Index, Simple Measure of Gobbledygook, Coleman–Liau Index, and Automated Readability Index are commonly used to objectively assess the linguistic complexity of a text [8,9]. Numerous studies employing these measures have shown that health-related materials directed toward patients often exceed the recommended reading level, typically corresponding to primary or secondary education, thereby reducing their comprehensibility and practical usefulness [10–13].

At the same time, with the rapid development of natural language processing (NLP) technologies and the widespread adoption of large language models (LLMs) such as ChatGPT, Gemini, and Copilot, a new line of research has emerged focusing on the readability and comprehensibility of medical responses generated by chatbots [14–16]. Preliminary analyses indicate that although AI-generated texts often demonstrate linguistic accuracy and logical coherence, their readability and alignment with patients' health literacy levels vary substantially [17–19]. In some cases, chatbots produce overly technical or specialized messages, which may limit their educational value and potentially lead to misinterpretation or incomplete understanding of health information [20,21].

The review was guided by the following research question: To what extent do chatbot-generated medical texts comply with recommended readability standards for patient-facing health communication when evaluated using classical linguistic readability indices? This question focuses on publicly accessible chatbot outputs intended for patient education and health communication. In light of the growing body of research on chatbot readability, this review further examines whether and how such systems generate responses to medical, preventive, or educational inquiries posed by health professionals.

Recent developments in generative AI have also reshaped the conceptual understanding of how language models interact with users' health literacy needs. Earlier works in health communication emphasized structural barriers, such as excessive medical terminology, syntactic density, and low plain-language compliance, in printed materials [22–25]. However, generative LLMs introduce new challenges related to style transfer, prompt sensitivity, and the composition of training data [26–28]. Because these models are trained on large biomedical corpora, scientific preprints, and clinician-oriented resources, they tend to internalize formal and information-dense linguistic patterns. This training bias partly explains why chatbot-generated texts remain difficult for lay audiences despite their apparent fluency and coherence.

Furthermore, LLMs exhibit substantial variability in linguistic register depending on prompting strategy, system parameters, and model architecture [29]. This variability raises important methodological questions for evaluating AI-driven patient communication, including the reproducibility of readability scores, the impact of system updates, and the degree to which model fine-tuning shapes the balance between comprehensiveness and accessibility.

Together, these factors highlight the need to view readability not only as an attribute of a finished text but as an emergent property of algorithmic systems that continuously adapt during interaction. Understanding this dynamic context is essential for developing robust evaluation frameworks and for designing future AI systems capable of aligning linguistic complexity with patient literacy demands. In recent years, generative artificial intelligence in healthcare has evolved from general-purpose large language models toward domain-specific architectures designed for clinical and patient-facing applications. In particular, retrieval-augmented generation (RAG) systems, which integrate language models with external clinical knowledge bases, have become increasingly prominent in healthcare settings. Recent reviews indicate that RAG-based approaches improve factual

grounding, transparency, and domain reliability in patient education and clinical decision support tasks compared to standalone LLMs [30–33].

In parallel, multimodal generative AI systems combining text with medical imaging, laboratory data, and electronic health records are rapidly expanding across clinical domains. These developments suggest that contemporary evaluations of chatbot-generated medical texts should be interpreted within a broader ecosystem of healthcare-oriented generative AI, in which readability interacts with knowledge grounding, modality integration, and clinical specialization. Achieving this objective will synthesize existing evidence and inform the development of guidelines for designing and evaluating AI-based tools for patient communication and health education.

2. Materials and Methods

Database searches were conducted between 1 July and 30 September 2025, covering all records available up to the final search date. This study is designed as a comprehensive literature review informed by PRISMA (PRISMA 2020 checklist: EQUATOR Network) principles rather than a full PRISMA-compliant systematic review. PRISMA guidelines were used as a framework to enhance transparency in study identification, screening, and reporting; however, given the heterogeneity of study designs, chatbot models, prompts, and readability metrics, several elements required for a formal systematic review-such as quantitative synthesis and standardized risk-of-bias assessment, were not applicable [34]. A complete overview of the PRISMA checklist items and their implementation in this review is provided in Table S1 in the Supplementary Materials. The review, therefore, aims to provide a broad, structured synthesis of the current evidence rather than a statistically pooled evaluation. Each database was selected for its relevance to clinical, social, and technical research. To address potential limitations in database coverage, the search strategy was expanded to include studies that assessed text readability using established readability indices and analyzed publicly accessible chatbots available to general internet users.

Four medical databases were systematically searched: PubMed, Cochrane Library, Scopus, and Web of Science. The search strategy included the following keywords: chatbot [Title/Abstract] AND readability [Title/Abstract], chatbot [Title/Abstract] AND Flesch–Kincaid Grade Level [Title/Abstract], chatbot [Title/Abstract] AND Flesch Reading Ease [Title/Abstract], chatbot [Title/Abstract] AND Gunning Fog Index [Title/Abstract], chatbot [Title/Abstract] AND Simple Measure of Gobbledygook [Title/Abstract], chatbot [Title/Abstract] AND Coleman–Liau Index [Title/Abstract], and chatbot [Title/Abstract] AND Automated Readability Index [Title/Abstract]. No additional filters or limits were applied. A total of 4209 records were initially retrieved. After applying predefined inclusion and exclusion criteria, including language, document type, and chatbot accessibility, 140 articles were included in the review [35–174]. Figure 1 presents the PRISMA flow diagram illustrating the process of study identification, screening, eligibility assessment, and inclusion in the final review. Studies were included if they met the following criteria: (1) peer-reviewed original research articles; (2) assessment of readability of chatbot-generated medical or health-related text using at least one established quantitative readability index; (3) analysis of publicly accessible chatbot systems; and (4) focus on patient-facing or health education content. Only English-language publications were included. Studies were excluded if they evaluated proprietary or non-public chatbot systems, focused solely on technical performance without readability assessment, or did not report quantitative readability outcomes. Interrater agreement between the two independent reviewers was assessed using Cohen's kappa coefficient. Data extraction was performed independently by two reviewers using a predefined extraction framework. Extracted variables included chatbot model, medical domain, text type, readability indices applied, and reported outcomes. Any discrepancies were resolved through structured discussion until consensus

was reached; no third reviewer was involved. The obtained level of agreement was substantial (κ = 0.78), indicating high consistency in the screening and eligibility assessments.

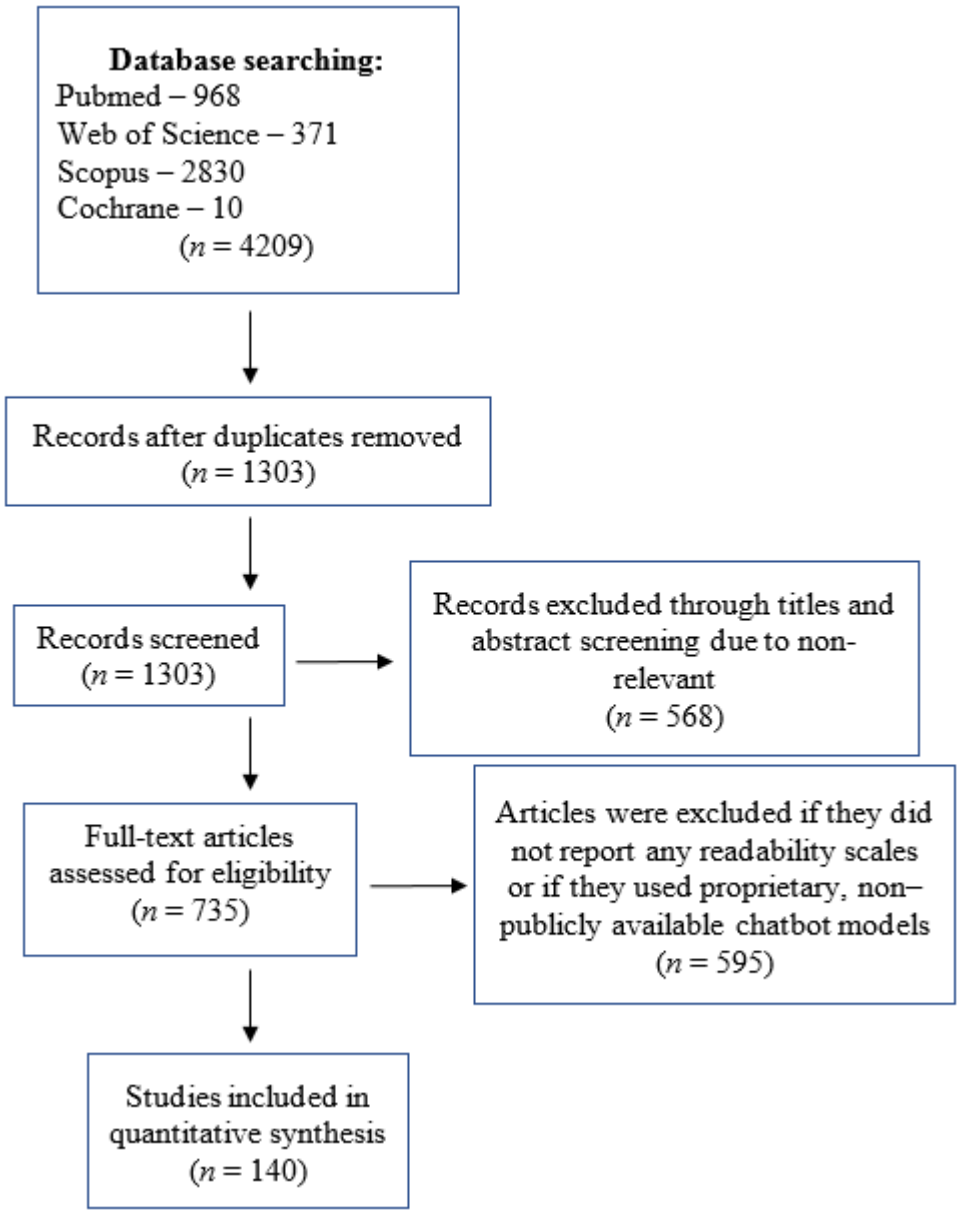

Figure 1. PRISMA Flowchart for this Review.

In addition to classical database searching, methodological attention was given to heterogeneity in prompt design, since prompting is increasingly recognized as a significant determinant of model output structure, tone, and complexity. The included studies displayed wide variation in whether prompts were phrased as open-ended questions, clinically oriented scenarios, or direct instructions to simplify language. Because readability scores are sensitive to such differences, prompt variability was treated as an important contextual factor; however, insufficient reporting of prompt formulations in the source studies precluded retrospective operationalization or stratified analysis. A meta-analysis was not conducted due to substantial heterogeneity in study designs, chatbot models, and readability outcomes, making quantitative pooling methodologically inappropriate. Across all analyzed studies, 21 chatbots and 14 readability indices were used. For consistency, chatbot nomenclature was standardized throughout the manuscript. Model

names referring to the same underlying architecture were consolidated (e.g., GPT-4 and GPT-4o), and Google Bard was treated as Gemini in studies published after the official rebranding.

Descriptive statistics were used to summarise the readability scores for each chatbot and the readability index. For every chatbot–scale pair, the mean (M) and standard deviation (SD) were calculated and reported as Mean ± SD. Calculations were performed using Python version 3.9 (Python Software Foundation, Wilmington, DE, USA). Although prompt variability is increasingly recognized as a key determinant of LLM output structure and linguistic complexity, most studies included in this review addressed prompting only descriptively. Few investigations systematically categorized prompts by intent, framing, or explicit readability constraints, thereby limiting reproducibility and cross-study comparability.

Future research would benefit from operationalizing prompt variability through standardized prompt taxonomies, for example, by distinguishing informational, instructional, reassurance-oriented, and simplification-focused prompts. Such an approach would enable clearer attribution of observed readability differences to model architecture versus prompting strategy and support longitudinal comparisons across model updates.

A formal risk-of-bias or quality assessment was not conducted due to substantial heterogeneity in study designs, chatbot architectures, prompting strategies, medical domains, and readability metrics. This heterogeneity also precluded quantitative synthesis and meta-analysis, as pooling results across fundamentally different models and outcome measures would not yield meaningful summary estimates. Instead, findings were synthesized qualitatively through comparative analysis.

3. Results

The comparative analysis revealed notable variability in the readability of chatbot-generated texts across models and readability indices. Overall, most chatbots produced content that would require at least a secondary or early tertiary education level to be fully comprehensible, suggesting that the linguistic complexity of current large language models (LLMs) remains relatively high for lay audiences. The review included studies published between 2023 and 2025, encompassing the most recent phase of research on the readability of AI-generated medical content and reflecting the rapid evolution of large language models used in healthcare communication. Supplementary Table S2 provides detailed characteristics of all included studies, including publication year, country, chatbot model, language of output, text type, readability indices used, and primary outcomes.

Most studies originated in the United States (n = 60; 42.8% of all publications) and Turkey (n = 34; 24% of all publications). Table 1 presents the complete distribution of countries from which the studies were derived, whereas Figure 2 illustrates their geographical dispersion on a world map.

Table 1. Geographical distribution of studies included in the review (n = 140).

Country	Count
USA	60
Turkey	34
China	6
India	6
Australia	5
Canada	5
Germany	3
Italy	2

Brazil	1
Denmark	2
Ireland	2
Belgium	1
Croatia	1
Egypt	1
Netherlands	1
Poland	1
Saudi Arabia	1
Singapore	1
South Korea	1
Spain	1
United Kingdom	1

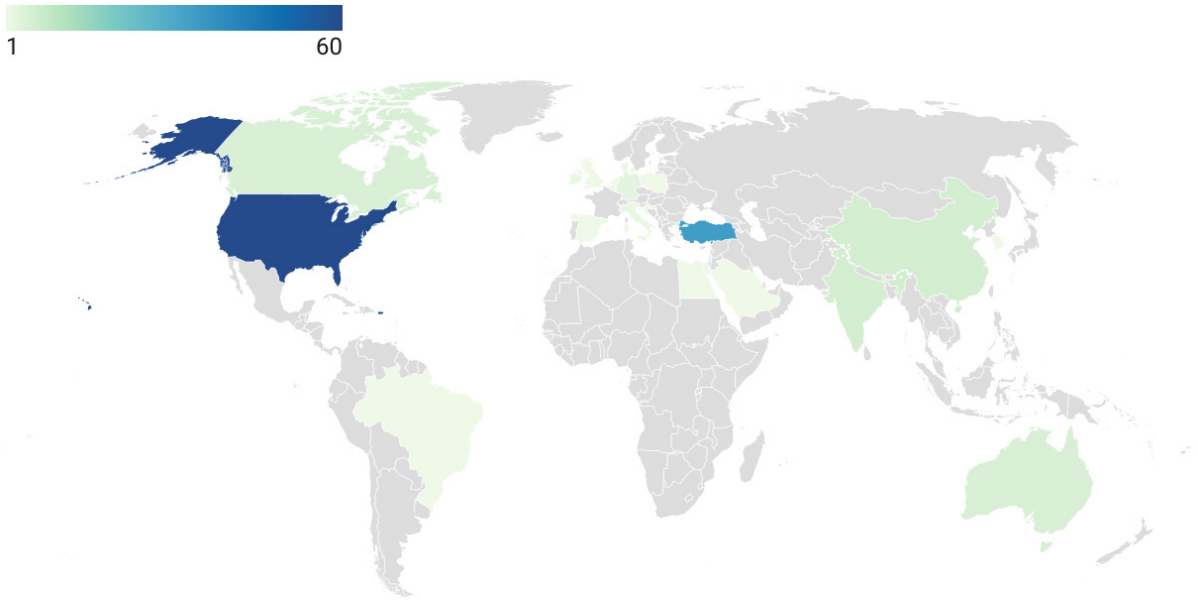

Figure 2. Global distribution of the studies included in the review.

A total of 21 chatbots were used across the publications included in the review. The most frequently used model was ChatGPT-4 (94 occurrences). The next most frequently used chatbot was ChatGPT-3.5 (83 occurrences). Table 2 presents all chatbots used in the publications included in the review.

Table 2. Chatbots analyzed across the included publications and their frequency of occurrence.

Chatbot	Count
ChatGPT-4/GPT-4o	94
ChatGPT-3.5	83
Google Bard/Gemini	52
Microsoft Copilot/Microsoft Copilot Pro/Bing AI	39
Perplexity AI/Perplexity Pro	26
Claude 2.0/Claude 3.5/Claude Sonnet	12
Meta AI Assistant	4
ChatSonic 1.0.2	3
DeepSeek-V3	2
DocsGPT 0.15.0—Changelog	2
DeepSeek-R1	2

Open Evidence 2.0	1
ChatSpot Alpha	1
DeepSeek-R1	1
Ernie Bot 4.0	1
LLaMA 3.1	1
Llama 3.1 Large	1
MediSearch Version 1.5.10	1
Pi AI 1.0.53	1
Vello	1
Vello Pro	1

The most frequently used readability measure across the analyzed publications was the Flesch–Kincaid Grade Level (used 117 times), followed by the Flesch Reading Ease Score (used 94 times). Table 3 presents all readability indices and the frequency of their use across the included studies.

Table 3. Readability indices used in the included studies and frequency of their application.

Readability Scale	Count
Flesch–Kincaid Grade Level	117
Flesch Reading Ease Score	95
Gunning Fog Index	41
Simple Measure of Gobbledygook	39
Coleman–Liau Index	22
Automated Readability Index	14
FORCAST	4
Dale–Chall Readability	3
Fry Readability Graph	2
Fry Readability Score	2
Läsbarhetsindex	2
Linsear Write	2
Raygor Readability Estimate	2
Lix Readability Index	1

The most frequently addressed topics in the chatbot queries were Patient Education/Health Communication (18 occurrences), followed by Oncology/Cancer (15 occurrences) and Otolaryngology (13 occurrences). Figure 3 presents a tabular distribution of all medical fields covered in the analyzed publications.

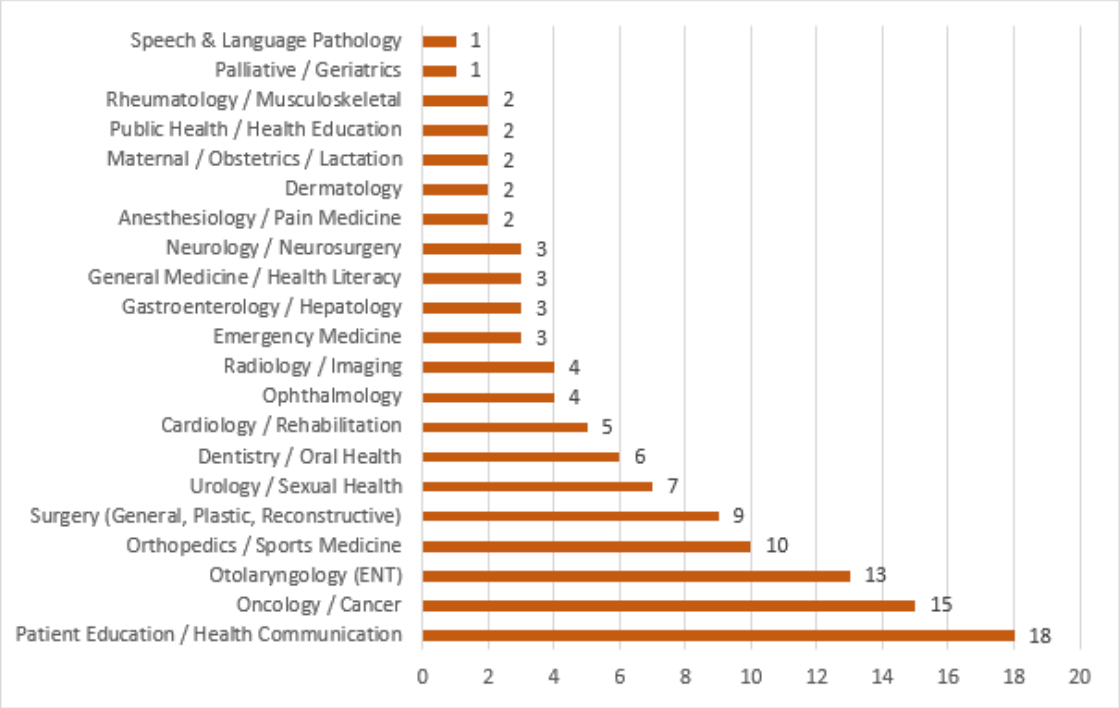

Figure 3. Medical fields covered by chatbot readability studies, grouped by topic category.

Readability Patterns Across Medical Specialities

Distinct readability patterns emerged across different medical domains. Topics such as oncology, cardiology, neurology, and orthopaedics exhibited consistently higher grade-level scores across multiple chatbot models. These fields are characterized by dense terminology, abstract pathophysiological concepts, and complex treatment algorithms, all of which tend to increase syntactic complexity and average sentence length. In contrast, domains such as patient education, public health, and maternal care yielded comparatively lower readability scores. These topics typically rely on more narrative, instruction-based language that is easier for LLMs to simplify.

Notably, oncology-related responses demonstrated some of the highest complexity values in the dataset. This may reflect both the inherent difficulty of the domain and LLMs' tendency to adopt cautious, legally conservative phrasing when discussing high-risk clinical conditions. Similarly, cardiology questions frequently elicited long, multi-clause sentences with numerous modifiers, suggesting that models may emphasize completeness over accessibility when addressing conditions perceived as clinically severe.

These speciality-level differences underscore the importance of contextualizing readability within the content domain, as the same model can yield dramatically different linguistic structures across clinical topics. Models producing lower grade-level estimates on one scale tended to score similarly across the others, reinforcing the robustness of the observed ranking patterns.

As shown in Table 4, readability scores vary significantly across chatbot models and readability indices, with most results exceeding the recommended 8th-grade level for patient-facing materials. Among the most frequently analyzed models, ChatGPT-4, Google Gemini, and Microsoft Copilot demonstrated the most balanced readability profiles. Their texts generally fell within the "difficult" category of the Flesch Reading Ease scale and corresponded to approximately college-level reading difficulty. These models showed relatively low variation across scales, indicating a consistent language structure and stable readability performance.

Table 4. Readability scores of medical texts generated by chatbots.

Chatbot	Flesch Reading Ease	Flesch–Kincaid Grade Level	Gunning Fog Index	SMOG Index	Coleman–Liau Index	Automated Readability Index	Linsear Write	Dale-Chall Score	FORCAST	Fry Graph	Fry Readability Score	Lesbarhetsindex	Lix Readability Index	Raygor Estimate
ChatGPT-4	37.55 ± 17.76	13.85 ± 8.10	14.49 ± 3.60	12.94 ± 2.74	14.61 ± 2.91	11.67 ± 2.38	9.61 ± 2.33	9.90	12.60 ± 0.42	13.55 ± 0.64	9.50 ± 0.71	36.49 ± 38.91	72.00	13.80 ± 0.28
ChatGPT-3.5	35.16 ± 13.59	15.45 ± 8.78	15.57 ± 3.26	13.11 ± 1.92	15.43 ± 2.16	14.06 ± 1.62	13.95 ± 1.81	10.25 ± 0.35	12.48 ± 0.12					
Microsoft Copilot	35.66 ± 12.01	13.66 ± 8.02	14.57 ± 2.94	13.64 ± 2.87	14.25 ± 2.38	11.95 ± 2.20	11.90 ± 1.27	10.30	12.30					
Google Gemini	39.61 ± 14.73	13.14 ± 8.31	14.29 ± 4.13	12.65 ± 2.41	13.33 ± 2.66	11.23 ± 2.45	11.71 ± 2.39	11.60	11.21 ± 1.41					
Perplexity	31.31 ± 11.27	19.62 ± 13.51	16.58 ± 2.63	14.02 ± 2.40	14.68 ± 2.07	14.06 ± 3.14	14.76 ± 5.11							
Meta AI	28.38 ± 21.83	11.97 ± 1.79	11.60	12.40	19.10	13.50			13.80					
Claude	40.11 ± 21.18	11.22 ± 2.87	10.31			10.31								
PiAI	16.30	15.90	20.00					11.90						
DeepSeek-V3	53.35 ± 7.00	8.45 ± 0.35		16.40	15.10									
ChatSpot	23.10	15.00	18.20					11.30						
DeepSeek	76.43			12.26	15.40									
DocsGPT	72.00	9.75 ± 5.73		12.10										
Llama 3.1 Large	20.10	24.10												
Llama 3.1	23.70	34.20												
Ernie Bot 4.0	37.50	12.90												
DeepSeek-R1	61.40	7.20												
MediSearch				18.30										
ChatSonic		21.65 ±												

	16.77	
Open Evidence		17.09 ± 0.56
Vello	29.00	
Vello Pro	17.40	

ChatGPT-3.5 and Perplexity, in contrast, generated content characterized by higher linguistic complexity, with longer sentences and more specialized vocabulary. Both models consistently scored higher on grade-level indices, implying that the information they produced would be challenging for audiences with average health literacy. Within the GPT family, the transition from version 3.5 to 4 was accompanied by a measurable improvement in readability, suggesting refinements in language coherence and sentence simplification in the newer model.

Models such as Claude and Meta AI showed intermediate readability, with scores fluctuating between moderate and strenuous across the scales. This variability likely reflects the heterogeneity of available prompts and text domains used in the analyzed studies.

DeepSeek-V3 and DeepSeek-R1 were the only models to produce outputs classified as readable or moderately easy, with text difficulty levels approximating those recommended for patient information materials. Their consistently lower grade-level scores suggest that these models may prioritize shorter sentences and simpler word choice, making them more accessible to a general audience.

Smaller or domain-specific chatbots, such as DocsGPT, PiAI, ChatSpot, Vello, and Open Evidence, were represented in fewer studies and across fewer readability indices. While their readability estimates varied widely, these systems tended to exhibit higher linguistic variability and less consistent results, likely due to narrower training data and differing use cases.

Chatbots that achieved higher grade-level scores on indices such as the Flesch–Kincaid Grade Level, Gunning Fog, or Linsear Write generally exhibited lower values on the Flesch Reading Ease scale. This alignment indicates that the indices captured similar dimensions of linguistic complexity, providing a coherent overall picture of relative readability across chatbot-generated texts. Figure 4 presents a heatmap demonstrating the comparative distribution of 14 readability metrics across 21 AI chatbots. Missing data reflects incomplete reporting in source studies. Darker colours indicate lower values, while yellow-green shades indicate higher values.

Lower values on grade-level indices (e.g., Flesch–Kincaid Grade Level, Gunning Fog Index) indicate greater readability. In contrast, higher values on the Flesch Reading Ease scale correspond to easier-to-read text. Scale directionality is explicitly indicated to facilitate interpretation by readers unfamiliar with readability metrics.

We also summarized the citation impact of all included publications. Figure 5 presents the 20 most cited articles in the dataset, while the complete citation ranking of all 140 studies is provided in the Supplementary Materials Table S3.

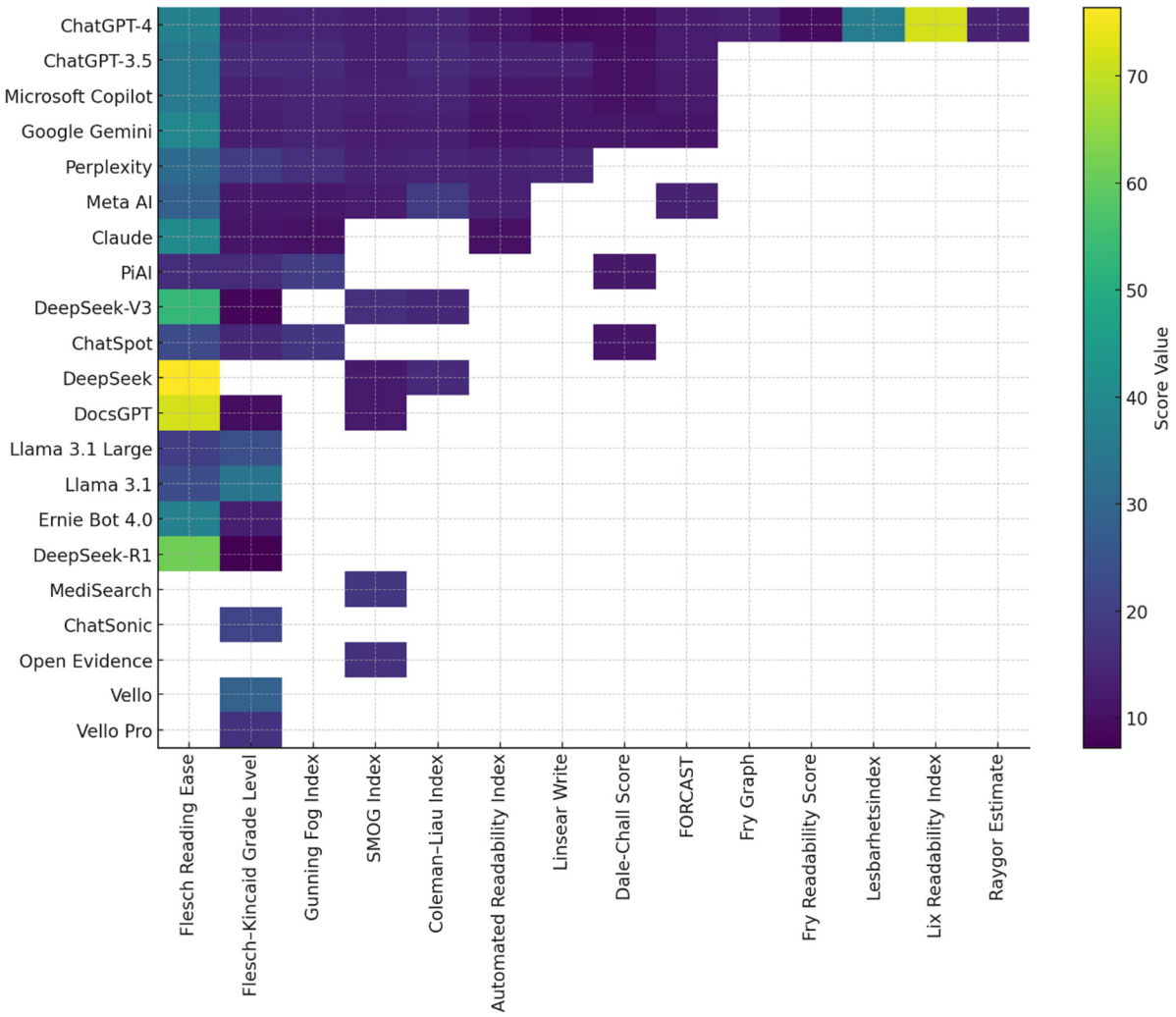

Figure 4. Comparative Heatmap of 14 Readability Metrics Across 21 AI Chatbots.

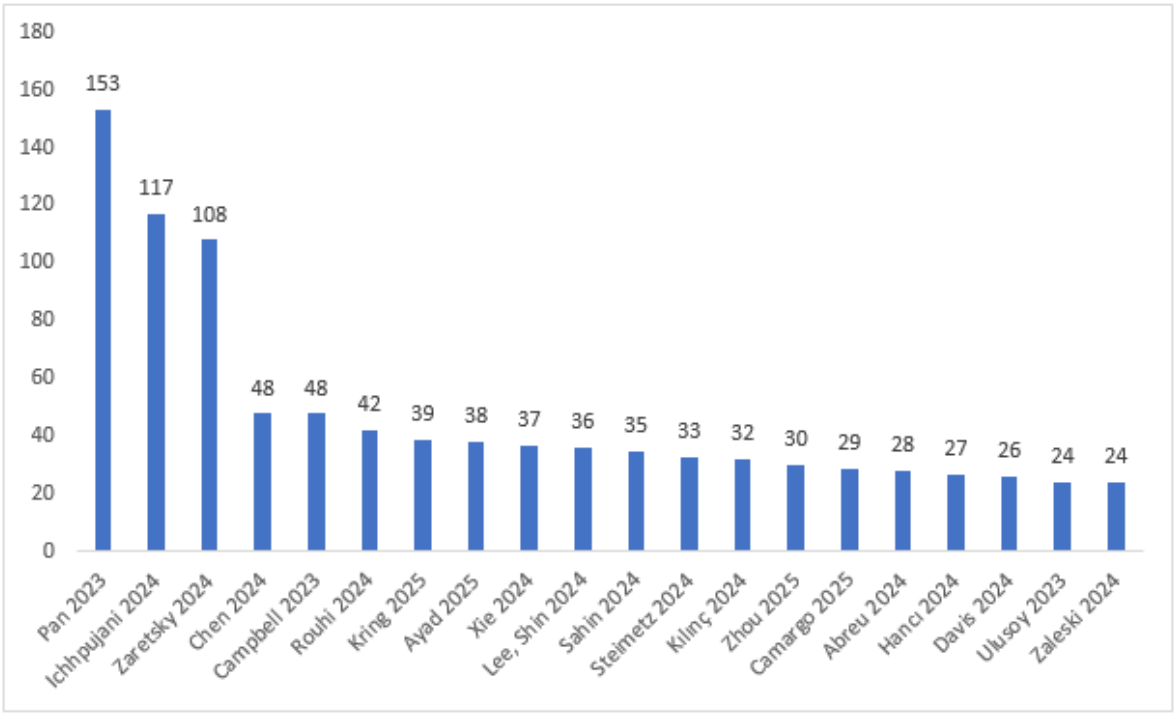

Figure 5. Top 20 Most Cited Publications Included in the Review.

4. Discussion

This review provides a comprehensive synthesis of existing studies assessing the readability of chatbot-generated medical texts using classical linguistic indices. In this review, readability is treated as a patient-facing communication dimension of AI-generated medical content, evaluated under the assumption of baseline informational correctness and considered complementary to, rather than a replacement for, accuracy and clinical validity assessments [175–177]. Readability should not be viewed as a purely linguistic attribute, as excessive textual complexity in healthcare contexts may directly compromise patient safety, healthcare reliability, and decision-making accuracy. AI-generated medical information that is difficult to read may increase the risk of misinterpreting treatment instructions, overlooking contraindications, or misunderstanding probabilistic risk information [178].

Significantly, readability interacts with known failure modes of generative AI systems, including hallucinations, overgeneralization, and omission of uncertainty markers. Linguistically dense or overly formal responses may obscure hedging statements and limitations, potentially fostering unwarranted trust in incorrect or incomplete information [179]. From this perspective, readability constitutes a core dimension of responsible AI deployment in healthcare, alongside accuracy, transparency, and domain alignment.

Recent domain-specific reviews further reinforce the importance of contextual grounding and clinical specialization in generative AI for healthcare. A comprehensive review of retrieval-augmented generation in healthcare suggests that grounding model outputs in curated clinical sources not only improves factual accuracy but may also constrain response scope, thereby indirectly enhancing communicative clarity [180–182]. Similarly, longitudinal analyses of generative AI applications in health care illustrate how domain-specific fine-tuning and guideline integration shape both informational quality and accessibility [183,184]. These findings suggest that readability assessments should explicitly account for whether standalone LLMs or augmented architectures generate chatbot responses, as this distinction may systematically influence linguistic complexity and clinical appropriateness.

While a growing number of publications have explored factual accuracy, empathy, or the reliability of AI-driven health information, the fundamental issue of linguistic accessibility has remained largely underexamined. By consolidating findings from 140 studies across 21 chatbot models, this review provides a comprehensive overview of the readability of chatbot-generated medical texts using classical linguistic indices.

Earlier research on online health communication—long before the advent of generative AI—consistently showed that most patient education materials were written at a level too advanced for the general population, typically above the 8th-grade level recommended by the American Medical Association and the U.S. Department of Health and Human Services [185,186]. Studies on web-based patient portals and hospital websites confirmed similar patterns, revealing that even materials intended for public education often demand college-level literacy.

Recent studies investigating chatbot-generated content, though limited in number and scope, have echoed these concerns. For example, it was reported that ChatGPT and Bard produced health information with Flesch–Kincaid Grade Levels of 12–14, substantially above recommended thresholds [187,188]. Similarly, another study found that ChatGPT's answers regarding cardiovascular health were syntactically correct but lexically dense, often employing specialized terminology [189–192]. The present review confirms and extends these observations by aggregating evidence across multiple models and domains, demonstrating that the issue of excessive linguistic complexity is systemic rather than model-specific.

However, some studies have suggested that newer model iterations, such as GPT-4, tend to produce slightly simpler, more structured responses than earlier versions, such as GPT-3.5 [193,194]. This review provides converging evidence for this trend, indicating incremental but insufficient progress toward readability improvement. These findings collectively suggest that advances in model architecture alone do not guarantee improved accessibility for end users without deliberate optimization for readability.

This variability is not merely linguistic but reflects system-level technical choices that shape the generated text. Readability in LLM-generated medical text should therefore be interpreted as a downstream outcome shaped by these design decisions rather than as an inherent property of a model label. Variation in readability across studies may reflect differences in decoding strategies (e.g., temperature, sampling constraints, output length limits), prompt and instruction design (e.g., explicit simplification constraints, disclaimer requirements), and alignment objectives [195]. In particular, safety-optimized alignment procedures (including RLHF) can promote conservative phrasing, hedging, and extensive disclaimers, which may increase sentence length and syntactic complexity. Conversely, instruction tuning that prioritizes clarity and user comprehension may yield more concise, accessible outputs. Retrieval-augmented generation further complicates interpretation: while retrieval can improve factual grounding, it may also introduce domain-specific terminology and longer guideline-like responses that inflate grade-level estimates [196]. These mechanisms imply that readability comparisons between chatbots are not causally interpretable without standardized reporting of technical parameters and interaction settings.

To enable technically meaningful interpretation and cross-study comparability, future readability evaluations should routinely report a minimal set of system-level indicators: (i) model identifier, version, and date of access; (ii) complete prompt templates and instruction constraints (including system prompts where available); (iii) decoding parameters and output-length settings; (iv) retrieval/tool-use configuration (if applicable); (v) interaction design (single-turn vs. multi-turn, context length, memory settings); and (vi) post-processing or safety filtering applied to responses [197–200]. The limited reporting of such parameters in most of the existing literature constitutes a significant methodological barrier to linking readability outcomes to specific LLM techniques and to establishing reproducible benchmarks [201].

The observed variability in readability across models likely stems from architectural and training differences. GPT-3.5 and Perplexity frequently produced longer and more syntactically intricate sentences, consistent with their tendency to generate verbose, detail-heavy responses. GPT-4 and Gemini, although more consistent, still align with formal scientific prose because their training corpora heavily represent academic texts. In contrast, DeepSeek-V3 and DeepSeek-R1-models intentionally optimized for brevity—generated significantly shorter sentences and simpler vocabulary. This suggests that model alignment strategies and fine-tuning objectives play a decisive role in shaping linguistic accessibility.

It should also be acknowledged that many chatbots evaluated in the included studies may rely on shared large language model backends, common APIs, or similar corporate infrastructures, despite being presented as distinct systems. Consequently, observed differences in readability across chatbot labels may reflect variations in prompting strategies, interface design, or response formatting rather than fundamental differences in underlying model technologies [202,203].

An additional factor is the influence of reinforcement learning from human feedback (RLHF), which may inadvertently increase linguistic complexity by promoting cautious, formal, and legally conservative phrasing. Systems optimized primarily for safety or factual correctness may therefore produce verbose outputs (e.g., hedging or extensive

disclaimers), whereas models fine-tuned with objectives emphasizing instructional clarity tend to generate more patient-friendly text. These findings support the need for fine-tuning pipelines that explicitly include readability as a core performance metric [204]. The persistence of high reading difficulty in chatbot-generated health communication underscores a broader challenge: technological sophistication does not automatically translate into information that patients can understand.

To address this, future chatbot design should incorporate mechanisms to monitor and adapt readability, including real-time complexity assessment and model optimization strategies that prioritize clarity over verbosity. Moreover, interdisciplinary collaboration between computer scientists, linguists, and health communication experts will be essential to ensure that AI systems are optimized not only for accuracy but also for comprehension and inclusivity.

Traditional readability metrics, while valuable, measure only the surface structure of text, such as sentence length, syllable count, and syntactic density. They do not capture semantic transparency, contextual coherence, or pragmatic appropriateness, all of which shape actual understanding. Several recent works have emphasized that comprehension depends on both linguistic and cognitive accessibility, including familiarity with medical terminology and the perceived credibility of the source [204–207].

Several high-impact studies within the dataset provide significant insights into how LLMs handle medical communication. For example, studies evaluating oncology- and cardiology-related materials demonstrated that even state-of-the-art models struggled to reach recommended reading levels, often producing content equivalent to college-level difficulty [208]. Research on low back pain, cataract surgery, and thyroid disorders found that although LLMs offer coherent, structurally organized explanations, they often introduce specialized terminology without simplifying or contextualizing it for lay readers [209–214]. Notably, several investigations comparing AI-generated content with expert-written materials revealed that AI models can surpass clinicians in structural clarity but still fall short in accessibility. This finding reinforces the duality between linguistic fluency and accurate readability [215–218].

These landmark studies collectively suggest that readability challenges are systemic across models and domains rather than isolated incidents. Their conclusions emphasize the need for computational approaches that extend beyond classical metrics toward more holistic, patient-centred evaluation frameworks. The geographic concentration of readability research in English-speaking or high-income countries limits the generalizability of findings. Chatbots operating in languages with complex morphology (e.g., Polish, Turkish, or Korean) may exhibit different readability patterns due to linguistic structure and translation effects. Expanding this line of inquiry to multilingual and multicultural contexts is therefore crucial to understanding global variations and equity implications. At the policy level, the findings highlight the need for evidence-based standards for AI-generated health communication, analogous to readability guidelines for printed materials. Institutions such as the WHO or national health agencies could issue frameworks defining acceptable linguistic thresholds for AI-based public health tools, ensuring that emerging technologies align with accessibility principles.

Given the multiple factors influencing chatbot-generated responses, including model architecture, prompting strategies, knowledge base design, and regional context, the statistical results summarized in this review should be interpreted descriptively rather than causally. Their reliability lies in the consistency of observed readability patterns across multiple studies and indices, not in precise attribution to specific technological components.

5. Future Directions

5.1. Dynamic, Readability-Aware Text Generation

Future LLMs should incorporate real-time control mechanisms that allow users or healthcare providers to specify a target readability range (e.g., FKGL 6–8). Such systems could include built-in constraints on sentence length, lexical complexity, and structural density, enabling models to adapt dynamically to each patient's literacy level. Integrating these features into user-facing interfaces would substantially improve the accessibility of AI-driven health communication.

5.2. Beyond Surface Metrics: Hybrid Readability Models

Classical readability indices capture syntactic and lexical features but fail to assess semantic transparency or conceptual load. Combining traditional metrics with embedding-based semantic measures, such as contextual coherence or terminology familiarity, would create more comprehensive tools for evaluating patient comprehension. Future research should explore hybrid frameworks that combine rule-based and machine-learning indicators to capture the multifactorial nature of readability.

5.3. Cross-Linguistic and Cross-Cultural Readability Evaluation

Most studies included in this review focused on English-language outputs, limiting the generalizability of findings. Languages with complex morphology, such as Polish, Turkish, Korean, or Finnish, may exhibit different readability patterns due to inflectional structure and word length. Expanding research to multilingual contexts is crucial for ensuring equitable access to AI-generated health information and for identifying cultural and linguistic factors that modulate readability.

5.4. User-Based Comprehension Studies

A critical next step involves shifting from purely text-based metrics to patient-centred comprehension research. Randomized controlled studies assessing users' understanding, recall, and decision-making accuracy after reading AI-generated texts would provide more actionable insights into real-world usability. Combining these behavioural outcomes with readability indices would help validate whether improvements in linguistic complexity translate into meaningful gains in patient comprehension.

An additional limitation of the current literature concerns the limited stratification of readability outcomes by use-case category. Chatbot-generated medical texts serve heterogeneous functions, including general health education, preventive counselling, disease-specific self-management, and post-discharge instructions. These use cases differ substantially in their tolerance for ambiguity, acceptable linguistic complexity, and clinical risk.

Aggregating readability scores across heterogeneous use cases may therefore obscure clinically meaningful differences and limit interpretability. Future evaluations should classify chatbot outputs into functional use-case categories to better align readability assessments with real-world healthcare applications.

6. Limitations and Strengths

Several limitations should be acknowledged. First, the review was restricted to studies that reported quantitative readability metrics and analyzed publicly available chatbot models. As a result, it may have excluded unpublished or domain-specific evaluations, particularly those conducted within clinical settings or using proprietary systems. Second, the included studies varied in methodology, prompt design, and thematic focus, limiting direct comparability and precluding meta-analytic synthesis. Some discrepancies in readability scores may therefore reflect differences in prompt structure rather than actual

model variation. Third, the findings should be interpreted with the understanding that classical readability indices capture surface linguistic features rather than semantic or cognitive comprehension. Fourth, the geographical and linguistic concentration of existing research (predominantly in English and in high-income countries) limits the generalizability of conclusions to other languages and health systems.

First, it is the first synthesis of studies assessing the readability of chatbot-generated medical texts across a wide range of models, indices, and medical domains. By including 140 publications and systematically analyzing 21 chatbots and 14 readability measures, the review offers a broad overview of how AI communicates health information to lay audiences. Second, the inclusion of multiple readability indices and cross-model comparisons enhances methodological robustness and interpretive depth. The convergence of findings across different indices (e.g., Flesch–Kincaid, SMOG, and Gunning Fog) strengthens the validity of observed trends and supports the reliability of the overall conclusions. Third, the study offers a clear conceptual framework for future investigations by linking linguistic readability with broader issues of health literacy, digital equity, and responsible AI design. This interdisciplinary perspective situates the findings not only within computational linguistics but also within public health and communication research, making the results relevant for both technical and health policy audiences.

7. Conclusions

This review is, to our knowledge, among the first to systematically synthesise evidence on the readability of chatbot-generated medical content. Despite advances in AI language models, most outputs remain too complex for typical patient audiences, highlighting a persistent communication gap. Readability should therefore be treated as a key quality criterion in the design and evaluation of health chatbots. Our findings highlight an emerging risk that general-purpose AI models may unintentionally widen the health communication gap unless readability-aware safety controls become standard in clinical and public-facing AI systems. Based on the reviewed evidence, future evaluations of AI-generated medical content should routinely report a minimum core set of readability indices and explicitly document the prompting strategies used. In addition, implementation of generative AI tools in healthcare should incorporate readability assessment and user testing as standard components of validation. At the policy level, public health agencies may consider developing guidelines and standards for readability in AI-generated patient communication. Future work should integrate standardized readability assessment with user-based comprehension testing as a routine component of evaluating AI-generated patient communication.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app16031423/s1, Table S1: Completed PRISMA 2020 Checklist; Table S2: Characteristics of all included studies. Table S3: Complete citation ranking of all 140 studies.

Author Contributions: Conceptualization, J.B. and R.O.; methodology, J.B.; software, K.W.; validation, R.O. and J.R.; formal analysis, K.W.; investigation, J.B.; resources, J.B.; data curation, J.B.; writing—original draft preparation, J.B., R.O. and K.W.; writing—review and editing, J.B., R.O. and J.R.; visualization, K.W.; supervision, R.O. and J.R.; project administration, J.B.; funding acquisition, none. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Some or all of the framework features that support the findings of this study are available from the corresponding author upon reasonable request.

https://doi.org/10.3390/app16031423

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fitzpatrick, P.J. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. *Front. Digit. Health* **2023**, *5*, 1264780. https://doi.org/10.3389/fdgth.2023.1264780.
2. Sharkiya, S.H. Quality communication can improve patient-centred health outcomes among older patients: A rapid review. *BMC Health Serv. Res.* **2023**, *23*, 886. https://doi.org/10.1186/s12913-023-09869-8.
3. Chen, X.; Hay, J.L.; Waters, E.A.; Kiviniemi, M.T.; Biddle, C.; Schofield, E.; Li, Y.; Kaphingst, K.; Orom, H. Health Literacy and Use and Trust in Health Information. *J. Health Commun.* **2018**, *23*, 724–734. https://doi.org/10.1080/10810730.2018.1511658.
4. Kwame, A.; Petrucka, P.M. A literature-based study of patient-centered care and communication in nurse-patient interactions: Barriers, facilitators, and the way forward. *BMC Nurs.* **2021**, *20*, 158. https://doi.org/10.1186/s12912-021-00684-2.
5. Al Shamsi, H.; Almutairi, A.G.; Al Mashrafi, S.; Al Kalbani, T. Implications of Language Barriers for Healthcare: A Systematic Review. *Oman Med. J.* **2020**, *35*, e122. https://doi.org/10.5001/omj.2020.40.
6. Coughlin, S.S.; Vernon, M.; Hatzigeorgiou, C.; George, V. Health Literacy, Social Determinants of Health, and Disease Prevention and Control. *J. Environ. Health Sci.* **2020**, *6*, 3061.
7. Pandey, M.; Maina, R.G.; Amoyaw, J.; Li, Y.; Kamrul, R.; Michaels, C.R.; Maroof, R. Impacts of English language proficiency on healthcare access, use, and outcomes among immigrants: A qualitative study. *BMC Health Serv. Res.* **2021**, *21*, 741. https://doi.org/10.1186/s12913-021-06750-4.
8. Yeung, A.W.K.; Goto, T.K.; Leung, W.K. Readability of the 100 Most-Cited Neuroimaging Papers Assessed by Common Readability Formulae. *Front. Hum. Neurosci.* **2018**, *12*, 308. https://doi.org/10.3389/fnhum.2018.00308.
9. Nash, E.; Bickerstaff, M.; Chetwynd, A.J.; Hawcutt, D.B.; Oni, L. The readability of parent information leaflets in paediatric studies. *Pediatr. Res.* **2023**, *94*, 1166–1171. https://doi.org/10.1038/s41390-023-02608-z.
10. Brega, A.G.; Freedman, M.A.; LeBlanc, W.G.; Barnard, J.; Mabachi, N.M.; Cifuentes, M.; Albright, K.; Weiss, B.D.; Brach, C.; West, D.R. Using the Health Literacy Universal Precautions Toolkit to Improve the Quality of Patient Materials. *J. Health Commun.* **2015**, *20*, 69–76. https://doi.org/10.1080/10810730.2015.1081997.
11. Rooney, M.K.; Santiago, G.; Perni, S.; Horowitz, D.P.; McCall, A.R.; Einstein, A.J.; Jagsi, R.; Golden, D.W. Readability of Patient Education Materials from High-Impact Medical Journals: A 20-Year Analysis. *J. Patient Exp.* **2021**, *8*, 2374373521998847. https://doi.org/10.1177/2374373521998847.
12. Eltorai, A.E.; Ghanian, S.; Adams, C.A., Jr.; Born, C.T.; Daniels, A.H. Readability of patient education materials on the american association for surgery of trauma website. *Arch. Trauma. Res.* **2014**, *3*, e18161. https://doi.org/10.5812/atr.18161.
13. Badarudeen, S.; Sabharwal, S. Assessing readability of patient education materials: Current role in orthopaedics. *Clin. Orthop. Relat. Res.* **2010**, *468*, 2572–2580. https://doi.org/10.1007/s11999-010-1380-y.
14. Geantă, M.; Bădescu, D.; Chirca, N.; Nechita, O.C.; Radu, C.G.; Rascu, Ș.; Rădăvoi, D.; Sima, C.; Toma, C.; Jinga, V. The Emerging Role of Large Language Models in Improving Prostate Cancer Literacy. *Bioengineering* **2024**, *11*, 654. https://doi.org/10.3390/bioengineering11070654.
15. Demir, G.; Sevri, M.; Hacıosmanoğlu, C.D.; Büyüktaşkın, D.; Özaslan, A. Comparative Evaluation of Large Language Models in Addressing Autism-Related Information Queries: Insights from ChatGPT, Gemini, and Copilot. *Gazi Med. J.* **2025**, *36*, 407–416. https://doi.org/10.12996/gmj.2025.4451.
16. Bolgova, O.; Ganguly, P.; Mavrych, V. Comparative analysis of LLMs performance in medical embryology: A cross-platform study of ChatGPT, Claude, Gemini, and Copilot. *Anat. Sci. Educ.* **2025**, *18*, 718–726. https://doi.org/10.1002/ase.70044.
17. Swisher, A.R.; Wu, A.W.; Liu, G.C.; Lee, M.K.; Carle, T.R.; Tang, D.M. Enhancing Health Literacy: Evaluating the Readability of Patient Handouts Revised by ChatGPT's Large Language Model. *Otolaryngol. Head Neck Surg.* **2024**, *171*, 1751–1757. https://doi.org/10.1002/ohn.927.
18. Nasra, M.; Jaffri, R.; Pavlin-Premrl, D.; Kok, H.K.; Khabaza, A.; Barras, C.; Slater, L.A.; Yazdabadi, A.; Moore, J.; Russell, J.; et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. *Intern. Med. J.* **2025**, *55*, 20–34. https://doi.org/10.1111/imj.16607.
19. Kirchner, G.J.; Kim, R.Y.; Weddle, J.B.; Bible, J.E. Can Artificial Intelligence Improve the Readability of Patient Education Materials? *Clin. Orthop. Relat. Res.* **2023**, *481*, 2260–2267. https://doi.org/10.1097/corr.0000000000002668.
20. Mokmin, N.A.M.; Ibrahim, N.A. The evaluation of chatbot as a tool for health literacy education among undergraduate students. *Educ. Inf. Technol.* **2021**, *26*, 6033–6049. https://doi.org/10.1007/s10639-021-10542-y.

21. Sezer, B.; Aydoğdu, T. Performance of Advanced Artificial Intelligence Models in Traumatic Dental Injuries in Primary Dentition: A Comparative Evaluation of ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7 in Terms of Accuracy, Completeness, Response Time, and Readability. *Appl. Sci.* **2025**, *15*, 7778. https://doi.org/10.3390/app15147778.
22. Tilton AK, Caplan BE, Cole BJ. Generative AI in consumer health: leveraging large language models for health literacy and clinical safety with a digital health framework. Front Digit Health. 2025 Aug 26;7:1616488. doi: 10.3389/fdgth.2025.1616488.
23. Randell, R.L.; Wilson, H.P.; Ragavan, M.I.; Collins, A.B.; Vail, J.; Ramirez, S.; Amodei, J.; Mickievicz, E.; Krieger, M.S.; Macon, E.C.; et al. Communicating Health Research with Plain Language. *Inq. J. Health Care Organ. Provis. Financ.* **2025**, *62*, 469580251357755. https://doi.org/10.1177/00469580251357755.
24. Giguère, A.; Zomahoun, H.T.V.; Carmichael, P.H.; Uwizeye, C.B.; Légaré, F.; Grimshaw, J.M.; Gagnon, M.P.; Auguste, D.U.; Massougbodji, J. Printed educational materials: Effects on professional practice and healthcare outcomes. *Cochrane Database Syst. Rev.* **2020**, *8*, CD004398. https://doi.org/10.1002/14651858.CD004398.pub4.
25. Yu, P.; Xu, H.; Hu, X.; Deng, C. Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare* **2023**, *11*, 2776. https://doi.org/10.3390/healthcare11202776.
26. Chen, B.; Zhang, Z.; Langrené, N.; Zhu, S. Unleashing the potential of prompt engineering for large language models. *Patterns* **2025**, *6*, 101260. https://doi.org/10.1016/j.patter.2025.101260.
27. Reddy, S. Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implement. Sci.* **2024**, *19*, 27. https://doi.org/10.1186/s13012-024-01357-9.
28. Warde, F.; Papadakos, J.; Papadakos, T.; Rodin, D.; Salhia, M.; Giuliani, M. Plain language communication as a priority competency for medical professionals in a globalized world. *Can. Med. Educ. J.* **2018**, *9*, e52–e59.
29. Delgado-Chaves, F.M.; Jennings, M.J.; Atalaia, A.; Wolff, J.; Horvath, R.; Mamdouh, Z.M.; Baumbach, J.; Baumbach, L. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2411962122. https://doi.org/10.1073/pnas.2411962122.
30. Yang, S.; Jing, M.; Wang, S.; Huang, Z.; Wang, J.; Kou, J.; Shi, M.; Xia, Z.; Wei, Q.; Xing, W.; et al. Building trustworthy large language model-driven generative recommender system for healthcare decision support: A scoping review of corpus sources, customization techniques, and evaluation frameworks. *Artif. Intell. Med.* **2026**, *171*, 103310. https://doi.org/10.1016/j.artmed.2025.103310.
31. Ozmen, B.B.; Singh, N.; Shah, K.; Berber, I.; Singh, D.; Pinsky, E.; Schulz, S.A.; Bishop, S.N.; Bernard, S.; Djohan, R.S.; et al. MicroRAG: Development of a Novel Artificial Intelligence Retrieval-Augmented Generation Model for Microsurgery Clinical Decision Support. *Microsurgery* **2025**, *45*, e70138. https://doi.org/10.1002/micr.70138.
32. Amugongo, L.M.; Mascheroni, P.; Brooks, S.; Doering, S.; Seidel, J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLoS Digit. Health* **2025**, *4*, e0000877. https://doi.org/10.1371/journal.pdig.0000877.
33. Busch, F.; Kaibel, L.; Nguyen, H.; Lemke, T.; Ziegelmayer, S.; Graf, M.; Marka, A.W.; Endrös, L.; Prucker, P.; Spitzl, D.; et al. Evaluation of a Retrieval-Augmented Generation-Powered Chatbot for Pre-CT Informed Consent: A Prospective Comparative Study. *J. Imaging Inform. Med.* **2025**, *38*, 4312–4323. https://doi.org/10.1007/s10278-025-01483-w.
34. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. https://doi.org/10.1136/bmj.n71.
35. Yurdakurban, E.; Topsakal, K.G.; Duran, G.S. A comparative analysis of AI-based chatbots: Assessing data quality in orthognathic surgery related patient information. *J. Stomatol. Oral Maxillofac. Surg.* **2024**, *125*, 101757. https://doi.org/10.1016/j.jormas.2023.101757.
36. Camargo, E.S.; Quadras, I.C.C.; Garanhani, R.R.; de Araujo, C.M.; Stuginski-Barbosa, J. A Comparative Analysis of Three Large Language Models on Bruxism Knowledge. *J. Oral Rehabil.* **2025**, *52*, 896–903. https://doi.org/10.1111/joor.13948.
37. Deveci, C.D.; Baker, J.J.; Sikander, B.; Rosenberg, J. A comparison of cover letters written by ChatGPT-4 or humans. *Dan. Med. J.* **2023**, *70*, A06230412.
38. Kring, T.; Prasad, S.; Dadi, S.; Sokhn, E.; Franzmann, E. A comparison of quality and readability of Artificial Intelligence chatbots in triage for head and neck cancer. *Am. J. Otolaryngol.* **2025**, *46*, 104710. https://doi.org/10.1016/j.amjoto.2025.104710.
39. Yun, J.Y.; Kim, D.J.; Lee, N.; Kim, E.K. A comprehensive evaluation of ChatGPT consultation quality for augmentation mammoplasty: A comparative analysis between plastic surgeons and laypersons. *Int. J. Med. Inform.* **2023**, *179*, 105219. https://doi.org/10.1016/j.ijmedinf.2023.105219.

40. Carlson, J.A.; Cheng, R.Z.; Lange, A.; Nagalakshmi, N.; Rabets, J.; Shah, T.; Sindhwani, P. Accuracy and Readability of Artificial Intelligence Chatbot Responses to Vasectomy-Related Questions: Public Beware. *Cureus* **2024**, *16*, e67996. https://doi.org/10.7759/cureus.67996.
41. Halawani, A.; Mitchell, A.; Saffarzadeh, M.; Wong, V.; Chew, B.H.; Forbes, C.M. Accuracy and Readability of Kidney Stone Patient Information Materials Generated by a Large Language Model Compared to Official Urologic Organizations. *Urology* **2024**, *186*, 107–113. https://doi.org/10.1016/j.urology.2023.11.042.
42. Yau, J.Y.; Saadat, S.; Hsu, E.; Murphy, L.S.; Roh, J.S.; Suchard, J.; Tapia, A.; Wiechmann, W.; Langdorf, M.I. Accuracy of Prospective Assessments of 4 Large Language Model Chatbot Responses to Patient Questions About Emergency Care: Experimental Comparative Study. *J. Med. Internet Res.* **2024**, *26*, e60291. https://doi.org/10.2196/60291.
43. Yıldız, H.A.; Söğütdelen, E. AI Chatbots as Sources of STD Information: A Study on Reliability and Readability. *J. Med. Syst.* **2025**, *49*, 43. https://doi.org/10.1007/s10916-025-02178-z.
44. Stephan, D.; Bertsch, A.; Burwinkel, M.; Vinayahalingam, S.; Al-Nawas, B.; Kämmerer, P.W.; Thiem, D.G. AI in Dental Radiology-Improving the Efficiency of Reporting with ChatGPT: Comparative Study. *J. Med. Internet Res.* **2024**, *26*, e60684. https://doi.org/10.2196/60684.
45. Hand, C.; Bohn, C.; Tannir, S.; Ulrich, M.; Saniei, S.; Girod-Hoffman, M.; Lu, Y.; Forsythe, B. American Academy of Orthopaedic Surgeons OrthoInfo provides more readable information regarding rotator cuff injury than ChatGPT. *J. ISAKOS* **2025**, *12*, 100841. https://doi.org/10.1016/j.jisako.2025.100841.
46. Bohn, C.; Hand, C.; Tannir, S.; Ulrich, M.; Saniei, S.; Girod-Hoffman, M.; Lu, Y.; Krych, A.; Forsythe, B. American academy of Orthopedic Surgeons' OrthoInfo provides more readable information regarding meniscus injury than ChatGPT-4 while information accuracy is comparable. *J. ISAKOS* **2025**, *11*, 100843. https://doi.org/10.1016/j.jisako.2025.100843.
47. Ichhpujani, P.; Parmar, U.P.S.; Kumar, S. Appropriateness and readability of Google Bard and ChatGPT-3.5 generated responses for surgical treatment of glaucoma. *Rom. J. Ophthalmol.* **2024**, *68*, 243-248. https://doi.org/10.22336/rjo.2024.45.
48. Azzopardi, M.; Ng, B.; Logeswaran, A.; Loizou, C.; Cheong, R.C.T.; Gireesh, P.; Ting, D.S.J.; Chong, Y.J. Artificial intelligence chatbots as sources of patient education material for cataract surgery: ChatGPT-4 versus Google Bard. *BMJ Open Ophthalmol.* **2024**, *9*, e001824. https://doi.org/10.1136/bmjophth-2024-001824.
49. Gondode, P.G.; Singh, R.; Mehta, S.; Singh, S.; Kumar, S.; Nayak, S.S. Artificial intelligence chatbots versus traditional medical resources for patient education on "Labor Epidurals": An evaluation of accuracy, emotional tone, and readability. *Int. J. Obstet. Anesth.* **2025**, *61*, 104302. https://doi.org/10.1016/j.ijoa.2024.104302.
50. Pradhan, F.; Fiedler, A.; Samson, K.; Olivera-Martinez, M.; Manatsathit, W.; Peeraphatdit, T. Artificial intelligence compared with human-derived patient educational materials on cirrhosis. *Hepatol. Commun.* **2024**, *8*, e0367. https://doi.org/10.1097/HC9.0000000000000367.
51. Ayad, O.; Yassa, A.; Patel, A.M.; Vengsarkar, V.A.; Ayad, S.; Ayad, S.; Mikhael, M. Artificial intelligence in patient care: Evaluating artificial intelligence's accuracy and accessibility in addressing blepharoplasty concerns. *Int. Ophthalmol.* **2025**, *45*, 244. https://doi.org/10.1007/s10792-025-03611-5.
52. Erden, Y.; Temel, M.H.; Bağcıer, F. Artificial intelligence insights into osteoporosis: Assessing ChatGPT's information quality and readability. *Arch. Osteoporos.* **2024**, *19*, 17. https://doi.org/10.1007/s11657-024-01376-5.
53. Shin, D.; Park, H.; Shaffrey, I.; Yacoubian, V.; Taka, T.M.; Dye, J.; Danisa, O. Artificial intelligence versus clinical judgement: How accurately do generative models reflect CNS guidelines for chiari malformation? *Clin. Neurol. Neurosurg.* **2025**, *248*, 108662. https://doi.org/10.1016/j.clineuro.2024.108662.
54. Andrikyan, W.; Sametinger, S.M.; Kosfeld, F.; Jung-Poppe, L.; Fromm, M.F.; Maas, R.; Nicolaus, H.F. Artificial intelligence-powered chatbots in search engines: A cross-sectional study on the quality and risks of drug information for patients. *BMJ Qual. Saf.* **2025**, *34*, 100–109. https://doi.org/10.1136/bmjqs-2024-017476.
55. De Rouck, R.; Wille, E.; Gilbert, A.; Vermeersch, N. Assessing artificial intelligence-generated patient discharge information for the emergency department: A pilot study. *Int. J. Emerg. Med.* **2025**, *18*, 85. https://doi.org/10.1186/s12245-025-00885-5.
56. Mondal, H.; Gupta, G.; Sarangi, P.K.; Sharma, S.; Choudhary, P.K.; Juhi, A.; Kumari, A.; Mondal, S. Assessing the Capability of Large Language Model Chatbots in Generating Plain Language Summaries. *Cureus* **2025**, *17*, e80976. https://doi.org/10.7759/cureus.80976.
57. Xu, Q.; Wang, J.; Chen, X.; Wang, J.; Li, H.; Wang, Z.; Li, W.; Gao, J.; Chen, C.; Gao, Y. Assessing the Efficacy of ChatGPT Prompting Strategies in Enhancing Thyroid Cancer Patient Education: A Prospective Study. *J. Med. Syst.* **2025**, *49*, 11. https://doi.org/10.1007/s10916-024-02129-0.

58. Scaff, S.P.S.; Reis, F.J.J.; Ferreira, G.E.; Jacob, M.F.; Saragiotto, B.T. Assessing the performance of AI chatbots in answering patients' common questions about low back pain. *Ann. Rheum. Dis.* **2025**, *84*, 143–149. https://doi.org/10.1136/ard-2024-226202.
59. Dharia, S.N.; Traversone, J.; Wortman, R.; Mulligan, M. Assessing the quality and readability of ChatGPT responses to frequently asked questions about trigger finger release. *J. Plast. Reconstr. Aesthet. Surg.* **2025**, *105*, 170–172. https://doi.org/10.1016/j.bjps.2025.04.033.
60. Stephenson-Moe, C.A.; Behers, B.J.; Gibons, R.M.; Behers, B.M.; Jesus Herrera, L.; Anneaud, D.; Rosario, M.A.; Wojtas, C.N.; Bhambrah, S.; Hamad, K.M. Assessing the quality and readability of patient education materials on chemotherapy cardiotoxicity from artificial intelligence chatbots: An observational cross-sectional study. *Medicine* **2025**, *104*, e42135. https://doi.org/10.1097/MD.0000000000042135.
61. Grilo, A.; Marques, C.; Corte-Real, M.; Carolino, E.; Caetano, M. Assessing the Quality and Reliability of ChatGPT's Responses to Radiotherapy-Related Patient Queries: Comparative Study with GPT-3.5 and GPT-4. *JMIR Cancer* **2025**, *11*, e63677. https://doi.org/10.2196/63677.
62. Gezer, M.C.; Armangil, M. Assessing the quality of ChatGPT's responses to commonly asked questions about trigger finger treatment. *Turk. J. Trauma Emerg. Surg. Ulus. Travma Acil Cerrahi Derg.* **2025**, *31*, 389–393. https://doi.org/10.14744/tjtes.2025.32735.
63. Keating, M.; Bollard, S.M.; Potter, S. Assessing the Quality, Readability, and Acceptability of AI-Generated Information in Plastic and Aesthetic Surgery. *Cureus* **2024**, *16*, e73874. https://doi.org/10.7759/cureus.73874.
64. Ozduran, E.; Hancı, V.; Erkin, Y.; Özbek, İ.C.; Abdulkerimov, V. Assessing the readability, quality and reliability of responses produced by ChatGPT, Gemini, and Perplexity regarding most frequently asked keywords about low back pain. *PeerJ* **2025**, *13*, e18847. https://doi.org/10.7717/peerj.18847.
65. Ömür Arça, D.; Erdemir, İ.; Kara, F.; Shermatov, N.; Odacioğlu, M.; İbişoğlu, E.; Hanci, F.B.; Sağiroğlu, G.; Hanci, V. Assessing the readability, reliability, and quality of artificial intelligence chatbot responses to the 100 most searched queries about cardiopulmonary resuscitation: An observational study. *Medicine* **2024**, *103*, e38352. https://doi.org/10.1097/MD.0000000000038352.
66. Olszewski, R.; Watros, K.; Mańczak, M.; Owoc, J.; Jeziorski, K.; Brzeziński, J. Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: A comparative study. *Int. J. Med. Inform.* **2024**, *190*, 105562. https://doi.org/10.1016/j.ijmedinf.2024.105562.
67. Saeedi, S.; Bakhtiar, M. Assessing the response quality and readability of ChatGPT in stuttering. *J. Fluen. Disord.* **2025**, *85*, 106149. https://doi.org/10.1016/j.jfludis.2025.106149.
68. Khabaz, K.; Newman-Hung, N.J.; Kallini, J.R.; Kendal, J.; Christ, A.B.; Bernthal, N.M.; Wessel, L.E. Assessment of Artificial Intelligence Chatbot Responses to Common Patient Questions on Bone Sarcoma. *J. Surg. Oncol.* **2025**, *131*, 719–724. https://doi.org/10.1002/jso.27966.
69. Pan, A.; Musheyev, D.; Bockelman, D.; Loeb, S.; Kabarriti, A.E. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. *JAMA Oncol.* **2023**, *9*, 1437–1440. https://doi.org/10.1001/jamaoncol.2023.2947.
70. Topdağı, B.; Kavaz, T. Assessment of information quality in contemporary artificial intelligence systems for digital smile design: A comparative analysis. *J. Prosthet. Dent.* **2025**, *134*, 1279.E1–1279.E8. https://doi.org/10.1016/j.prosdent.2025.06.030.
71. Hancı, V.; Ergün, B.; Gül, Ş.; Uzun, Ö.; Erdemir, İ.; Hancı, F.B. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. *Medicine* **2024**, *103*, e39305. https://doi.org/10.1097/MD.0000000000039305.
72. Cao, H.; Hao, C.; Zhang, T.; Zheng, X.; Gao, Z.; Wu, J.; Gan, L.; Liu, Y.; Zeng, X.; Wang, W. Battle of the artificial intelligence: A comprehensive comparative analysis of DeepSeek and ChatGPT for urinary incontinence-related questions. *Front. Public Health* **2025**, *13*, 1605908. https://doi.org/10.3389/fpubh.2025.1605908.
73. Özer Aslan, İ.; Aslan, M.T. Benchmarking AI Chatbots for Maternal Lactation Support: A Cross-Platform Evaluation of Quality, Readability, and Clinical Accuracy. *Healthcare* **2025**, *13*, 1756. https://doi.org/10.3390/healthcare13141756.
74. Rouhi, A.D.; Ghanem, Y.K.; Yolchieva, L.; Saleh, Z.; Joshi, H.; Moccia, M.C.; Suarez-Pierre, A.; Han, J.J. Can Artificial Intelligence Improve the Readability of Patient Education Materials on Aortic Stenosis? A Pilot Study. *Cardiol. Ther.* **2024**, *13*, 137–147. https://doi.org/10.1007/s40119-023-00347-0.
75. Dursun, D.; Bilici Geçer, R. Can artificial intelligence models serve as patient information consultants in orthodontics? *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 211. https://doi.org/10.1186/s12911-024-02619-8.
76. Lack, B.T.; Mouhawasse, E.; Childers, J.T.; Jackson, G.R.; Daji, S.V.; Yerke-Hansen, P.; Familiari, F.; Knapik, D.M.; Sabesan, V.J. Can ChatGPT answer patient questions regarding reverse shoulder arthroplasty? *J. ISAKOS* **2024**, *9*, 100323. https://doi.org/10.1016/j.jisako.2024.100323.

77. Hones, K.; Krisanda, E.; Chim, H. Caution Regarding ChatGPT's Appropriateness and Reliability Regarding Surgery for Wrist Arthritis. *Hand* **2025**, *20*, 910–916. https://doi.org/10.1177/15589447241265519.
78. Dias, R.; Castan, A.; Gotoff, K.; Kadkoy, Y.; Ippolito, J.; Beebe, K.; Benevenia, J. ChatGPT 3.5 Better Improves Comprehensibility of English, than Spanish, Generated Responses to Osteosarcoma Questions. *J. Surg. Oncol.* **2025**, *131*, 1692–1695. https://doi.org/10.1002/jso.28109.
79. Nian, P.P.; Umesh, A.; Jones, R.H.; Adhiyaman, A.; Williams, C.J.; Goodbody, C.M.; Heyer, J.H.; Doyle, S.M. ChatGPT and Google Gemini are Clinically Inadequate in Providing Recommendations on Management of Developmental Dysplasia of the Hip Compared to American Academy of Orthopaedic Surgeons Clinical Practice Guidelines. *J. Pediatr. Orthop. Soc. N. Am.* **2024**, *10*, 100135. https://doi.org/10.1016/j.jposna.2024.100135.
80. Siu, A.H.Y.; Gibson, D.P.; Chiu, C.; Kwok, A.; Irwin, M.; Christie, A.; Koh, C.E.; Keshava, A.; Reece, M.; Suen, M.; et al. ChatGPT as a patient education tool in colorectal cancer-An in-depth assessment of efficacy, quality and readability. *Color. Dis.* **2025**, *27*, e17267. https://doi.org/10.1111/codi.17267.
81. Deng, J.; Li, L.; Oosterhof, J.J.; Malliaras, P.; Silbernagel, K.G.; Breda, S.J.; Eygendaal, D.; Oei, E.H.; de Vos, R.J. ChatGPT is a comprehensive education tool for patients with patellar tendinopathy, but it currently lacks accuracy and readability. *Musculoskelet. Sci. Pract.* **2025**, *76*, 103275. https://doi.org/10.1016/j.msksp.2025.103275.
82. Mathes, S.; Seurig, S.; Bluhme, F.; Beyer, K.; Heizmann, F.; Wagner, M.; Neugärtner, I.; Biedermann, T.; Darsow, U. ChatGPT Performance on 120 Interdisciplinary Allergology Questions-Systematic Evaluation with Clinical Error Impact Assessment for Critical Erroneous AI-Guided Chatbot Advice. *J. Allergy Clin. Immunol. Pract.* **2025**, *13*, 1350–1357.e4. https://doi.org/10.1016/j.jaip.2025.03.030.
83. AlShehri, Y.; McConkey, M.; Lodhia, P. ChatGPT Provides Satisfactory but Occasionally Inaccurate Answers to Common Patient Hip Arthroscopy Questions. *Arthroscopy* **2025**, *41*, 1337–1347. https://doi.org/10.1016/j.arthro.2024.06.017.
84. Ho, R.A.; Shaari, A.L.; Cowan, P.T.; Yan, K. ChatGPT Responses to Frequently Asked Questions on Ménière's Disease: A Comparison to Clinical Practice Guideline Answers. *OTO Open* **2024**, *8*, e163. https://doi.org/10.1002/oto2.163.
85. Shen, S.A.; Perez-Heydrich, C.A.; Xie, D.X.; Nellis, J.C. ChatGPT vs. web search for patient questions: What does ChatGPT do better? *Eur. Arch. Otorhinolaryngol.* **2024**, *281*, 3219–3225. https://doi.org/10.1007/s00405-024-08524-0.
86. Sikander, B.; Baker, J.J.; Deveci, C.D.; Lund, L.; Rosenberg, J. ChatGPT-4 and Human Researchers Are Equal in Writing Scientific Introduction Sections: A Blinded, Randomized, Non-inferiority Controlled Study. *Cureus* **2023**, *15*, e49019. https://doi.org/10.7759/cureus.49019.
87. Browne, R.; Gull, K.; Hurley, C.M.; Sugrue, R.M.; O'Sullivan, J.B. ChatGPT-4 Can Help Hand Surgeons Communicate Better with Patients. *J. Hand Surg. Glob. Online* **2024**, *6*, 436–438. https://doi.org/10.1016/j.jhsg.2024.03.008.
88. Akyol Onder, E.N.; Ensari, E.; Ertan, P. ChatGPT-4o's performance on pediatric Vesicoureteral reflux. *J. Pediatr. Urol.* **2025**, *21*, 504–509. https://doi.org/10.1016/j.jpurol.2024.12.002.
89. Najafali, D.; Galbraith, L.G.; Camacho, J.M.; Stoffel, V.; Herzog, I.; Moss, C.; Taiberg, S.L.; Knoedler, L. Class in Session: Analysis of GPT-4-created Plastic Surgery In-service Examination Questions. *Plast. Reconstr. Surg. Glob. Open* **2024**, *12*, e6185. https://doi.org/10.1097/GOX.0000000000006185.
90. Bahçeci, T.; Elmaağaç, B.; Ceyhan, E. Comparative analysis of the effectiveness of microsoft copilot artificial intelligence chatbot and google search in answering patient inquiries about infertility: Evaluating readability, understandability, and actionability. *Int. J. Impot. Res.* **2025**, *37*, 1002–1007. https://doi.org/10.1038/s41443-025-01056-z.
91. Maron, C.M.; Emile, S.H.; Horesh, N.; Freund, M.R.; Pellino, G.; Wexner, S.D. Comparing answers of ChatGPT and Google Gemini to common questions on benign anal conditions. *Tech. Coloproctol.* **2025**, *29*, 57. https://doi.org/10.1007/s10151-024-03096-x.
92. Du, K.; Li, A.; Zuo, Q.H.; Zhang, C.Y.; Guo, R.; Chen, P.; Du, W.S.; Li, S.M. Comparing Artificial Intelligence-Generated and Clinician-Created Personalized Self-Management Guidance for Patients with Knee Osteoarthritis: Blinded Observational Study. *J. Med. Internet Res.* **2025**, *27*, e67830. https://doi.org/10.2196/67830.
93. Gondode, P.; Duggal, S.; Garg, N.; Sethupathy, S.; Asai, O.; Lohakare, P. Comparing patient education tools for chronic pain medications: Artificial intelligence chatbot versus traditional patient information leaflets. *Indian. J. Anaesth.* **2024**, *68*, 631–636. https://doi.org/10.4103/ija.ija_204_24.
94. Shanmugam, S.K.; Browning, D.J. Comparison of Large Language Models in Diagnosis and Management of Challenging Clinical Cases. *Clin. Ophthalmol.* **2024**, *18*, 3239–3247. https://doi.org/10.2147/OPTH.S488232.
95. Roy, J.M.; Atallah, E.; Piper, K.; Majmundar, S.; Mouchtouris, N.; Self, D.M.; Kaul, A.; Sizdahkhani, S.; Musmar, B.; Tjoumakaris, S.I.; et al. Comparison of quality, empathy and readability of physician responses versus chatbot responses to common

cerebrovascular neurosurgical questions on a social media platform. *Clin. Neurol. Neurosurg.* **2025**, *255*, 108986. https://doi.org/10.1016/j.clineuro.2025.108986.

96. Zaleski, A.L.; Berkowsky, R.; Craig, K.J.T.; Pescatello, L.S. Comprehensiveness, Accuracy, and Readability of Exercise Recommendations Provided by an AI-Based Chatbot: Mixed Methods Study. *JMIR Med. Educ.* **2024**, *10*, e51308. https://doi.org/10.2196/51308.
97. Singh, S.; Errampalli, E.; Errampalli, N.; Miran, M.S. Enhancing Patient Education on Cardiovascular Rehabilitation with Large Language Models. *Mo. Med.* **2025**, *122*, 67–71.
98. Abreu, A.A.; Murimwa, G.Z.; Farah, E.; Stewart, J.W.; Zhang, L.; Rodriguez, J.; Sweetenham, J.; Zeh, H.J.; Wang, S.C.; Polanco, P.M. Enhancing Readability of Online Patient-Facing Content: The Role of AI Chatbots in Improving Cancer Information Accessibility. *J. Natl. Compr. Canc. Netw.* **2024**, *22*, e237334. https://doi.org/10.6004/jnccn.2023.7334.
99. Mondal, H.; Tiu, D.N.; Mondal, S.; Dutta, R.; Naskar, A.; Podder, I. Evaluating Accuracy and Readability of Responses to Midlife Health Questions: A Comparative Analysis of Six Large Language Model Chatbots. *J. Midlife Health* **2025**, *16*, 45–50. https://doi.org/10.4103/jmh.jmh_182_24.
100. Zhan, Y.; Chen, X.; Ye, F.; Wu, Z.; Usman, M.; Yuan, Z.; Wu, H.; Huang, J.; Yu, H. Evaluating AI Chatbot Responses to Postkidney Transplant Inquiries. *Transplant. Proc.* **2025**, *57*, 394–405. https://doi.org/10.1016/j.transproceed.2024.12.028.
101. Kayra, M.V.; Anil, H.; Ozdogan, I.; Baradia, S.M.A.; Toksoz, S. Evaluating AI chatbots in penis enhancement information: A comparative analysis of readability, reliability and quality. *Int. J. Impot. Res.* **2025**, *37*, 558–563. https://doi.org/10.1038/s41443-025-01098-3.
102. Kacer, E.O. Evaluating AI-based breastfeeding chatbots: Quality, readability, and reliability analysis. *PLoS ONE* **2025**, *20*, e0319782. https://doi.org/10.1371/journal.pone.0319782.
103. Zhou, M.; Pan, Y.; Zhang, Y.; Song, X.; Zhou, Y. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. *Int. J. Med. Inform.* **2025**, *198*, 105871. https://doi.org/10.1016/j.ijmedinf.2025.105871.
104. Helvacioglu-Yigit, D.; Demirturk, H.; Ali, K.; Tamimi, D.; Koenig, L.; Almashraqi, A. Evaluating artificial intelligence chatbots for patient education in oral and maxillofacial radiology. *Oral. Surg. Oral. Med. Oral. Pathol. Oral Radiol.* **2025**, *139*, 750–759. https://doi.org/10.1016/j.oooo.2025.01.001.
105. Dincer, H.A.; Dogu, D. Evaluating Artificial Intelligence in Patient Education: DeepSeek-V3 Versus ChatGPT-4o in Answering Common Questions on Laparoscopic Cholecystectomy. *ANZ J. Surg.* **2025**, *95*, 2322–2328. https://doi.org/10.1111/ans.70198.
106. Sina, E.M.; Campbell, D.J.; Duffy, A.; Mandloi, S.; Benedict, P.; Farquhar, D.; Unsal, A.; Nyquist, G. Evaluating ChatGPT as a Patient Education Tool for COVID-19-Induced Olfactory Dysfunction. *OTO Open* **2024**, *8*, e70011. https://doi.org/10.1002/oto2.70011.
107. Lee, T.J.; Campbell, D.J.; Rao, A.K.; Hossain, A.; Elkattawy, O.; Radfar, N.; Lee, P.; Gardin, J.M. Evaluating ChatGPT Responses on Atrial Fibrillation for Patient Education. *Cureus* **2024**, *16*, e61680. https://doi.org/10.7759/cureus.61680.
108. Campbell, D.J.; Estephan, L.E.; Mastrolonardo, E.V.; Amin, D.R.; Huntley, C.T.; Boon, M.S. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. *J. Clin. Sleep Med.* **2023**, *19*, 1989–1995. https://doi.org/10.5664/jcsm.10728.
109. Pandey, V.K.; Munshi, A.; Mohanti, B.K.; Bansal, K.; Rastogi, K. Evaluating ChatGPT to test its robustness as an interactive information database of radiation oncology and to assess its responses to common queries from radiotherapy patients: A single institution investigation. *Cancer Radiother.* **2024**, *28*, 258–264. https://doi.org/10.1016/j.canrad.2023.11.005.
110. Sahin, S.; Erkmen, B.; Duymaz, Y.K.; Bayram, F.; Tekin, A.M.; Topsakal, V. Evaluating ChatGPT-4's performance as a digital health advisor for otosclerosis surgery. *Front. Surg.* **2024**, *11*, 1373843. https://doi.org/10.3389/fsurg.2024.1373843.
111. Alapati, R.; Campbell, D.; Molin, N.; Creighton, E.; Wei, Z.; Boon, M.; Huntley, C. Evaluating insomnia queries from an artificial intelligence chatbot for patient education. *J. Clin. Sleep Med.* **2024**, *20*, 583–594. https://doi.org/10.5664/jcsm.10948.
112. Fazilat, A.Z.; Brenac, C.; Kawamoto-Duran, D.; Berry, C.E.; Alyono, J.; Chang, M.T.; Liu, D.T.; Patel, Z.M.; Tringali, S.; Wan, D.C.; et al. Evaluating the quality and readability of ChatGPT-generated patient-facing medical information in rhinology. *Eur. Arch. Otorhinolaryngol.* **2025**, *282*, 1911–1920. https://doi.org/10.1007/s00405-024-09180-0.
113. Giammanco, P.A.; Collins, C.E.; Zimmerman, J.; Kricfalusi, M.; Rice, R.C.; Trumbo, M.; Carlson, B.A.; Rajfer, R.A.; Schneiderman, B.A.; Elsissy, J.G. Evaluating the Quality and Readability of Information Provided by Generative Artificial Intelligence Chatbots on Clavicle Fracture Treatment Options. *Cureus* **2025**, *17*, e77200. https://doi.org/10.7759/cureus.77200.
114. Singavarapu, J.; Khemlani, A.; Jacobs, M.; Berglas, E.; Lazar, J.; Kabarriti, A. Evaluating the Quality of Cardiovascular Disease Information from AI Chatbots: A Comparative Study. *Cureus* **2025**, *17*, e88085. https://doi.org/10.7759/cureus.88085.

115. Kara, M.; Ozduran, E.; Kara, M.M.; Özbek, İ.C.; Hancı, V. Evaluating the readability, quality, and reliability of responses generated by ChatGPT, Gemini, and Perplexity on the most commonly asked questions about Ankylosing spondylitis. *PLoS ONE* **2025**, *20*, e0326351. https://doi.org/10.1371/journal.pone.0326351.
116. Karaagac, M.; Carkit, S. Evaluation of AI-Based Chatbots in Liver Cancer Information Dissemination: A Comparative Analysis of GPT, DeepSeek, Copilot, and Gemini. *Oncology* **2025**, 1–10. https://doi.org/10.1159/000546726.
117. Spina, A.; Andalib, S.; Flores, D.; Vermani, R.; Halaseh, F.F.; Nelson, A.M. Evaluation of Generative Language Models in Personalizing Medical Information: Instrument Validation Study. *JMIR AI* **2024**, *3*, e54371. https://doi.org/10.2196/54371.
118. Şahin, M.F.; Keleş, A.; Özcan, R.; Doğan, Ç.; Topkaç, E.C.; Akgül, M.; Yazıcı, C.M. Evaluation of information accuracy and clarity: ChatGPT responses to the most frequently asked questions about premature ejaculation. *Sex. Med.* **2024**, *12*, qfae036. https://doi.org/10.1093/sexmed/qfae036.
119. Öztürk, Z.; Bal, C.; Çelikkaya, B.N. Evaluation of Information Provided by ChatGPT Versions on Traumatic Dental Injuries for Dental Students and Professionals. *Dent. Traumatol.* **2025**, *41*, 427–436. https://doi.org/10.1111/edt.13042.
120. Casciato, D.; Mateen, S.; Cooperman, S.; Pesavento, D.; Brandao, R.A. Evaluation of Online AI-Generated Foot and Ankle Surgery Information. *J. Foot Ankle Surg.* **2024**, *63*, 680–683. https://doi.org/10.1053/j.jfas.2024.06.009.
121. Davis, R.J.; Ayo-Ajibola, O.; Lin, M.E.; Swanson, M.S.; Chambers, T.N.; Kwon, D.I.; Kokot, N.C. Evaluation of Oropharyngeal Cancer Information from Revolutionary Artificial Intelligence Chatbot. *Laryngoscope* **2024**, *134*, 2252–2257. https://doi.org/10.1002/lary.31191.
122. Meyer, M.K.R.; Kandathil, C.K.; Davis, S.J.; Durairaj, K.K.; Patel, P.N.; Pepper, J.P.; Spataro, E.A.; Most, S.P. Evaluation of Rhinoplasty Information from ChatGPT, Gemini, and Claude for Readability and Accuracy. *Aesthetic Plast. Surg.* **2025**, *49*, 1868–1873. https://doi.org/10.1007/s00266-024-04343-0.
123. Gupta, A.; Basha, A.; Sontam, T.R.; Hlavinka, W.J.; Croen, B.J.; Abdou, C.; Abdullah, M.; Hamilton, R. Evolution of patient education materials from large-language artificial intelligence models on complex regional pain syndrome: Are patients learning? *Bayl. Univ. Med. Cent. Proc.* **2025**, *38*, 221–226. https://doi.org/10.1080/08998280.2025.2470033.
124. Kılınç, D.D.; Mansız, D. Examination of the reliability and readability of Chatbot Generative Pretrained Transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. *Am. J. Orthod. Dentofacial Orthop.* **2024**, *165*, 546–555. https://doi.org/10.1016/j.ajodo.2023.11.012.
125. Canillas Del Rey, F.; Canillas Arias, M. Exploring the potential of Artificial Intelligence in Traumatology: Conversational answers to specific questions. *Rev. Esp. Cir. Ortop. Traumatol.* **2025**, *69*, 38–46. https://doi.org/10.1016/j.recot.2024.05.004. (In English, Spanish)
126. Park, K.U.; Lipsitz, S.; Dominici, L.S.; Lynce, F.; Minami, C.A.; Nakhlis, F.; Waks, A.G.; Warren, L.E.; Eidman, N.; Frazier, J.; et al. Generative artificial intelligence as a source of breast cancer information for patients: Proceed with caution. *Cancer* **2025**, *131*, e35521. https://doi.org/10.1002/cncr.35521.
127. Zaretsky, J.; Kim, J.M.; Baskharoun, S.; Zhao, Y.; Austrian, J.; Aphinyanaphongs, Y.; Gupta, R.; Blecker, S.B.; Feldman, J. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Netw. Open* **2024**, *7*, e240357. https://doi.org/10.1001/jamanetworkopen.2024.0357.
128. Lee, Y.; Shin, T.; Tessier, L.; Javidan, A.; Jung, J.; Hong, D.; Strong, A.T.; McKechnie, T.; Malone, S.; ASMBS Artificial Intelligence and Digital Surgery Task Force; et al. Harnessing artificial intelligence in bariatric surgery: Comparative analysis of ChatGPT-4, Bing, and Bard in generating clinician-level bariatric surgery recommendations. *Surg. Obes. Relat. Dis.* **2024**, *20*, 603–608. https://doi.org/10.1016/j.soard.2024.03.011.
129. Asfuroğlu, Z.M.; Yağar, H.; Gümüşoğlu, E. High accuracy but limited readability of large language model-generated responses to frequently asked questions about Kienböck's disease. *BMC Musculoskelet. Disord.* **2024**, *25*, 879. https://doi.org/10.1186/s12891-024-07983-0.
130. Gül, Ş.; Erdemir, İ.; Hanci, V.; Aydoğmuş, E.; Erkoç, Y.S. How artificial intelligence can provide information about subdural hematoma: Assessment of readability, reliability, and quality of ChatGPT, BARD, and perplexity responses. *Medicine* **2024**, *103*, e38009. https://doi.org/10.1097/MD.0000000000038009.
131. Ulusoy, I.; Yılmaz, M.; Kıvrak, A. How Efficient Is ChatGPT in Accessing Accurate and Quality Health-Related Information? *Cureus* **2023**, *15*, e46662. https://doi.org/10.7759/cureus.46662.
132. Akkan, H.; Seyyar, G.K. Improving readability in AI-generated medical information on fragility fractures: The role of prompt wording on ChatGPT's responses. *Osteoporos. Int.* **2025**, *36*, 403–410. https://doi.org/10.1007/s00198-024-07358-0.

133. Tan, C.W.; Chan, J.C.Y.; Chan, J.J.I.; Nagarajan, S.; Sng, B.L. Information about labor epidural analgesia: An updated evaluation on the readability, accuracy, and quality of ChatGPT responses incorporating patient preferences and complex clinical scenarios. *Int. J. Obstet. Anesth.* **2025**, *63*, 104688. https://doi.org/10.1016/j.ijoa.2025.104688.
134. Xie, Y.; Seth, I.; Hunter-Smith, D.J.; Rozen, W.M.; Seifman, M.A. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: A comprehensive analysis. *ANZ J. Surg.* **2024**, *94*, 68–77. https://doi.org/10.1111/ans.18666.
135. Cao, J.J.; Kwon, D.H.; Ghaziani, T.T.; Kwo, P.; Tse, G.; Kesselman, A.; Kamaya, A.; Tse, J.R. Large language models' responses to liver cancer surveillance, diagnosis, and management questions: Accuracy, reliability, readability. *Abdom. Radiol.* **2024**, *49*, 4286–4294. https://doi.org/10.1007/s00261-024-04501-7.
136. Singh, S.P.; Jamal, A.; Qureshi, F.; Zaidi, R.; Qureshi, F. Leveraging Generative Artificial Intelligence Models in Patient Education on Inferior Vena Cava Filters. *Clin. Pract.* **2024**, *14*, 1507–1514. https://doi.org/10.3390/clinpract14040121.
137. Andreadis, K.; Newman, D.R.; Twan, C.; Shunk, A.; Mann, D.M.; Stevens, E.R. Mixed methods assessment of the influence of demographics on medical advice of ChatGPT. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 2002–2009. https://doi.org/10.1093/jamia/ocae086.
138. Shukla, I.Y.; Sun, M.Z. Online and ChatGPT-generated patient education materials regarding brain tumor prognosis fail to meet readability standards. *J. Clin. Neurosci.* **2025**, *138*, 111410. https://doi.org/10.1016/j.jocn.2025.111410.
139. Hunter, N.; Allen, D.; Xiao, D.; Cox, M.; Jain, K. Patient education resources for oral mucositis: A google search and ChatGPT analysis. *Eur. Arch. Otorhinolaryngol.* **2025**, *282*, 1609–1618. https://doi.org/10.1007/s00405-024-08913-5.
140. Yalla, G.R.; Hyman, N.; Hock, L.E.; Zhang, Q.; Shukla, A.G.; Kolomeyer, N.N. Performance of Artificial Intelligence Chatbots on Glaucoma Questions Adapted from Patient Brochures. *Cureus* **2024**, *16*, e56766. https://doi.org/10.7759/cureus.56766.
141. Alasker, A.; Alsalamah, S.; Alshathri, N.; Almansour, N.; Alsalamah, F.; Alghafees, M.; AlKhamees, M.; Alsaikhan, B. Performance of large language models (LLMs) in providing prostate cancer information. *BMC Urol.* **2024**, *24*, 177. https://doi.org/10.1186/s12894-024-01570-0.
142. Chen, D.; Parsa, R.; Hope, A.; Hannon, B.; Mak, E.; Eng, L.; Liu, F.F.; Fallah-Rad, N.; Heesters, A.M.; Raman, S. Physician and Artificial Intelligence Chatbot Responses to Cancer Questions from Social Media. *JAMA Oncol.* **2024**, *10*, 956–960. https://doi.org/10.1001/jamaoncol.2024.0836.
143. Zhang, J.; Sun, Y.; Rong, Y.; Li, H.; Jiang, B.; Zhao, C.; Liu, H. Potential of AI Chatbots in Online Hair Transplantation Consultations: A Multi-metric Assessment of Three Models. *Aesthet. Plast. Surg.* **2025**, *49*, 6155–6161. https://doi.org/10.1007/s00266-025-05103-4.
144. Bragazzi, N.L.; Buchinger, M.; Atwan, H.; Tuma, R.; Chirico, F.; Szarpak, L.; Farah, R.; Khamisy-Farah, R. Proficiency, Clarity, and Objectivity of Large Language Models Versus Specialists' Knowledge on COVID-19's Impacts in Pregnancy: Cross-Sectional Pilot Study. *JMIR Form. Res.* **2025**, *9*, e56126. https://doi.org/10.2196/56126.
145. Warren, C.J.; Edmonds, V.S.; Payne, N.G.; Voletti, S.; Wu, S.Y.; Colquitt, J.; Sadeghi-Nejad, H.; Punjani, N. Prompt matters: Evaluation of large language model chatbot responses related to Peyronie's disease. *Sex. Med.* **2024**, *12*, qfae055. https://doi.org/10.1093/sexmed/qfae055.
146. Warren, C.J.; Payne, N.G.; Edmonds, V.S.; Voleti, S.S.; Choudry, M.M.; Punjani, N.; Abdul-Muhsin, H.M.; Humphreys, M.R. Quality of Chatbot Information Related to Benign Prostatic Hyperplasia. *Prostate* **2025**, *85*, 175–180. https://doi.org/10.1002/pros.24814.
147. Stapleton, P.; Santucci, J.; Cundy, T.P.; Sathianathen, N. Quality of Information on Wilms Tumor from Artificial Intelligence Chatbots: What Are Your Patients and Their Families Reading? *Urology* **2025**, *198*, 130–134. https://doi.org/10.1016/j.urology.2025.01.054.
148. Boscolo-Rizzo, P.; Marcuzzo, A.V.; Lazzarin, C.; Giudici, F.; Polesel, J.; Stellin, M.; Pettorelli, A.; Spinato, G.; Ottaviano, G.; Ferrari, M.; et al. Quality of Information Provided by Artificial Intelligence Chatbots Surrounding the Reconstructive Surgery for Head and Neck Cancer: A Comparative Analysis Between ChatGPT4 and Claude2. *Clin. Otolaryngol.* **2025**, *50*, 330–335. https://doi.org/10.1111/coa.14261.
149. Aydın, F.O.; Aksoy, B.K.; Ceylan, A.; Akbaş, Y.B.; Ermiş, S.; Kepez Yıldız, B.; Yıldırım, Y. Readability and Appropriateness of Responses Generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in Refractive Surgery. *Turk. J. Ophthalmol.* **2024**, *54*, 313–317. https://doi.org/10.4274/tjo.galenos.2024.28234.
150. Musheyev, D.; Pan, A.; Gross, P.; Kamyab, D.; Kaplinsky, P.; Spivak, M.; Bragg, M.A.; Loeb, S.; Kabarriti, A.E. Readability and Information Quality in Cancer Information from a Free vs Paid Chatbot. *JAMA Netw. Open* **2024**, *7*, e2422275. https://doi.org/10.1001/jamanetworkopen.2024.22275.

151. Alsabawi, Y.; Quesada, P.R.; Rouse, D.T. Readability of custom chatbot vs. GPT-4 responses to otolaryngology-related patient questions. *Am. J. Otolaryngol.* **2025**, *46*, 104717. https://doi.org/10.1016/j.amjoto.2025.104717.
152. Gawey, L.; Dagenet, C.B.; Tran, K.A.; Park, S.; Hsiao, J.L.; Shi, V. Readability of Information Generated by ChatGPT for Hidradenitis Suppurativa. *JMIR Dermatol.* **2024**, *7*, e55204. https://doi.org/10.2196/55204.
153. Büker, M.; Mercan, G. Readability, accuracy and appropriateness and quality of AI chatbot responses as a patient information source on root canal retreatment: A comparative assessment. *Int. J. Med. Inform.* **2025**, *201*, 105948. https://doi.org/10.1016/j.ijmedinf.2025.105948.
154. Ozduran, E.; Akkoc, I.; Büyükçoban, S.; Erkin, Y.; Hanci, V. Readability, reliability and quality of responses generated by ChatGPT, gemini, and perplexity for the most frequently asked questions about pain. *Medicine* **2025**, *104*, e41780. https://doi.org/10.1097/MD.0000000000041780.
155. Alamleh, S.; Mavedatnia, D.; Francis, G.; Le, T.; Davies, J.; Lin, V.; Lee, J.J.W. Readability, Reliability, and Quality Analysis of Internet-Based Patient Education Materials and Large Language Models on Meniere's Disease. *J. Otolaryngol. Head Neck Surg.* **2025**, *54*, 19160216251360651. https://doi.org/10.1177/19160216251360651.
156. Şan, H.; Bayrakcı, Ö.; Çağdaş, B.; Serdengeçti, M.; Alagöz, E. Reliability and readability analysis of ChatGPT-4 and Google Bard as a patient information source for the most commonly applied radionuclide treatments in cancer patients. *Rev. Esp. Med. Nucl. Imagen. Mol. Engl. Ed.* **2024**, *43*, 500021. https://doi.org/10.1016/j.remnie.2024.500021.
157. Aydinbelge-Dizdar, N.; Dizdar, K. Evaluación de la fiabilidad y legibilidad de las respuestas de los chatbots como recurso de información al paciente para las exploraciones PET-TC más communes. *Rev. Esp. Med. Nucl. Imagen. Mol. Engl. Ed.* **2025**, *44*, 500065. https://doi.org/10.1016/j.remnie.2024.500065.
158. Şahin, M.F.; Ateş, H.; Keleş, A.; Özcan, R.; Doğan, Ç.; Akgül, M.; Yazıcı, C.M. Responses of Five Different Artificial Intelligence Chatbots to the Top Searched Queries About Erectile Dysfunction: A Comparative Analysis. *J. Med. Syst.* **2024**, *48*, 38. https://doi.org/10.1007/s10916-024-02056-0.
159. Yassa, A.; Ayad, O.; Cohen, D.A.; Patel, A.M.; Vengsarkar, V.A.; Hegazin, M.S.; Filimonov, A.; Hsueh, W.D.; Eloy, J.A. Search for medical information for chronic rhinosinusitis through an artificial intelligence ChatBot. *Laryngoscope Investig. Otolaryngol.* **2024**, *9*, e70009. https://doi.org/10.1002/lio2.70009.
160. Shin, D.; Tang, T.; Carson, J.; Isaac, R.; Dinh, C.; Im, D.; Fay, A.; Isaac, A.; Cho, S.; Brandt, Z.; et al. Subthalamic nucleus or globus pallidus internus deep brain stimulation for the treatment of parkinson's disease: An artificial intelligence approach. *J. Clin. Neurosci.* **2025**, *138*, 111393. https://doi.org/10.1016/j.jocn.2025.111393.
161. Anıl, H.; Kayra, M.V. The digital dialogue on premature ejaculation: Evaluating the efficacy of artificial intelligence-driven responses. *Int. Urol. Nephrol.* **2025**, *57*, 2829–2836. https://doi.org/10.1007/s11255-025-04461-x.
162. Liu, X.; Shi, S.; Zhang, X.; Gao, Q.; Wang, W. The role of ChatGPT-4o in differential diagnosis and management of vertigo-related disorders. *Sci. Rep.* **2025**, *15*, 18688. https://doi.org/10.1038/s41598-025-96309-8.
163. Taka, T.M.; Collins, C.E.; Miner, A.; Overfield, I.; Shin, D.; Seo, L.; Danisa, O. The role of generative artificial intelligence in deciding fusion treatment of lumbar degeneration: A comparative analysis and narrative review. *Eur. Spine J.* **2025**, *34*, 3901–3910. https://doi.org/10.1007/s00586-025-09052-z.
164. Arzu, U.; Gencer, B. To Self-Treat or Not to Self-Treat: Evaluating the Diagnostic, Advisory and Referral Effectiveness of ChatGPT Responses to the Most Common Musculoskeletal Disorders. *Diagnostics* **2025**, *15*, 1834. https://doi.org/10.3390/diagnostics15141834.
165. Ayo-Ajibola, O.; Davis, R.J.; Lin, M.E.; Vukkadala, N.; O'Dell, K.; Swanson, M.S.; Johns, M.M., 3rd; Shuman, E.A. TrachGPT: Appraisal of tracheostomy care recommendations from an artificial intelligent Chatbot. *Laryngoscope Investig. Otolaryngol.* **2024**, *9*, e1300. https://doi.org/10.1002/lio2.1300.
166. Kerkütlüoğlu, M.; Kaya, E.; Gökmen, R. Trustworthiness, Value, Danger, and Readability of ChatGPT-Generated Responses to Health Questions Related to Pulmonary Arterial Hypertension. *Cureus* **2024**, *16*, e71472. https://doi.org/10.7759/cureus.71472.
167. Lee, T.J.; Campbell, D.J.; Patel, S.; Hossain, A.; Radfar, N.; Siddiqui, E.; Gardin, J.M. Unlocking Health Literacy: The Ultimate Guide to Hypertension Education from ChatGPT Versus Google Gemini. *Cureus* **2024**, *16*, e59898. https://doi.org/10.7759/cureus.59898.
168. Covington, E.W.; Watts Alexander, C.S.; Sewell, J.; Hutchison, A.M.; Kay, J.; Tocco, L.; Hyte, M. Unlocking the future of patient Education: ChatGPT vs. LexiComp® as sources of patient education materials. *J. Am. Pharm. Assoc.* **2025**, *65*, 102119. https://doi.org/10.1016/j.japh.2024.102119.

169. Steimetz, E.; Minkowitz, J.; Gabutan, E.C.; Ngichabe, J.; Attia, H.; Hershkop, M.; Ozay, F.; Hanna, M.G.; Gupta, R. Use of Artificial Intelligence Chatbots in Interpretation of Pathology Reports. *JAMA Netw. Open* **2024**, *7*, e2412767. https://doi.org/10.1001/jamanetworkopen.2024.12767.
170. Patel, T.A.; Michaelson, G.; Morton, Z.; Harris, A.; Smith, B.; Bourguillon, R.; Wu, E.; Eguia, A.; Maxwell, J.H. Use of ChatGPT for patient education involving HPV-associated oropharyngeal cancer. *Am. J. Otolaryngol.* **2025**, *46*, 104642. https://doi.org/10.1016/j.amjoto.2025.104642.
171. Burns, C.; Bakaj, A.; Berishaj, A.; Hristidis, V.; Deak, P.; Equils, O. Use of Generative AI for Improving Health Literacy in Reproductive Health: Case Study. *JMIR Form. Res.* **2024**, *8*, e59434. https://doi.org/10.2196/59434.
172. ELSenbawy, O.M.; Patel, K.B.; Wannakuwatte, R.A.; Thota, A.N. Use of generative large language models for patient education on common surgical conditions: A comparative analysis between ChatGPT and Google Gemini. *Updates Surg.* **2025**, 1–7. https://doi.org/10.1007/s13304-025-02074-8.
173. Šuto Pavičić, J.; Marušić, A.; Buljan, I. Using ChatGPT to Improve the Presentation of Plain Language Summaries of Cochrane Systematic Reviews About Oncology Interventions: Cross-Sectional Study. *JMIR Cancer* **2025**, *11*, e63347. https://doi.org/10.2196/63347.
174. Tran, Q.L.; Huynh, P.P.; Le, B.; Jiang, N. Utilization of Artificial Intelligence in the Creation of Patient Information on Laryngology Topics. *Laryngoscope* **2025**, *135*, 1295–1300. https://doi.org/10.1002/lary.31891.
175. Sönmezoğlu, H.İ.; Güner Sönmezoğlu, B.; Temel, M.H.; Çakir, B. Comprehensibility and readability of selected artificial intelligence chatbots in providing uveitis-related information. *Medicine* **2025**, *104*, e45135. https://doi.org/10.1097/MD.0000000000045135.
176. Baur, D.; Ansorg, J.; Heyde, C.E.; Voelker, A. Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study. *JMIR AI* **2025**, *4*, e75262. https://doi.org/10.2196/75262.
177. Prabha, S.; Gomez-Cabello, C.A.; Haider, S.A.; Genovese, A.; Trabilsy, M.; Wood, N.G.; Bagaria, S.; Tao, C.; Forte, A.J. Enhancing Clinical Decision Support with Adaptive Iterative Self-Query Retrieval for Retrieval-Augmented Large Language Models. *Bioengineering* **2025**, *12*, 895. https://doi.org/10.3390/bioengineering12080895.
178. Cross, J.L.; Choma, M.A.; Onofrey, J.A. Bias in medical AI: Implications for clinical decision-making. *PLoS Digit. Health* **2024**, *3*, e0000651. https://doi.org/10.1371/journal.pdig.0000651.
179. Alli, S.R.; Hossain, S.Q.; Das, S.; Upshur, R. The Potential of Artificial Intelligence Tools for Reducing Uncertainty in Medicine and Directions for Medical Education. *JMIR Med. Educ.* **2024**, *10*, e51446. https://doi.org/10.2196/51446.
180. Gomez-Cabello, C.A.; Prabha, S.; Haider, S.A.; Genovese, A.; Collaco, B.G.; Wood, N.G.; Bagaria, S.; Forte, A.J. Comparative Evaluation of Advanced Chunking for Retrieval-Augmented Generation in Large Language Models for Clinical Decision Support. *Bioengineering* **2025**, *12*, 1194. https://doi.org/10.3390/bioengineering12111194.
181. Abo El-Enen, M.; Saad, S.; Nazmy, T. A survey on retrieval-augmentation generation (RAG) models for healthcare applications. *Neural Comput. Appl.* **2025**, *37*, 28191–28267. https://doi.org/10.1007/s00521-025-11666-9.
182. Wada, A.; Tanaka, Y.; Nishizawa, M.; Yamamoto, A.; Akashi, T.; Hagiwara, A.; Hayakawa, Y.; Kikuta, J.; Shimoji, K.; Sano, K.; et al. Retrieval-augmented generation elevates local LLM quality in radiology contrast media consultation. *npj Digit. Med.* **2025**, *8*, 395. https://doi.org/10.1038/s41746-025-01802-z.
183. Maity, S.; Saikia, M.J. Large Language Models in Healthcare and Medical Applications: A Review. *Bioengineering* **2025**, *12*, 631. https://doi.org/10.3390/bioengineering12060631.
184. Weiss, B.D. *Health Literacy and Patient Safety: Help Patients Understand. Manual for Clinicians*, 2nd ed.; American Medical Association Foundation and American Medical Association: Chicago, IL, USA, 2007.
185. US Department of Health and Human Services; Office of Disease Prevention and Health Promotion. *National Action Plan to Improve Health Literacy*; US Department of Health and Human Services :Washington, DC, USA, 2010.
186. DeTemple, D.E.; Meine, T.C. Comparison of the readability of ChatGPT and Bard in medical communication: A meta-analysis. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 325. https://doi.org/10.1186/s12911-025-03035-2.
187. Moons, P.; Van Bulck, L. Using ChatGPT and Google Bard to improve the readability of written patient information: A proof of concept. *Eur. J. Cardiovasc. Nurs.* **2024**, *23*, 122–126. https://doi.org/10.1093/eurjcn/zvad087.
188. Andrew, A. Accuracy of ChatGPT in answering cardiology board-style questions. *J. Educ. Eval. Health Prof.* **2025**, *22*, 9. https://doi.org/10.3352/jeehp.2025.22.9.

189. Uchmanowicz, I.; Jędrzejczyk, M.; Vellone, E.; Janczak, S.; Mirkowski, K.; Uchmanowicz, B.M.; Czapla, M. ChatGPT in cardiovascular medicine: Revolution, hype, or helper? *Front. Public Health* **2025**, *13*, 1622561. https://doi.org/10.3389/fpubh.2025.1622561.
190. Harskamp, R.E.; De Clercq, L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: A proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). *Acta Cardiol.* **2024**, *79*, 358–366. https://doi.org/10.1080/00015385.2024.2303528.
191. Lautrup, A.D.; Hyrup, T.; Schneider-Kamp, A.; Dahl, M.; Lindholt, J.S.; Schneider-Kamp, P. Heart-to-heart with ChatGPT: The impact of patients consulting AI for cardiovascular health advice. *Open Heart* **2023**, *10*, e002455. https://doi.org/10.1136/openhrt-2023-002455.
192. Meyer, A.; Riese, J.; Streichert, T. Comparison of the Performance of GPT-3.5 and GPT-4 with That of Medical Students on the Written German Medical Licensing Examination: Observational Study. *JMIR Med. Educ.* **2024**, *10*, e50965. https://doi.org/10.2196/50965.
193. Lahat, A.; Sharif, K.; Zoabi, N.; Shneor Patt, Y.; Sharif, Y.; Fisher, L.; Shani, U.; Arow, M.; Levin, R.; Klang, E. Assessing Generative Pretrained Transformers (GPT) in Clinical Decision-Making: Comparative Analysis of GPT-3.5 and GPT-4. *J. Med. Internet Res.* **2024**, *26*, e54571. https://doi.org/10.2196/54571.
194. Bolliger, L.S.; Haller, P.; Cretton, I.C.R.; Reich, D.R.; Kew, T.; Jäger, L.A. EMTeC: A corpus of eye movements on machine-generated texts. *Behav. Res. Methods* **2025**, *57*, 189. https://doi.org/10.3758/s13428-025-02677-4.
195. James, A.; Trovati, M.; Bolton, S. Retrieval-Augmented Generation to Generate Knowledge Assets and Creation of Action Drivers. *Appl. Sci.* **2025**, *15*, 6247. https://doi.org/10.3390/app15116247.
196. Nastoska, A.; Jancheska, B.; Rizinski, M.; Trajanov, D. Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries. *Electronics* **2025**, *14*, 2717. https://doi.org/10.3390/electronics14132717.
197. Novelo, R.; Silva, R.R.; Bernardino, J. A Literature Review of Personalized Large Language Models for Email Generation and Automation. *Future Internet* **2025**, *17*, 536. https://doi.org/10.3390/fi17120536.
198. Di Martino, F.; Delmastro, F. Explainable AI for clinical and remote health applications: A survey on tabular and time series data. *Artif. Intell. Rev.* **2023**, *56*, 5261–5315. https://doi.org/10.1007/s10462-022-10304-3.
199. Wagner, N.; Kraus, M.; Minker, W.; Griol, D.; Callejas, Z. A Survey on Multi-User Conversational Interfaces. *Appl. Sci.* **2025**, *15*, 7267. https://doi.org/10.3390/app15137267.
200. Lai, X.; Lai, Y.; Chen, J.; Huang, S.; Gao, Q.; Huang, C. Evaluation Strategies for Large Language Model-Based Models in Exercise and Health Coaching: Scoping Review. *J. Med. Internet Res.* **2025**, *27*, e79217. https://doi.org/10.2196/79217.
201. Lv, X.; Zhang, X.; Li, Y.; Ding, X.; Lai, H.; Shi, J. Leveraging Large Language Models for Improved Patient Access and Self-Management: Assessor-Blinded Comparison Between Expert- and AI-Generated Content. *J. Med. Internet Res.* **2024**, *26*, e55847. https://doi.org/10.2196/55847.
202. Singh, S.U.; Namin, A.S. A survey on chatbots and large language models: Testing and evaluation techniques. *Nat. Lang. Process. J.* **2025**, *10*, 100128. https://doi.org/10.1016/j.nlp.2025.100128.
203. Dahlgren Lindström, A.; Methnani, L.; Krause, L.; Ericson, P.; de Rituerto de Troya, Í.M.; Coelho Mollo, D.; Dobbe, R. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. *Ethics Inf. Technol.* **2025**, *27*, 28. https://doi.org/10.1007/s10676-025-09837-2.
204. Shao, Y.; Yang, X.; Chen, Q.; Guo, H.; Duan, X.; Xu, X.; Yue, J.; Zhang, Z.; Zhao, S.; Zhang, S. Determinants of digital health literacy among older adult patients with chronic diseases: A qualitative study. *Front. Public Health* **2025**, *13*, 1568043. https://doi.org/10.3389/fpubh.2025.1568043.
205. Zolfaghari, Z.; Karimian, Z.; Zarifsanaiey, N.; Farahmandi, A.Y. Navigating challenges in medical english learning: Leveraging technology and gamification for interactive education—A qualitative study. *BMC Med. Educ.* **2025**, *25*, 1045. https://doi.org/10.1186/s12909-025-07511-1.
206. Khojasteh, L.; Kafipour, R.; Pakdel, F.; Mukundan, J. Empowering medical students with AI writing co-pilots: Design and validation of AI self-assessment toolkit. *BMC Med. Educ.* **2025**, *25*, 159. https://doi.org/10.1186/s12909-025-06753-3.
207. Ahmed, A.; Leroy, G.; Kauchak, D.; Barai, P.; Harber, P.; Rains, S. Parallel Corpus Analysis of Text and Audio Comprehension to Evaluate Readability Formula Effectiveness: Quantitative Analysis. *J. Med. Internet Res.* **2025**, *27*, e69772. https://doi.org/10.2196/69772.
208. Joseph, S.; Bhardwaj, A.; Skariah, J.; Aggarwal, I.; Shah, V.; Harris, R.A. Effects of education level on natural language processing in cardiovascular health communication. *Front. Public Health* **2025**, *13*, 1688173. https://doi.org/10.3389/fpubh.2025.1688173.

209. Gao, Y.; Xu, Q.; Zhang, O.; Wang, H.; Wang, Y.; Wang, J.; Chen, X. Large language models: Unlocking new potential in patient education for thyroid eye disease. *Endocrine* **2025**, *90*, 689–698. https://doi.org/10.1007/s12020-025-04339-z.
210. Zhang, Z.; Zhang, H.; Pan, Z.; Bi, Z.; Wan, Y.; Song, X.; Fan, X. Evaluating Large Language Models in Ophthalmology: Systematic Review. *J. Med. Internet Res.* **2025**, *27*, e76947. https://doi.org/10.2196/76947.
211. Zhang, J.; Song, X.; Tian, B.; Tian, M.; Zhang, Z.; Wang, J.; Fan, T. Large language models in the management of chronic ocular diseases: A scoping review. *Front. Cell Dev. Biol.* **2025**, *13*, 1608988. https://doi.org/10.3389/fcell.2025.1608988.
212. Betzler, B.K.; Chen, H.; Cheng, C.Y.; Lee, C.S.; Ning, G.; Song, S.J.; Lee, A.Y.; Kawasaki, R.; van Wijngaarden, P.; Grzybowski, A.; et al. Large language models and their impact in ophthalmology. *Lancet Digit. Health* **2023**, *5*, e917–e924. https://doi.org/10.1016/S2589-7500(23)00201-7.
213. Bacco, L.; Russo, F.; Ambrosio, L.; D'Antoni, F.; Vollero, L.; Vadalà, G.; Dell'Orletta, F.; Merone, M.; Papalia, R.; Denaro, V. Natural language processing in low back pain and spine diseases: A systematic review. *Front. Surg.* **2022**, *9*, 957085. https://doi.org/10.3389/fsurg.2022.957085.
214. Shah, R.; Schwab, J.H. Large Language Models in Spine Surgery: A Promising Technology. *HSS J.* **2025**, *21*, 15563316251340696. https://doi.org/10.1177/15563316251340696.
215. Croxford, E.; Gao, Y.; First, E.; Pellegrino, N.; Schnier, M.; Caskey, J.; Oguss, M.; Wills, G.; Chen, G.; Dligach, D.; et al. Evaluating clinical AI summaries with large language models as judges. *npj Digit. Med.* **2025**, *8*, 640. https://doi.org/10.1038/s41746-025-02005-2.
216. Alshammari, A.F.; Madfa, A.A.; Anazi, B.A.; Alenezi, Y.E.; Alkurdi, K.A. Comparison of accuracy and consistency of AI Language models when answering standardised dental MCQs. *BMC Med. Educ.* **2025**, *25*, 1507. https://doi.org/10.1186/s12909-025-07624-7.
217. Martos, M.; Fields, B.; Finlayson, S.G.; Hartell, N.; Kim, T.; Larimer, E.; Lau, J.J.; Lin, Y.H.; Salaguinto, T.; Tran, N.; et al. Accuracy of Artificial Intelligence vs Professionally Translated Discharge Instructions. *JAMA Netw. Open* **2025**, *8*, e2532312. https://doi.org/10.1001/jamanetworkopen.2025.32312.
218. Lee, C.; Britto, S.; Diwan, K. Evaluating the Impact of Artificial Intelligence (AI) on Clinical Documentation Efficiency and Accuracy Across Clinical Settings: A Scoping Review. *Cureus* **2024**, *16*, e73994. https://doi.org/10.7759/cureus.73994.