Review

Automated Approaches of Text Simplification of Patient Education Materials: Scoping Review

Cornelia Krenn[1], BSc, MSc, Dr.scient.med.; Christine Loder[1], Mag., MPH; Natalie Berger[1], BSc, MSc; Klaus Jeitler[1,2], Dr med; Thomas Semlitsch[1], Mag.rer.nat.; Andrea Siebenhofer[1,3], Dr med, MBA; Denise Wilfling[1,4], BSc, MSc, Dr.rer.hum.biol.

[1]Institute of General Practice and Evidence-based Health Services Research, Medical University of Graz, Graz, Austria

[2]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria

[3]Institute of General Practice, Goethe University Frankfurt am Main, Frankfurt, Germany

[4]Institute of Nursing and Health Sciences, Medical University of Graz, Graz, Austria

Corresponding Author:
Cornelia Krenn, BSc, MSc, Dr.scient.med.
Institute of General Practice and Evidence-based Health Services Research, Medical University of Graz
Neue Stiftingtalstraße 6
Graz, 8010
Austria
Phone: 43 316385 ext 73567
Email: cornelia.krenn@medunigraz.at

Abstract

Background: Patient education materials (PEMs) often exceed the American Medical Association's (AMA) recommended sixth-grade reading grade level (RGL). While artificial intelligence (AI) offers potential for automated text simplification, concerns persist regarding linguistic quality, content fidelity, and the understandability of simplified PEMs by laypeople.

Objective: This scoping review maps existing evidence on automated language processing technologies for simplifying PEMs for laypeople.

Methods: Following the Joanna Briggs Institute (JBI) methodology and the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline, 5 bibliographic databases (Ovid MEDLINE, Embase, CINAHL, PsycInfo, and IEEE Xplore) were systematically searched from 2019 to May 2025, supplemented by reference screening and gray literature searches. Eligible sources were peer-reviewed empirical studies published in English that examined large language models (LLMs), AI-supported writing assistants, AI-based conversational agents, or AI-supported tools designed for automatic text simplification of PEMs. Targeted outcomes included linguistic quality (ie, linguistic comprehensibility, linguistic correctness) and content fidelity (ie, factual accuracy, factual completeness) of simplified PEMs. Excluded sources comprised rule-based systems, manual text simplification, non-laypeople target groups, and technology-focused performance metrics. Results were synthesized via thematic analysis across the domains of targeted outcomes. In accordance with JBI methodology, a risk-of-bias assessment was not performed.

Results: A total of 31 eligible studies met the inclusion criteria, examining various LLMs, including OpenAI's GPT series, Gemini, Bard, Claude, Copilot, and Llama. Specifically, GPT-4.0 achieved the most consistent improvements in standardized readability metrics (eg, the Flesch-Kincaid Grade Level [FKGL]). However, achieving predefined target RGLs remained challenging across all LLMs, particularly at lower RGLs. Findings on content fidelity were inconsistent: despite high content similarity scores, content accuracy was often compromised.

Conclusions: This is the first scoping review to comprehensively synthesize evidence on automated technologies for text simplification in PEMs. The review identified 2 critical validation gaps. First, no study examined the linguistic correctness (eg, grammar and typographical errors) of automatically simplified PEMs. Second, and most notably, the understandability of the simplified PEMs was assessed exclusively by experts, with no empirical evaluation involving laypeople. Although LLMs effectively reduce text complexity as measured by objective readability metrics, reliance on these formulas represents a critical limitation, as they serve merely as structural proxies. Improvements in readability do not guarantee the maintenance of content accuracy or laypeople's understandability. Current evidence is further limited by the lack of systematic prompt quality evaluation and the predominant focus on English-language PEMs in US contexts, restricting generalizability. This review provides a foundation for

future research by highlighting the need for validated evaluation frameworks that encompass layperson testing and content verification. For clinical practice, LLMs should currently serve as assistive tools, with mandatory expert review remaining essential to verify content fidelity before disseminating LLM-simplified PEMs to laypeople.

(J Med Internet Res 2026;28:e88365) doi: 10.2196/88365

KEYWORDS

artificial intelligence; large language models; patient education materials; automatic text simplification; linguistic quality; content fidelity

Introduction

Rationale

Effective health communication is recognized as a public health priority [1]. Health communication aims to improve health by ensuring effective understanding and application of health information. Central to this is health literacy, defined as the degree to which individuals have the capacity to obtain and understand health information needed to make appropriate health decisions [2]. Patients with higher health literacy are more likely to engage in health-promoting behaviors, utilize health care services, and effectively manage chronic diseases [3]. However, the use of complex medical language poses a significant barrier to patient understanding. Health care professionals are encouraged to use plain language tailored to patients' comprehension levels, which can be challenging, especially in time-sensitive clinical care settings [4,5]. Patient education materials (PEMs) play a central role in supporting patient-physician interactions by providing clear and accessible information on health conditions, treatments, and health promotion [6]. Personalized PEMs have been shown to improve patient care through shared decision-making, enhanced patient satisfaction, and better physical and psychosocial well-being [7]. To maximize accessibility of PEMs, leading organizations such as the National Institutes of Health and the American Medical Association (AMA) recommend writing PEMs at or below a sixth-grade reading level (RGL) [8,9]. However, numerous studies have demonstrated that a significant portion of existing PEMs fail to meet this benchmark, often being written at a level too complex for many patients to understand. An analysis of PEMs showed average readability scores ranging from 8th to 15th RGL across various medical fields [10-16]. This readability concern has not improved between 2001 and 2022. These findings underscore that simplified versions of PEMs are necessary, and health care professionals are encouraged to provide easy-to-read PEMs to patients [17,18].

Artificial intelligence (AI) offers promising opportunities to enhance effective health communication. In particular, large language models (LLMs) have emerged as transformative tools in natural language processing (NLP), enabling diverse applications such as answering patient questions; summarizing, translating, or simplifying medical texts; supporting clinical paperwork; and providing individualized medical guidance [19-21]. Importantly, LLMs have the potential to enhance the accessibility of medical knowledge by making complex medical language more comprehensible to laypeople, thereby enabling patients to better understand their health conditions [22,23]. Text simplification as an NLP task has advanced significantly in recent years, especially driven by developments in LLMs since 2019. Whereas earlier technologies relied primarily on rule-based systems or machine learning models, LLMs have enabled a paradigm shift in NLP, making these capabilities more widely accessible to health care professionals and health researchers. In addition, LLM capabilities have rapidly evolved, demonstrating increasingly sophisticated language understanding and generation abilities [24,25].

Despite promising results, important challenges remain. These include the risk of factual errors (hallucinations), critical omissions of information, and the unintended loss of meaning between original and simplified text versions. Furthermore—and most critically—the difficulty of verifying understandability persists, as strong performance on standard quality metrics does not guarantee that simplified texts are actually understandable to laypeople [26]. Moreover, the rapid evolution of AI language processing technologies makes it challenging for researchers and health care providers to maintain an up-to-date overview of available tools and supporting evidence.

Objectives

This scoping review aimed to identify and map existing evidence on the use of automated language processing technologies for the simplification of PEMs into layperson-friendly language. For this review, layperson-friendly language is defined as text characterized by an accessible reading level, simple sentence structures, and explanation or avoidance of medical jargon. This scoping review is guided by the following research questions:

- What automated language processing technologies are currently used to simplify PEMs into layperson-friendly language?
- How are the linguistic quality and content fidelity of these simplified texts evaluated?

Methods

Review Principles and Protocol

This scoping review followed the methodological framework of the Joanna Briggs Institute (JBI) [27] and is reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [28] (Multimedia Appendix 1) and the PRISMA-S (PRISMA Statement for Reporting Literature Searches in Systematic Reviews) [29] (Multimedia Appendix 2) guidelines. The review methodology was registered in the Open Science Framework [30].

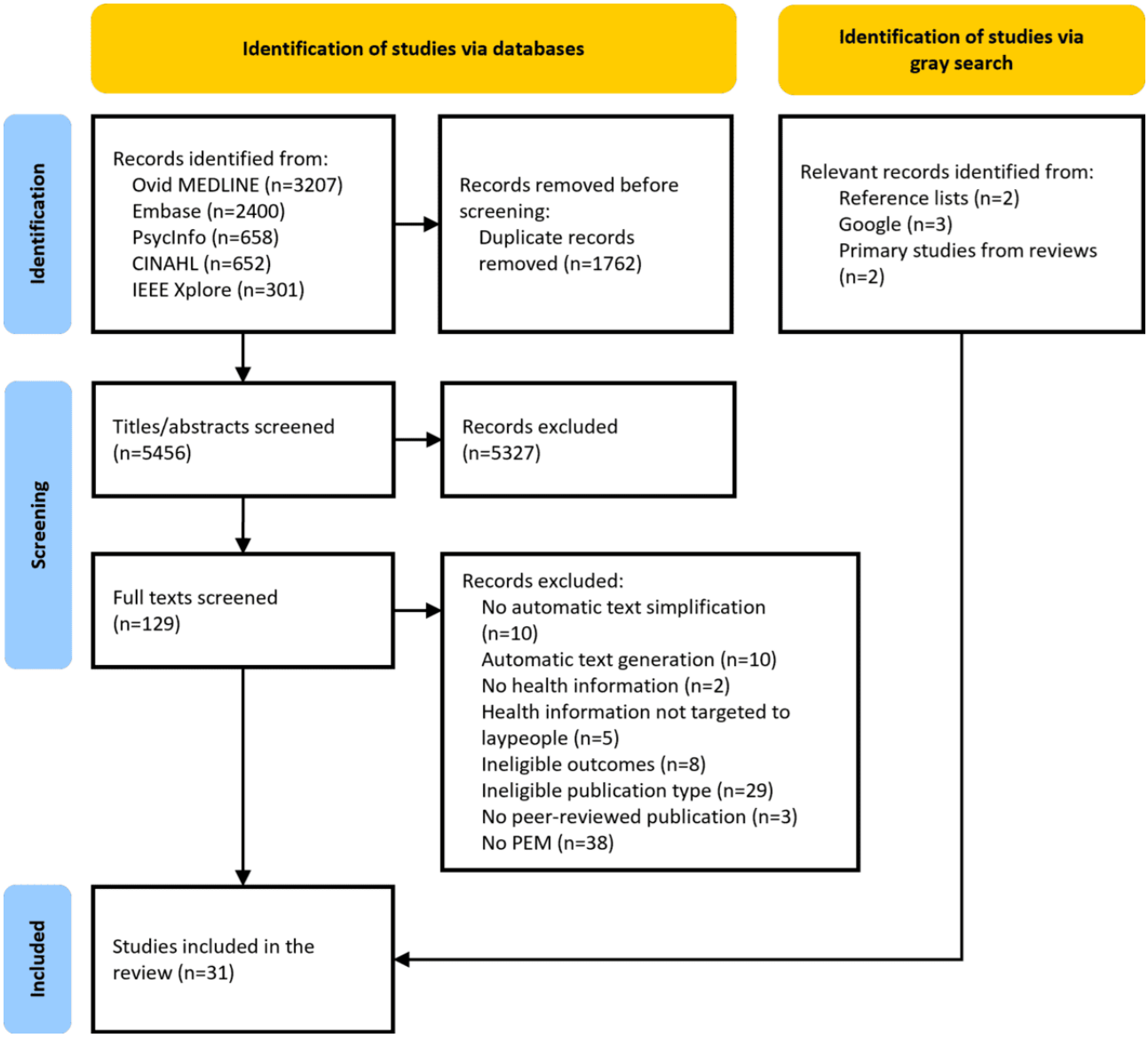
Deviation From the Protocol

The preregistered protocol specified the inclusion of studies on automatic text simplification of any complex medical or health-related text. As Figure 1 shows, full-text screening revealed a large and highly heterogeneous body of literature, encompassing different document types such as radiology reports, electronic health records, discharge letters, PEMs, informed consent forms, and scientific papers. This heterogeneity presents challenges for meaningful synthesis and comparison across fundamentally different health information materials with varying purposes, audiences, and complexity levels.

Additionally, this study was framed as part of a larger project, the A+CHIS project [31], which aims to develop a system that provides users with a selection of diverse health documents and preliminary PEMs tailored to their individual information needs and cognitive prerequisites. To present a more focused and coherent synthesis of evidence that would yield insights for patient-facing health communication systems such as A+CHIS, a post hoc decision was made to refine the scope to studies concerning PEM simplification specifically. This narrowed focus aligns with the primary goal of improving health information accessibility for laypeople and ensures methodological consistency in evaluating simplification approaches.

This post hoc decision was made after the initial full-text screening but before full-text data extraction. All other methods remained unchanged from the preregistered protocol.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection. PEM: patient education material.



Screening and Identifying Relevant Studies: Information Sources and Search

A comprehensive literature search was conducted in May 2025 across 5 bibliographic databases: MEDLINE, Embase, CINAHL, PsycINFO, and IEEE Xplore. The MEDLINE and Embase strategies were run simultaneously as a multifile search in Ovid, and results were deduplicated using the Ovid deduplication tool. The search strategy was based on database-specific controlled vocabulary and free-text terms. To identify relevant search

terms, the MeSH Browser, a keyword analysis tool [32], and a synonym identification tool [33] were used. The search was limited to the period from 2019 to 2025, as this period marks significant advances in automated language processing technologies, particularly with the emergence of LLMs that have substantially enhanced automatic text simplification capabilities. No search filter was applied. The search block for "automated language processing technologies" was based on 2 existing search strategies [34,35]. Detailed search strategies for the databases are provided in Multimedia Appendix 3. The MEDLINE search strategy was peer reviewed by another research team member (KJ).

Reference lists of included studies were manually screened to identify additional studies. We also searched gray literature in July 2025 using the search string "patient education material" AND ("large language model" OR "ChatGPT") AND "simplifying" in Google. The first 5 pages of each search result were screened. We did not contact the authors of included studies for clarification, as the available data were sufficient to conduct the scoping review. No other methods were used to locate additional studies. Eligible studies were imported into EndNote 21.5 (Clarivate Plc) to identify and remove duplicates.

Eligibility Criteria

Framework Applied

The Population-Concept-Context (PCC) framework was used to define the eligibility criteria and guide the search strategy [27].

- Population:
  The target population was defined as laypeople requiring PEMs, including patients and their families, the general public, individuals with limited health literacy, and health care professionals without specific medical expertise.
- Concept:
  The central concept was the use of automated language processing technologies for the simplification of PEMs.
- Context:
  The context was limited to PEM simplification, defined as any text (full text or excerpt) designed to inform laypeople about medical or health-related topics. Nontext formats, such as video, audio materials, or infographics, were not considered in this review.

Building upon the PCC framework, the following inclusion and exclusion criteria were applied for study selection:

Inclusion Criteria

- Studies that assessed automated language processing technologies, specifically LLMs such as AI-supported writing assistants (eg, DeepL Write) and AI-based conversational agents (eg, ChatGPT, Gemini), or other AI-supported tools designed for automatic text simplification.
- Studies that assessed text quality indicators of simplified PEMs in at least one of two domains: (1) linguistic quality, encompassing linguistic comprehensibility (eg, readability, text complexity, word choice, structural clarity, medical jargon) and linguistic correctness (eg, grammatical accuracy, typographical errors); and (2) content fidelity, encompassing factual correctness (eg, hallucinations, exaggerations, falsifications, understatements, misinterpretations) and factual completeness (eg, content retention).
- Studies that measured outcomes either objectively (eg, FKGL) or subjectively (eg, expert ratings, user feedback).
- Peer-reviewed empirical study designs (quantitative, qualitative, and case studies) published in English, including primary studies identified through relevant reviews.

Exclusion Criteria

- Studies that evaluated nonlearning or rule-based systems, manual text simplification or human postediting of AI output, automated language translation without simplification intent, or text analysis software tools.
- Studies that evaluated the effectiveness of automated simplified PEMs targeted at health care professionals.
- Studies that evaluated only technology-oriented performance metrics (eg, Bilingual Evaluation Understudy [BLEU] or Recall-Oriented Understudy for Gisting Evaluation [ROUGE] scores).
- Non–peer-reviewed publications (eg, preprints, editorials, commentaries, letters to the editor, opinions), studies lacking a complete methodological description, and publications in languages other than English.

Selection of Sources of Evidence

Following JBI guidelines [27], 2 authors (CK, DW, CL, or NB) independently screened all titles and abstracts against the inclusion criteria. A pilot stage involving 25 titles and abstracts was conducted to ensure consistent application of the criteria, with conflicts resolved through discussion. Full texts of potentially eligible studies were subsequently assessed by 2 independent reviewers (CK, DW, CL, or NB) using the same eligibility criteria. Disagreements were resolved through discussion or by involving a third reviewer (CK, DW, CL, or NB).

Data Charting

A standardized data extraction form was developed and piloted on 3 eligible studies to assess its clarity and completeness. Following the pilot test and discussion within the research team, 1 additional item (the language of simplified PEMs) was added to the final form.

Data Items

The following data from the included studies were extracted: bibliographic details (authors, year of publication, and country), technology details (name and version of LLM and prompts used), source text (medical field of PEMs, number of analyzed materials, and language), outcomes (text quality indicators and measurement methods), and key findings (main results related to linguistic quality and content fidelity). As recommended by the JBI methodology for scoping reviews [27], 1 reviewer (CK) extracted the data, which was verified by a second reviewer (DW).

Critical Appraisal

In line with JBI guidelines [27], critical appraisal of eligible studies was not required.

Collating, Summarizing, and Reporting the Results

Following the JBI scoping review methodology [27], we conducted a descriptive synthesis of the extracted data. We did not perform an analytical synthesis of outcomes but instead mapped and summarized findings descriptively. The extracted data were collated iteratively and organized using a framework-based approach, categorizing findings into the 2 predefined outcome domains: linguistic quality (encompassing linguistic comprehensibility and linguistic correctness) and content fidelity (encompassing factual correctness and factual completeness). Within each domain, descriptive qualitative content analysis was conducted by the first author (CK) to summarize the types of automated language processing technologies evaluated, the measurement methods applied, and the reported direction of effect (eg, improvements in readability scores or identified factual inaccuracies). Results were presented descriptively through narrative summaries, tables, and figures to address the review questions. Frequency counts were used where appropriate to quantify the occurrence of specific technologies or outcome measures. All descriptive analyses were conducted using Microsoft Excel.

Results

Selection of Sources of Evidence

The systematic database search identified 7218 references. Following deduplication, 5456 titles and abstracts were screened, and 129 full texts were subsequently assessed for eligibility. This process resulted in the inclusion of 24 [36-59] studies specifically addressing PEMs. In addition, 2 studies [60,61] identified from relevant reviews were included. Two further studies [62,63] were identified through reference list screening of the included studies, and a supplementary Google search yielded 3 additional relevant studies [64-66]. In total, this scoping review included 31 studies focusing on PEMs. Figure 1 shows the study selection process in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

Studies excluded at the full-text screening stage and their primary reasons for exclusion are provided in Multimedia Appendix 4.

Characteristics of Sources of Evidence

The main characteristics of all included studies are presented in Table 1. All included studies investigated LLMs exclusively, evaluating 8 different models. OpenAI's GPT series emerged as the most extensively studied models: GPT-4.0 (n=10) [36-43,62,64], GPT-3.5 (n=8) [44-49,60,61], and GPT-3.0 (n=1) [50]. Other evaluated LLMs included Google's Gemini (n=4) [51,52,65,66] and Bard (n=5) [53-56,63], Anthropic's Claude (n=2) [57,65], Microsoft's Copilot (n=1) [66], and Meta's Llama (n=1) [57], with 12 studies conducting comparative analyses across multiple models [51-59,63,65,66]. As illustrated in Multimedia Appendix 5, all included studies were published very recently, with 11 in 2025 [36-40,44,51,57,64-66], 19 in 2024 [41-43,45-48,50,52-56,58-63], and 1 in 2023 [49]. Although the literature search covered the period from 2019 onward, no relevant studies were identified before 2023, highlighting the emerging nature of this field.

Table 1. Summary of main characteristics of the included studies.

Study	Country	Study design	LLM[a]; number of PEMs[b] analyzed (language of PEMs)	Medical field in PEMs	Readability scale	Target RGL[c,d]
Spina et al [36]	United States	Not reported	GPT-4.0; 9 (English)	Glaucoma	FKGL[e] and FRE[f]	5
Reaver et al [37]	United States	Cross-sectional study	GPT-4.0; 57 (English) and 56 (Spanish)	Orthopedic surgery	FRE, Fry, LIX[g], RIX[h], SMOG[i], GPM[j] Fry, FHRI[k], and SOL[l]	6
Picton et al [38]	United States	Not reported	GPT-4.0; 340[m] (English)	Neurology and neurological surgery	FKGL and FRE	5
Li et al [44]	United States	Cross-sectional comparative study	GPT-3.5; 50 (English)	Cataract surgery	FKGL, SMOG, GFI[n], and CLI[o]	5
Dihan et al [51]	United States	Cross-sectional comparative study	GPT-3.5, GPT-4.0, and Gemini Advanced; 20 (English)	Dry eye disease	FKGL and SMOG	6
Chandra et al [39]	United States	Not reported	GPT-4.0; 30 (English)	Orthopedic surgery	SMOG	6-8
Busigo Torres et al [40]	United States	Not reported	GPT-4.0; 77 (Spanish)	Orthopedic	FHRI and SOL	5
Andalib et al [57]	United States	Not reported	GPT-3.5, GPT-4.0, Claude 2, and Llama 2; 48 (English)	Orthopedic	FKGL and FRE	5
Will et al [65]	United States	Cross-sectional study	GPT-4.0, Gemini 1.5-flash, and Claude 3.5 Sonnet; 60 (English)	Heart disease, cancer, and stroke	FKGL, FRE, SMOG, and GFI	5
Naghdi et al [66]	The Netherlands	Comparative observational study	GPT-3.5, GPT-4.0, Copilot, and Gemini; 30 (English)	Reproductive genetics	FKGL, FRE, SMOG, GFI, CLI, and LWF[p]	6-8
Singh et al [64]	United States	Not reported	GPT-4.0; 25 (English)	Neurosurgery	FKGL, FRE, SMOG, CLI, and ARI[q]	8
Zaki et al [41]	United States	Not reported	GPT-4.0; 73 (English)	Interventional radiology procedures	FRE, GFI, and ARI	5
Vallurupalli et al [45]	United States	Not reported	GPT-3.5; 18[r] (English)	Hand surgery	Combined calculator: FKGL, FRE, SMOG, ARI, GFI, LWF, and CLI	6-8
Shehab et al [42]	United States	Cross-sectional study	GPT-4.0; 124 (English)	Cleft lip and palate	FKGL and FRE	6
Patel et al [43]	United States	Proof‐of‐concept study	GPT-4.0; 71 (English)	Otolaryngology	FKGL, FRE, SMOG, and GFI	6
Oliva et al [46]	United States	Cross-sectional study	GPT-3.5; 109 (English)	Otolaryngology	FKGL and FRE	5
Kianian et al [53]	United States	Not reported	GPT-4.0 and Bard; 9 (English)	Uveitis	FKGL	6
Rasika et al [47]	United States	Not reported	GPT-3.5; 15 (English and Spanish)	Ophthalmology	FKGL, FRE, GFI, FHRI, Crawford Nivel-de-Grado, Gutiérrez, Szigriszt-Pazos/INFLESZ[s], and Legibilidad-µ	Nr
Gupta et al [52]	United States	Not reported	GPT-4.0 and Gemini; 7 (English)	Radiology	FKGL, FRE, SMOG, and GFI	6
Garcia Valencia et al [58]	United States and Thailand	Not reported	GPT-3.5 and GPT-4.0; 27[t] (English)	Living kidney donation	FKGL	8
Fanning et al [59]	United States	Not reported	GPT-3.5 and GPT-4.0; 75 (English)	Plastic surgery	FKGL, FRE, Fry, SMOG, GFI, and Raygor Estimate	6

Study	Country	Study design	LLM[a]; number of PEMs[b] analyzed (language of PEMs)	Medical field in PEMs	Readability scale	Target RGL[c,d]
Dihan et al [54]	United States	Cross-sectional comparative study	GPT-3.5, GPT-4.0, and Bard; 20 (English)	Idiopathic intracranial hypertension	FKGL and SMOG	6
Dihan et al [55]	United States	Cross-sectional comparative study	GPT-3.5, GPT-4.0, and Bard; 20 (English)	Childhood glaucoma	FKGL and SMOG	6
Dihan et al [56]	United States	Cross-sectional comparative study	GPT-3.5, GPT-4.0, and Bard; 20 (English)	Pediatric cataract	FKGL and SMOG	6
Baldwin [50]	United Kingdom	Not reported	GPT-3.0; 50 (English)	Burns first aid	FKGL, FRE, SMOG, GFI, and CLI	6
Ayre et al [48]	Australia	Observational study	GPT-3.5; 26[u] (English)	Not restricted[v]	SMOG	8
Manasyan et al [60]	United States	Not reported	GPT-3.5; 34 (English)	Alveolar bone grafting	FKGL, FRE, and GFI	5
Vallurupalli et al [61]	United States	Not reported	GPT-3.5; 18 (English)	Craniofacial procedures	Combined calculator: FKGL, FRE, SMOG, ARI, GFI, LWF, and CLI	8
Abreu et al [62]	United States	Cross-sectional study	GPT-4.0; 34 (English)	Cancer	FKGL, SMOG, Fry, and GFI	6
Rouhi et al [63]	United States	Pilot study	GPT-3.5 and Bard; 21 (English)	Aortic stenosis	FKGL, FRE, SMOG, and GFI	5
Kirchner et al [49]	United States	Proof-of-concept study	GPT-3.5; 20 (English)	Orthopedic	FKGL and FRE	5

[a]LLM: large language model.

[b]PEM: patient education material.

[c]RGL: reading grade level.

[d]According to the US grade school-level system.

[e]FKGL: Flesch-Kincaid Grade Level.

[f]FRE: Flesch Reading Ease.

[g]LIX: Läsbarhetsindex Index.

[h]RIX: Rate Index.

[i]SMOG: Simple Measure of Gobbledygook.

[j]GPM Fry: Gilliam Peña Mountain Fry Graph.

[k]FHRI: Fernandez-Huerta Readability Index.

[l]SOL: Spanish Orthographic Length.

[m]274 PEMs about neurology, 66 PEMs about neurological surgery.

[n]GFI: Gunning Fog Index.

[o]CLI: Coleman-Liau Index.

[p]LWF: Linsear Write Formula.

[q]ARI: Automated Readability Index.

[r]Excerpts of PEMs.

[s]INFLESZ: Índice Flesch-Szigriszt.

[t]Frequently asked questions.

[u]Extracts were at least 300 words.

[v]Online health information published by recognized national and international health information provider websites.

The geographical distribution showed a strong US dominance (n=27) [36-47,49,51-57,59-65], with additional contributions from the United Kingdom [50], Australia [48], the Netherlands [66], and a US-Thailand collaboration [58]. This geographic focus is reflected in the languages studied, with 30 studies examining English PEMs [36-39,41-66], 2 evaluating both English and Spanish PEMs [37,47], and 1 focusing solely on Spanish PEMs [40].

Sample sizes varied considerably, ranging from 7 to 340 PEMs per study, including full texts, text excerpts [45,48], and frequently asked question sections [58].

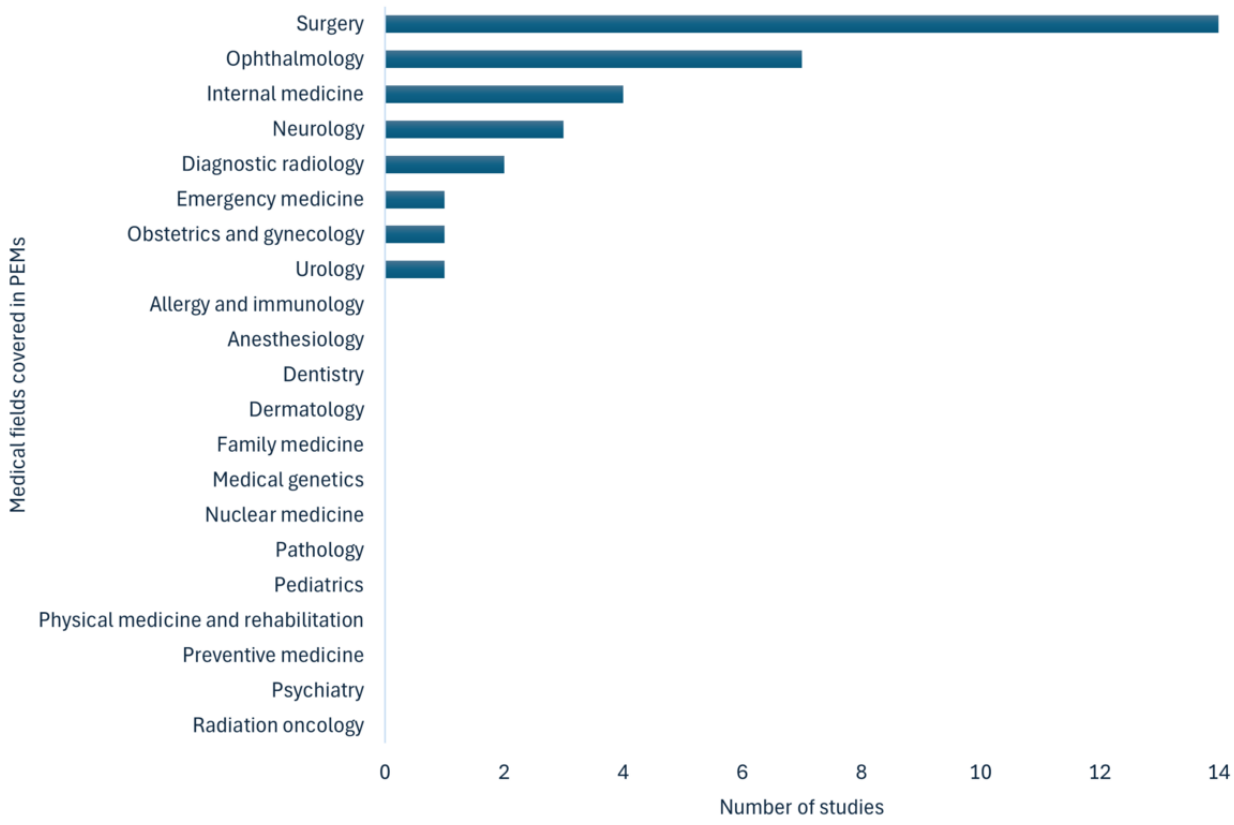
Figure 2 illustrates the distribution of medical fields addressed in the PEMs in the included studies. Surgery was the most frequently investigated field (n=14), encompassing orthopedics, neurosurgery, plastic surgery, oral and maxillofacial surgery, craniofacial surgery, and otolaryngology. Ophthalmology was the second most common field (n=7), covering topics such as

glaucoma, cataract, uveitis, and dry eye disease. Additional fields included internal medicine (n=4; heart disease, aortic stenosis, cancer), neurology (n=3; general neurology, stroke, idiopathic intracranial hypertension), and diagnostic radiology (n=2). Single studies addressed emergency medicine (burns first aid) [50], obstetrics and gynecology (reproductive genetics) [66], and urology (renal transplantation) [58]. No studies were identified in the following medical fields: allergy and immunology, anesthesiology, dermatology, dentistry, family medicine, medical genetics, nuclear medicine, pathology, pediatrics, physical medicine and rehabilitation, preventive medicine, psychiatry, or radiation oncology.

To evaluate the effect of prompt engineering, 1 study [59] compared the outputs of GPT-3.5 and GPT-4.0 using 2 different prompts. The first was a simple prompt with a general instruction to simplify the text while maintaining its structure. The second, more detailed prompt provided explicit constraints, referencing specific readability scales and including examples to guide the model's output [59]. All prompts used in the included studies are presented in Multimedia Appendix 6.

A comprehensive overview of the reported outcomes across all included studies is provided in Multimedia Appendix 7.

Figure 2. Distribution of medical fields represented in patient education materials across the included studies.



Linguistic Quality

Assessment of Linguistic Comprehensibility Versus Absence of Linguistic Correctness Evaluation

All 31 included studies assessed the linguistic quality of simplified PEMs using at least one indicator of linguistic comprehensibility. Notably, none of the studies evaluated linguistic correctness, such as grammatical accuracy or typographical errors.

Linguistic Comprehensibility

Readability

All included studies assessed linguistic comprehensibility using at least one readability scale. The most frequently used scales for English-language PEMs were the FKGL (n=23 studies) [38,42-44,46,47,49-60,62-66], Flesch Reading Ease (FRE; n=18) [36-38,41-43,46,47,49,50,52,57,59,60,63-66], and Simple Measure of Gobbledygook (SMOG; n=17) [37,39,43,44,48,50-52,54-56,59,62-66]. Additional metrics included the Gunning Fog Index (GFI; n=12) [41,43,44,47,50,52,59,60,62,63,65,66], Coleman-Liau Index (CLI; n=4) [44,50,64,66], Automated Readability Index (ARI; n=3) [41,52,64], Fry score (n=3) [37,59,62], and single-use scales such as the Raygor Estimate, Rate Index (RIX) [59], Läsbarhetsindex (LIX) [37], and Linsear Write Formula (LWF) [66]. Two studies used a composite readability scoring system based on the average of 7 readability indices (FRE, FKGL, SMOG, GFI, CLI, ARI, and LWF) [45,61]. Several studies employed multiple readability scales for comparative analysis. Table 2 provides an overview of the readability scales used to assess English-language PEMs in the included studies, along with their underlying scoring components used to measure readability.

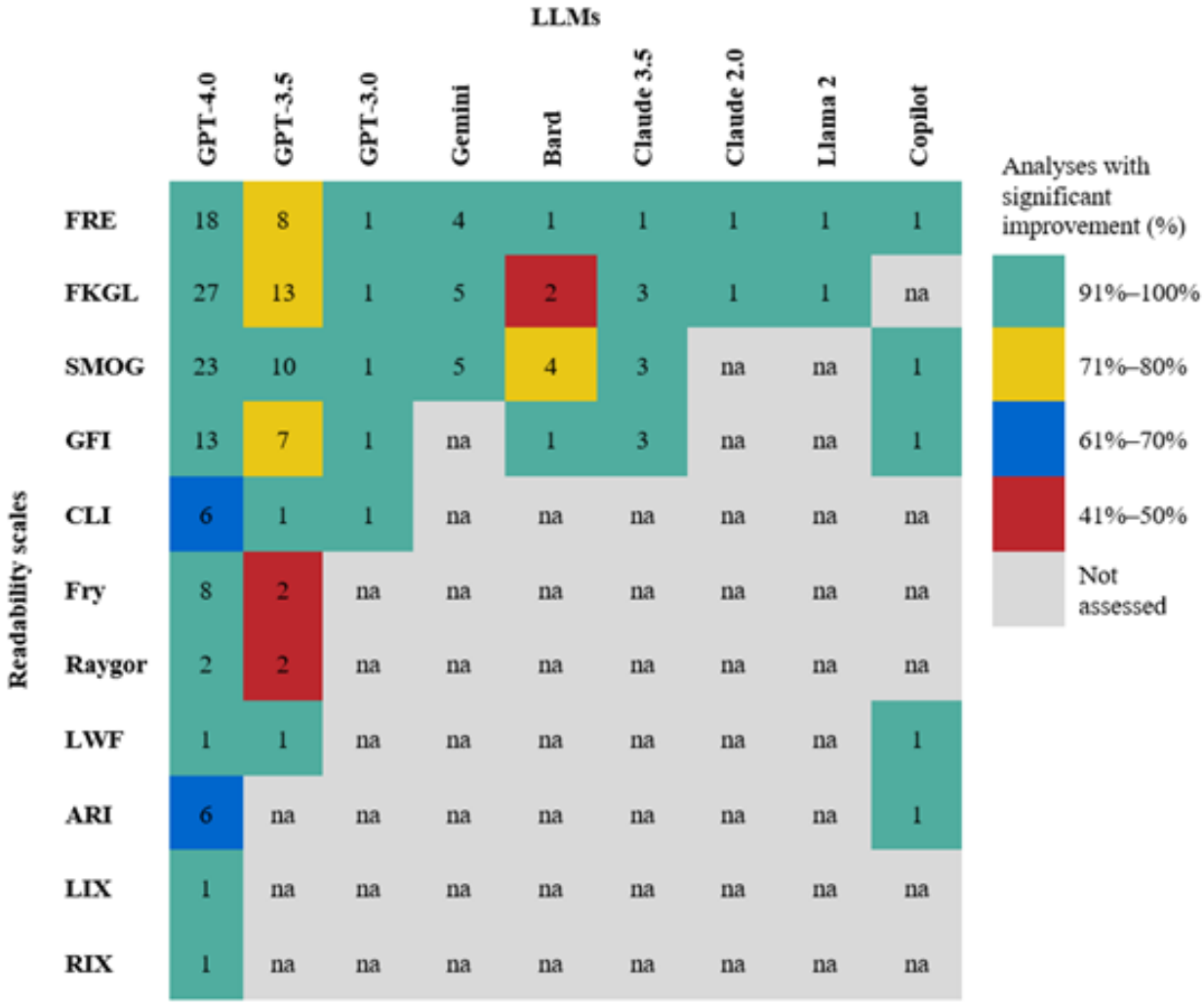
As Figure 3 shows, readability improvement varied by LLM and readability scale in English-language PEMs. GPT-4.0, the most frequently evaluated model, demonstrated the most consistent performance, with a proportion of analyses showing readability improvement falling into the "91%-100%" category across all metrics, except CLI and ARI (67%). By contrast, GPT-3.5 showed more variable results: a 100% success rate on SMOG but lower rates of improvement for FKGL (77%), FRE (75%), and GFI (71%). Other LLMs (GPT-3.0, Gemini, Bard, Claude, Llama 2, and Copilot) also achieved high proportions of analyses showing significant improvements; however, evidence remains limited to single studies for most models.

For Spanish-language PEMs, 3 studies [37,40,47] employed language-specific readability scales such as the Fernández-Huerta Readability Index, GPM Fry Graph, LIX, RIX, Spanish Orthographic Length (SOL), Crawford Nivel-de-Grado, Gutiérrez, Szigriszt-Pazos/Índice Flesch-Szigriszt (INFLESZ), and Legibilidad-µ, achieving improvements in 9 out of 12 (75%) analyses.

Table 2. Readability scales used in included studies and corresponding scoring components [67,68].

Scale	Components to score readability
Flesch-Kincaid Grade Level	• Average number of syllables per word • Average number of words per sentence
Flesch Reading Ease	• Average number of syllables • Average number of words per sentence • Average number of sentences
Simple Measure of Gobbledygook	• Average number of words with ≥3 syllables • Average number of sentences
Gunning Fog Index	• Number of sentences • Number of words • Number of words with ≥3 syllables
Coleman-Liau Index	• Average number of letters per 100 words • Average number of sentences per 100 words
Automated Readability Index	• Average number of characters per word (eg, any letters, numbers, symbols) • Average number of words per sentence
Fry	• Average number of sentences • Syllables per 100 words
Raygor Estimate	• Average number of sentences • Long words (≥6 characters) per 100 words
Rate Index	• Number of long words • Number of sentences
Läsbarhetsindex Index	• Average number of words per sentence • Percentage of words with >6 letters
Linsear Write Readability	• Number of easy words (words with ≤2 syllables) • Number of hard words (words with ≥3 syllables)

Figure 3. Map analysis of significant readability improvements in English-language patient education materials across all studies, achieved using various large language models and assessed with different readability scales. ARI: Automated Readability Index; CLI: Coleman-Liau Index; FKGL: Flesch-Kincaid Grade Level; FRE: Flesch Reading Ease; GFI: Gunning Fog Index; LIX: Läsbarhetsindex Index; LWF: Linsear Write Formula; RIX: Rate Index; SMOG: Simple Measure of Gobbledygook.



Reading Grade Levels

In addition to objective measurements of readability, nearly all included studies (n=30) investigated whether LLMs could achieve predefined target reading grade levels (RGLs) specified in the prompts, typically ranging from fifth to eighth grade. The detailed original, target, and achieved RGLs after prompting LLMs to rewrite PEMs to a specific RGL across all included studies can be found in Figure 4 (also see [38-40,42-46,48-52,54-59,61-63,66-68]).

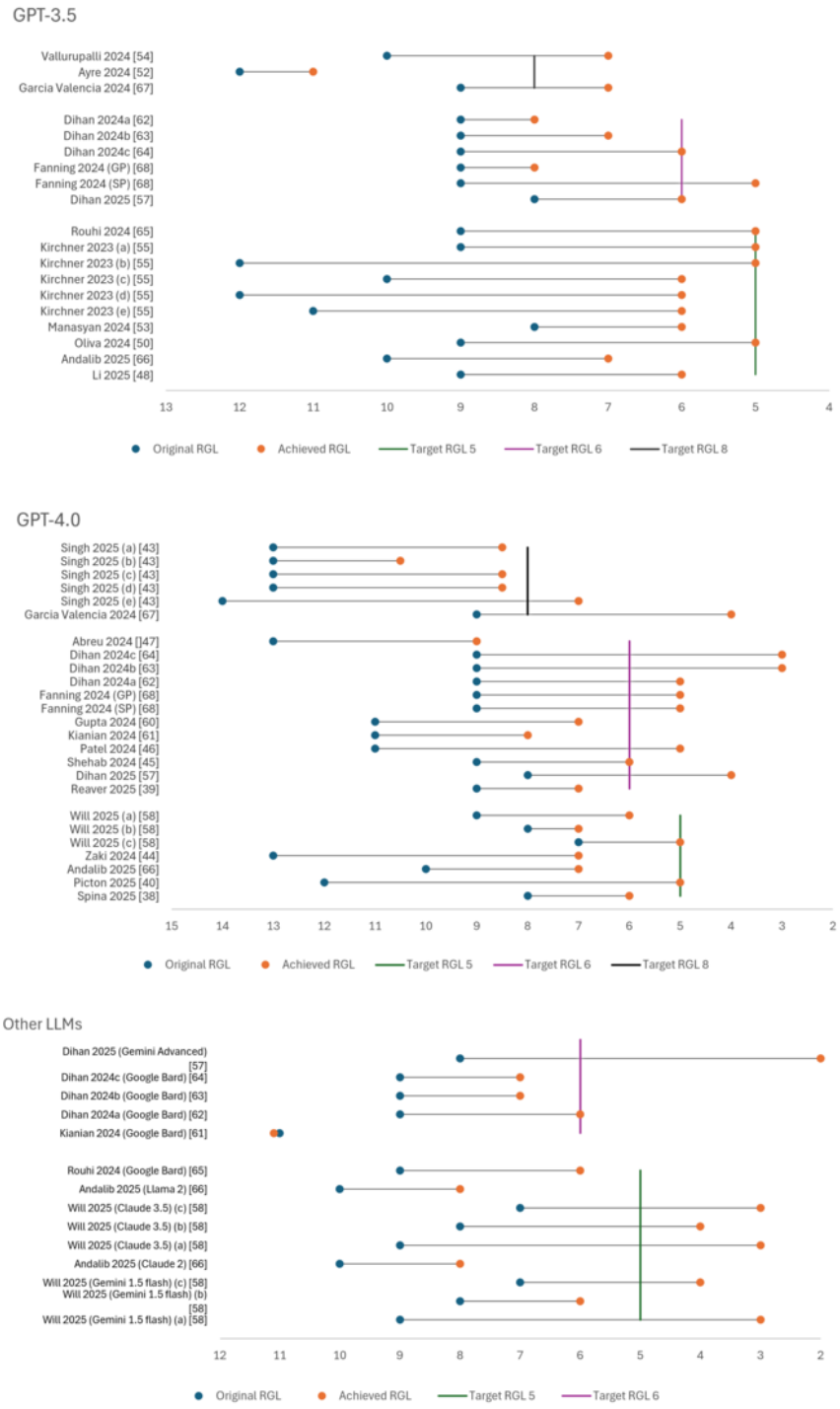
As Figure 4 shows, the studies demonstrated variable success rates for GPT models. At the fifth-grade target level, GPT-3.5 and GPT-4.0 demonstrated their lowest performance, achieving success in 4 out of 10 (40%) and 2 out of 7 (29%) attempts, respectively. Performance improved notably at the sixth-grade level, where GPT-3.5 reached the target in 3 out of 6 (50%) cases, whereas GPT-4.0 showed success in 8 out of 12 (67%) cases. At the eighth-grade level, an inverse pattern emerged: GPT-3.5 achieved a success rate of over 65% (2/3, 67%), whereas GPT-4.0 succeeded in only <35% of attempts (2/6, 33%).

Among other LLMs, Gemini demonstrated the strongest performance, with an 83% success rate, followed by Claude models at 75%. Conversely, Bard exhibited notably poor performance, achieving target RGLs in only 20% of attempts. Evidence for GPT-3.0 and Llama 2 is limited to single studies, precluding meaningful comparison.

Three studies explored flexible target ranges (sixth to eighth grade) rather than fixed targets, with only 2 of 6 analyses achieving the sixth-grade level recommended by the AMA [39,45,66].

Analysis of Spanish-language PEMs achieving specific RGLs showed mixed results: 1 study successfully reduced readability from sixth to fifth grade using GPT-4.0 [40], whereas another achieved only a minimal reduction (ninth to eighth grade) despite targeting the sixth grade [37]. The third study did not report a specific target RGL [52].

Figure 4. Comparison of original and achieved RGL after prompting various LLMs to simplify PEMs to a predefined target RGL. LLMs: large language models, PEM: patient education material, RGL: reading grade level.



Other Linguistic Comprehensibility Indicators

The ability to reduce *text length* was assessed through word count in 14 studies [37,38,40,42,48,51,52,54-57,60,65,66] and sentence count in 5 studies [51,54,55,57,66]. GPT-3.5 reduced word count in all 8 analyses, and GPT-4.0 achieved reductions in nearly 80% of cases (15/19, 79%). Both models maintained readability improvements alongside text reduction in most instances: GPT-3.5 in 75% and GPT-4.0 in 89%. Other LLMs, such as Bard, Claude (2.0 and 3.5), Llama 2, and Copilot, similarly achieved word count reductions while maintaining readability. Sentence count decreased in 13 out of 17 (76%) analyses, with no notable differences between LLMs.

Lexical complexity was assessed in 8 studies [37,51,54-57,60,62]. All tested LLMs (GPT-3.5, GPT-4.0, Bard, Gemini, and Llama 2) reduced syllable-based metrics (polysyllabic words, syllables per word, and syllables per

sentence), except Claude 2 in 1 analysis [57]. GPT-4.0 also reduced the number of words exceeding 6 characters [62].

Syntactic complexity was assessed in 3 studies [37,62,66]. Multiple LLMs (GPT-3.5, GPT-4.0, Gemini, and Copilot) successfully reduced long sentences (>20-22 words) and passive voice usage.

Vocabulary complexity was assessed in 2 studies. LLMs effectively simplified vocabulary, reducing medical jargon, acronyms, uncommon words, and complex words [48,66].

Spanish-language studies also confirmed reductions in word count, syllable count, and long sentence frequency [37,40].

Content Fidelity

Identified Studies

A total of 20 studies examined whether the simplified texts remained factually correct and complete [36-38,40,43,46,48,49,51,52,54,55,57,58,61-66].

Factual Correctness

Factual correctness was assessed through content similarity in 6 studies [36-38,57,58,62] and content accuracy in 16 studies [37,38,40,43,46,49,51,54,55,58,61-66].

Content similarity was evaluated through automated and human assessment methods. Automated analyses using latent semantic analysis with cosine similarity values consistently demonstrated high semantic preservation. GPT-4.0 showed the strongest evidence base, producing "near identical" outputs in all 4 analyses [36,37,57,62], with 3 additional studies reporting "high similarity" [37,38,58]. GPT-3.5, Claude 2, and Llama 2 each achieved "near identical" similarity in single analyses [57]. Expert evaluations, conducted by either research team members or specialized health care professionals, further supported these findings. Human raters consistently confirmed that simplified texts did not contain extraneous information across all tested LLMs [57]. One study reporting quantitative expert assessment of GPT-4.0 yielded a mean similarity score of 0.72 on a 0-1 scale (where 1 represents identical content) [38].

Content accuracy was generally high but varied across studies. Automated F_1-scores, which reflect precision and recall (ie, whether a statement is true or false and whether it is present or absent), ranged from 72% to 92% for GPT-4.0 across different studies [37,38,62]. Expert assessments were inconsistent. For GPT-4.0, 8 studies reported no factual errors [43,51,53-55,58,64,65], 1 reported 90%-100% accuracy across 3 raters [52], and 2 studies using 5-point Likert scales reported accuracy between 3.55 and 4.1 (5=completely accurate) [38,66]. One notable outlier reported only 52% factual correctness [40]. For GPT-3.5, several studies reported no factual inaccuracies [46,49,51,54,55,61,63], and where a 5-point Likert scale was used, it scored 3.8 [66]. Other LLMs showed similar variability: Gemini received expert accuracy or suitability ratings of 48%-71% (with up to 14% judged inaccurate) [52], whereas other studies reported error-free outputs [51,65] or 10% inaccuracy rates [65]. Claude showed 5% inaccuracy in some analyses; Copilot received a mean score of 3.51; and Bard achieved 88.9% accuracy and suitability in 1 study [53] while being reported as error-free in other studies [54,55,63].

Factual Completeness

Factual completeness was assessed through content retention [43,46,48,49,52,66] and content omission [66] in fewer studies.

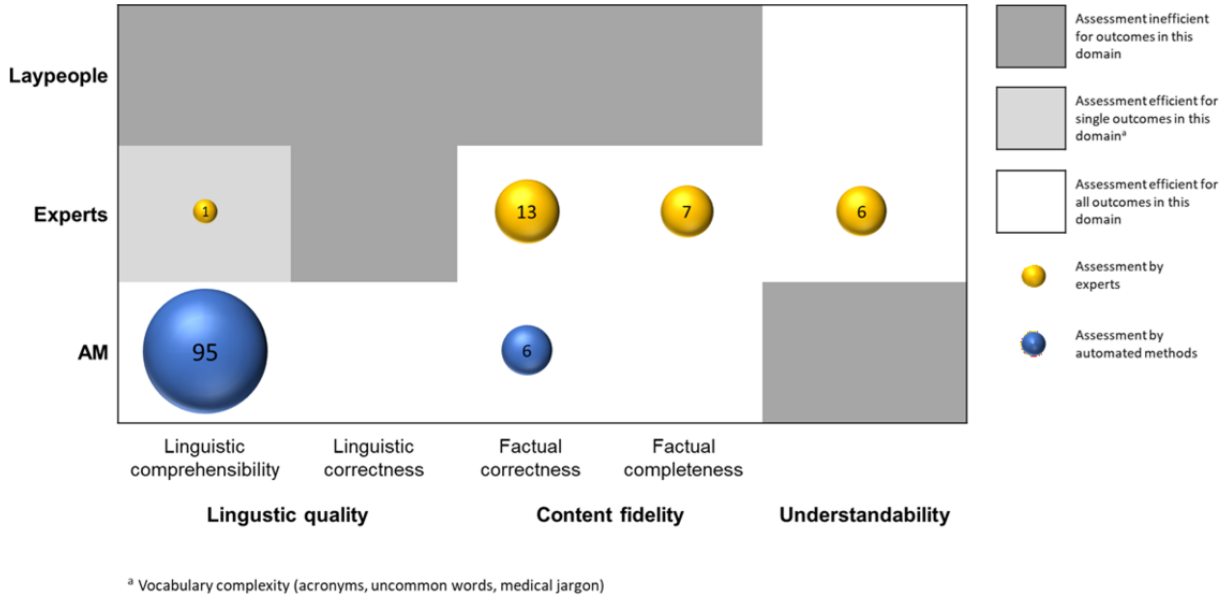
Content retention varied considerably across the evaluated LLMs. GPT-4.0 demonstrated the most robust performance when compared with GPT-3.5, Gemini, and Copilot [66]. Furthermore, GPT-4.0 maintained sufficient detail with some redundancies (eg, chronic cough was mentioned as a symptom of bronchitis 4 times) [43], and 95% of its outputs retained at least 75% of information [52]. Conversely, Gemini demonstrated reduced retention, with only 68% of outputs achieving the same retention threshold [52]. GPT-3.5 showed inconsistent retention, with studies reporting sufficient detail preservation [49], 80% average content retention [48], or truncation issues in 9% of outputs [46].

Content omission was evaluated in 1 study [66]. GPT-4.0 achieved the highest ratings for preserving essential information while simplifying, followed by Copilot. Conversely, Gemini received the lowest scores [66].

Overview of Text Quality Indicator Assessment Methods

As highlighted in Figure 5, linguistic comprehensibility was predominantly assessed using automated methods, as the manual assessment of objective indicators such as readability, sentence length, or word count is inefficient. By contrast, verifying factual correctness and completeness, as well as evaluating the understandability of simplified PEMs, inherently requires human judgment and cannot be fully automated. While content fidelity of simplified PEMs in the included studies was assessed by experts, this occurred less frequently overall. Crucially, the analysis reveals 2 major validation gaps: first, no study assessed the linguistic correctness of the simplified PEMs. Second, and most notably, no study involved laypeople in assessing the actual understandability of the simplified PEMs.

Figure 5. Evidence gap map of assessment methods used across studies for outcome domains (linguistic quality, content fidelity, and understandability) of simplified patient education materials, stratified by assessment type (laypeople, experts, and automated methods). AM: automated method.



Relationship Between Readability, Content Accuracy, and Understandability of Simplified PEMs

A total of 6 studies evaluated the understandability and clarity of simplified PEMs through human expert assessments, including research team members or health care professionals, using validated tools such as the Patient Education Materials Assessment Tool [51,54,55,65] and subjective assessments of whether the PEM was understandable for an average patient or successfully improved accessibility without losing key information [52,58]. However, only 4 studies directly examined the relationship between readability improvement, content accuracy, and overall expert-rated understandability of simplified PEMs [54,55,58,65] (Table 3).

Table 3. Relationship between readability, content accuracy, and understandability.

Large language model	Readability improvement[a], n/N (%)	Content accuracy maintenance[b], n/N (%)	Understandability success[c], n/N (%)	Medical topic	Studies
GPT-4.0	9/9 (100)	6/6 (100)	3/6 (50)	✓ [d]Childhood glaucoma, cancer, and living kidney donation ✗ [e]Idiopathic intracranial hypertension, heart disease, and stroke	[54,55,58,65]
GPT-3.5	3/4 (75)	2/2 (100)	0/2 (0)	✗ Idiopathic intracranial hypertension and childhood glaucoma	[54,55]
Bard	4/4 (100)	2/2 (100)	0/2 (0)	✗ Idiopathic intracranial hypertension and childhood glaucoma	[54,55]
Gemini 1.5	4/4 (100)	2/3 (67)	1/3 (33)	✓ Cancer ✗ Heart disease and stroke	[65]
Claude 3.5	4/4 (100)	1/3 (33)	1/3 (33)	✓ Heart disease ✗ Cancer and stroke	[65]

[a]Number of analyses with significant improvement in at least one readability metric (Flesch-Kincaid Grade Level, Flesch Reading Ease, Simple Measure of Gobbledygook, or Gunning Fog Index).

[b]Number of analyses where no factual errors were detected.

[c]Judgment by human experts

[d]Understandable.

[e]Not understandable.

The studies confirmed that GPT-3.5, GPT-4.0, Bard, Gemini 1.5, and Claude 3.5 consistently improved readability scores across standard metrics, except for GPT-3.5 using FKGL in 1 analysis [54]. Content accuracy was generally high but varied by LLM and medical field. GPT-3.5, GPT-4.0, and Bard achieved 100% content accuracy across multiple conditions.

However, Gemini 1.5 demonstrated 10% inaccuracy for heart disease content while maintaining 100% accuracy for cancer and stroke. Claude 3.5 showed the opposite pattern: 100% accuracy for heart disease but 5% inaccuracy for cancer and stroke. Understandability proved most challenging. GPT-4.0 emerged as the most capable, successfully generating readable, accurate, and understandable materials for childhood glaucoma [55] and kidney donation [62]; however, it failed for idiopathic intracranial hypertension [45]. Other models demonstrated greater limitations: GPT-3.5 and Bard generated readable and accurate outputs that experts deemed not understandable across multiple evaluations. Disease-specific performance varied substantially: GPT-4.0 generated understandable PEMs for cancer but not for cardiovascular conditions, whereas Claude 3.5 achieved understandability only for heart disease content [65].

Discussion

Principal Findings

This scoping review synthesized evidence on automated language processing technologies currently used to simplify PEMs into layperson-friendly language. A total of 31 studies met the inclusion criteria and examined LLMs for automatic text simplification. Notably, despite a comprehensive literature search, no studies specifically examined AI-supported writing assistants (eg, DeepL Write) or other AI-supported tools designed for automated text simplification. This scoping review has 3 key findings.

First, LLMs consistently improve readability metrics, with GPT-4.0 demonstrating the most reliable performance in comparison with other LLMs. However, achieving predefined RGLs remains challenging across all LLMs, particularly for GPT models at the lowest target level (fifth grade). Notably, GPT-4.0 does not consistently outperform GPT-3.5; rather, performance varies by target RGL. This indicates that newer models are not automatically better at achieving specific RGLs than older ones. Although Gemini and Claude show higher success rates, limited analyses preclude definitive conclusions. At the current stage of technological development, LLMs struggle to consistently achieve the recommended RGL for optimal lay comprehension, particularly the sixth-grade level recommended by the AMA. However, based on the available data, it remains unclear whether the failure to consistently achieve target RGLs stems from inherent model limitations or from heterogeneity in prompting strategies across the included studies.

Second, content fidelity assessments showed considerable variability. Although content similarity scores were generally high, content accuracy ranged from 48% to 100% across studies. This level of inaccuracy is concerning in medical contexts, where even minor errors can compromise patient safety.

Third, and most critically, this scoping review reveals a fundamental validation gap: no study has evaluated linguistic correctness (eg, grammar or typographical errors), and no study has assessed whether laypeople actually understand the simplified PEMs. The absence of reporting on linguistic correctness indicators suggests that authors implicitly assume that LLM-simplified texts are grammatically correct. However, this assumption can be problematic in medical contexts, where ambiguous phrasing or grammatical errors can alter the meaning of health instructions and potentially cause harm to patients. For instance, even minor errors in negation or word order can result in dangerous misinterpretations. Therefore, the failure to systematically assess and report linguistic correctness represents both a methodological oversight and a patient safety concern. Furthermore, the few studies that assessed overall understandability relied exclusively on expert assessments. However, experts such as health care professionals cannot reliably predict what laypeople understand or do not understand. This validation gap reveals a critical conceptual conflation: readability, as measured by formulas, captures surface-level text characteristics such as sentence length, word length, and syllable counts. However, it does not measure whether a reader actually understands the content. By contrast, understandability refers to the extent to which readers can extract meaning from text, apply the information to their own situation, and make informed decisions based on what they read. Consequently, improved readability metric scores do not guarantee that medical content is also factually correct, contextually appropriate, or genuinely understandable to the target group. This distinction is particularly significant in medical contexts, where patients must not only read but also correctly interpret and act upon health information. The complete absence of patient-centered outcome assessment represents the most significant finding of this scoping review.

Comparison With Previous Work

To the best of our knowledge, this is the first scoping review focusing specifically on automatic text simplification methods for PEMs. This distinguishes our work from previous research on medical AI, which has predominantly focused on text generation [69-78] rather than text simplification or the simplification of clinical records [79-82], rather than PEMs.

Nguyen et al [83] conducted a systematic review and meta-analysis of online PEMs related to cleft lip and palate. In line with our findings, they identified the critical absence of understandability testing, as only 1 study directly assessed patient understanding of simplified texts. In contrast to their disease-specific approach, our review spans a broader range of medical topics, evaluates linguistic quality and content fidelity, and employs a broader search strategy.

In a broader scoping review, Aydin et al [84] examined LLM applications across multiple domains of patient care, including education, engagement, workload reduction, patient-centered health customization, and communication. However, their search was limited to PubMed and was conducted in June 2024. They similarly reported readability improvements in studies addressing automatic text simplification. Their conclusion that LLMs can create accessible materials, help interpret complex information, and enhance patient-provider communication—while also noting that accuracy, readability issues, and ethical concerns require further development—aligns with the findings of our review.

The complete absence of direct testing of understandability with laypeople in all 31 studies is significant, reflecting a systemic methodological limitation in this field: the uncritical acceptance of readability formulas as valid measures of text simplification. These findings are further supported by studies indicating that standard readability formulas have important limitations. They primarily count variables such as sentence length, word length, or polysyllabic words, but do not assess whether the text aligns with the actual comprehension, context, expertise, semantic understanding, and textual coherence of the target audience. Common readability formulas cannot adequately assess the actual understandability of medical texts [85]. Furthermore, readability formulas may judge short technical terms and abbreviations as simple, despite being far less understandable than longer, everyday descriptions [86,87]. For example, the term "HbA_{1c}" is very short but often unknown, whereas the term "long-term blood sugar" is longer but more self-explanatory. Additionally, in-text explanations of terms are often needed for comprehension; however, they can lengthen and complicate sentences [88].

Research Implications

Overview

The findings of this review highlight several research gaps, constituting an urgent call to action for future studies. The validation gap identified in this review—namely, the complete absence of patient-centered comprehension testing and linguistic correctness assessment—should be addressed as a priority in future research. The following research gaps should guide the field forward:

Patient-Centered Evaluation

This represents the most significant research gap. Future studies must shift from purely algorithmic readability evaluation to patient-centered understandability testing [37,46,47,55-57,60,62,66]. Future studies should involve laypeople with varying health literacy levels, age, education, and cultural backgrounds in understandability assessments using validated methods, rather than relying solely on expert judgment or readability formulas [44,48,58]. For instance, Ondov et al [89] suggested 2 methods of comprehension assessment:

- Multiple-choice questions: After reading the original or simplified text, users answer constructed items that require understanding of the text to respond correctly. These questions are reliable when developed and validated properly, but are labor-intensive and require domain and assessment expertise.
- Cloze tests: These tests involve deleting words in a passage and asking users to supply the missing terms from context. They correlate well with other comprehension measures and are largely automatable. Common variants include basic cloze (masking important content words) and multiple-choice cloze (using distractors for each blank).

Prompt Engineering

Prompt engineering requires systematic investigation. In this scoping review, only 1 study [59] systematically investigated prompt design. Fanning et al [59] compared general and specific prompts using GPT-3.5 and GPT-4.0 and demonstrated that prompt design can markedly influence output quality. The study shows that well-designed, specific prompts enabled GPT-3.5 to achieve readability improvements comparable to GPT-4.0, suggesting that prompt optimization may be more influential than model generation for text simplification tasks. Their specific prompts included explicit references to readability measures, detailed instructions for writing style (eg, shortening sentences and simplifying words), specification of target readability levels, and explicit constraints requiring that rewritten texts maintain the original meaning and information. While these findings are based on a single study, they offer a possible explanation for the inconsistent results observed across the included studies. The wide variability in content accuracy (48%-100%) and frequent failure to achieve specific RGLs could be attributed to differences in prompt design rather than solely to inherent model limitations. However, further research is needed to confirm this. Future studies should establish best practices for prompt engineering [36,37,39,41,42,45,47,51,53,59,61,65]. This involves systematically comparing different prompting strategies. For instance, moving beyond basic zero-shot prompts to incorporate advanced techniques such as few-shot prompting (providing models with examples of desired simplified outputs) [37,65] or chain-of-thought prompting (guiding the model through a step-by-step reasoning process). Chain-of-thought prompting could instruct the LLM to first identify complex medical terms, then suggest layperson equivalents, then shorten sentences, and finally check for factual accuracy [90]. Similarly, exploring the impact of stylistic constraints, detailed context-specific instructions, and the use of word-substitution lists is essential [37]. Further, future studies should focus on iterative prompt design with feedback loops for continuous refinement based on expert and patient evaluations, as well as evaluating prompt stability across LLM generations [47,61]. The reliability of prompts should also be assessed through test-retest consistency, ensuring that identical prompts yield consistent, high-quality outputs over time [91]. Finally, a key long-term goal should be the development of a validated, open-access prompt library with demonstrated efficacy for medical text simplification, enabling standardization across research and clinical applications. While existing technical prompt collections in the medical domain primarily address reasoning and question-answering tasks (eg, Microsoft MedPrompt), they were neither designed nor clinically validated for medical text simplification [92].

Technological Advancements

The rapid progress in LLM technology requires continuous evaluation. Future studies should systematically compare emerging LLMs, particularly beyond the GPT series, to identify models with unique strengths (eg, long-context handling with Claude 2) [44,47,48,59,65]. Furthermore, cost-effectiveness comparisons between commercial and open-source LLMs are also needed [47,55,63].

Geographic and Linguistic Diversity

The included studies predominantly focused on English-language PEMs within the US health care context, with only 3 studies [37,40,47] evaluating Spanish texts.

Consequently, the generalizability of these findings to other languages and cultural health care systems is limited. Languages with complex morphology (eg, German or Slavic languages) or distinct syntactic structures may present different challenges for automated simplification than English. Furthermore, cultural differences in medical communication styles may affect how "layperson-friendly" is defined and achieved in other regions. Therefore, future studies should investigate simplification performance across diverse linguistic and cultural contexts [48,49,62,65,66].

Medical Field Scope

The included studies predominantly focused on surgical specialties and ophthalmology, while numerous medical fields remain unexplored with regard to LLM-based text simplification of PEMs. No studies addressed, for example, dentistry, dermatology, psychiatry, or preventive medicine. This is notable, as these fields frequently rely on PEMs containing complex terminology. Dentistry, for example, involves jargon-heavy procedures in areas such as orthodontics that may be difficult for patients to understand [93]. Future research should investigate LLM-based text simplification across a broader range of medical fields, especially rare diseases and less prevalent health conditions, where LLMs with limited training data may be more susceptible to errors [48,65].

Linguistic Correctness

None of the included studies assessed text quality indicators related to linguistic correctness. Thus, future evaluations should incorporate metrics for grammar and typographical errors, as these factors are fundamental to establishing credibility and trust, independent of their direct effect on comprehension [94,95].

Practice Implications

Based on the findings of this scoping review, this study has several recommendations for clinical practice. First, LLMs should be used as assistive rather than autonomous tools for simplifying PEMs used in clinical practice [51]. This implication is supported by a recent study showing that LLMs can approximate targeted RGLs when simplifying health information materials about psychiatry, but their outputs are inconsistent, with significant variability in reading levels and deviations from the intended content, making them unsuitable for standalone deployment in health care settings [96]. Second, all AI-simplified PEMs must undergo mandatory expert review by qualified health care professionals to verify content fidelity and clinical appropriateness before dissemination to laypeople [38-41,45-49,52,59-61,66]. Third, health care organizations using AI-assisted text simplification tools should establish clear protocols defining quality checkpoints and approval workflows. Continuous monitoring of performance can track both readability improvements and error rates across different medical domains [59].

Although LLMs hold considerable promise for enhancing patient health literacy through simplified text versions, their use in clinical practice requires careful attention to inherent limitations, including the risk of reinforcing biases and stereotypes embedded in training data. The use of AI in health care communication requires an ongoing commitment to the accuracy of generated content [50].

Limitations

Although this scoping review followed rigorous and systematic methods, some limitations must be acknowledged. First, rule-based technologies were excluded from this review. This decision was made to focus on the contemporary and rapidly evolving landscape dominated by generative AI, which offers greater scalability and adaptability for text simplification. Second, although different prompting strategies were used across the included studies, the quality of these prompts was not evaluated. In accordance with scoping review methodology, which aims to map the extent and nature of available evidence rather than critically appraise methodological quality, a systematic assessment of prompt design was not conducted. Consequently, it was not possible to determine whether the observed failure to meet specific target RGLs reflects model limitations or suboptimal prompting strategies. However, such an analysis could have provided deeper insights into optimal prompt engineering for PEM simplification. Lastly, the database search was conducted in May 2025, and the gray literature search in July 2025. Given the rapid pace of AI development, particularly regarding LLMs, new relevant studies may have emerged since, and some findings may already become outdated shortly after publication. Finally, the included studies were heavily skewed toward English-language PEMs and US-based contexts, limiting the generalizability of the findings, as text simplification approaches may perform differently across languages with varying grammatical complexity (eg, morphology, syntax). Moreover, LLMs are predominantly trained on English data [97], which may lead to lower performance in other languages. Additionally, the readability metrics used (eg, US RGLs) are specific to the US educational system and may not directly translate to plain-language guidelines in other cultural health care systems.

Conclusions

To our knowledge, this is the first scoping review to comprehensively synthesize evidence on automated language processing technologies for PEM simplification, systematically mapping linguistic, quality, and content fidelity outcomes.

The findings of our scoping review indicate that although LLMs can improve the readability of PEMs, they predominantly fail to achieve the recommended sixth-grade RGL. However, the most significant finding is the identification of a critical validation gap: no study has assessed whether laypeople actually understood the simplified PEMs, nor has any study evaluated linguistic correctness. Coupled with variable content accuracy and reliance solely on readability formulas that poorly predict actual understanding by laypeople, these findings have important implications for clinical practice. Currently, LLMs should serve as assistive tools rather than autonomous solutions. All AI-simplified materials must undergo mandatory expert review to verify content fidelity before dissemination to laypeople. Future research must urgently shift from purely algorithmic evaluation to patient-centered validation to directly assess laypeople's comprehension.

Acknowledgments

The authors declare the use of generative artificial intelligence (GAI) in the research and writing process. According to the Generative AI Disclosure and Transparency (GAIDeT) taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: proofreading and editing, translation, and reformatting. The GAI tool used was DeepL Write. Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Funding

This study was funded by the Austrian Science Fund (FWF; grant FG 11-B), as part of the project “Human-Centered Interactive Adaptive Visual Approaches in High-Quality Health Information.” The funder had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

All extracted data relevant to the research aims are included in the manuscript and its multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 84 KB-Multimedia Appendix 1]

Multimedia Appendix 2

PRISMA-S (PRISMA Statement for Reporting Literature Searches in Systematic Reviews) checklist.
[DOCX File , 17 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Search strategies for bibliographic databases.
[DOCX File , 21 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Excluded studies.
[DOCX File , 38 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Trend analysis of publication dates for studies evaluating large language models for the text simplification of patient education materials (2019-2025).
[PNG File , 84 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Overview of prompts used for automated text simplification in the included studies.
[DOCX File , 62 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Overview of reported outcomes in included studies.
[DOCX File , 66 KB-Multimedia Appendix 7]

References

1. Niederdeppe J, Boyd AD, King AJ, Rimal RN. Strategies for effective public health communication in a complex information environment. Annu Rev Public Health. Apr 2025;46(1):411-431. [FREE Full text] [doi: 10.1146/annurev-publhealth-071723-120721] [Medline: 39656948]

2. Berkman ND, Davis TC, McCormack L. Health literacy: what is it? J Health Commun. 2010;15 Suppl 2:9-19. [doi: 10.1080/10810730.2010.499985] [Medline: 20845189]
3. Sudirman. The role of health literacy in improving health outcomes: challenges, interventions, and policies. J Health Lit Qual Res. Mar 31, 2022;2(1):15-30. [doi: 10.61194/jhlqr.v2i1.530]
4. Wahrenbrock T, Landry K, Amin DP, Rizvanolli L, Chandrasekaran R, Mycyk MB, et al. Medical jargon is often misunderstood by emergency department patients. Am J Emerg Med. Oct 2025;96:25-29. [doi: 10.1016/j.ajem.2025.06.012] [Medline: 40513550]
5. Kumar PM, Nagate RR, Alqahtani SM, Sarma GSN, Supraja S. Role of jargon in the patient-doctor communication in the dental healthcare sector-a systematic review and meta-analysis. J Educ Health Promot. 2023;12:198. [FREE Full text] [doi: 10.4103/jehp.jehp_1442_22] [Medline: 37545998]
6. Delaney FT, Doinn T, Broderick JM, Stanley E. Readability of patient education materials related to radiation safety: what are the implications for patient-centred radiology care? Insights Imaging. Oct 21, 2021;12(1):148. [FREE Full text] [doi: 10.1186/s13244-021-01094-3] [Medline: 34674063]
7. Bhattad PB, Pacifico L. Empowering patients: promoting patient education and health literacy. Cureus. Jul 2022;14(7):e27336. [FREE Full text] [doi: 10.7759/cureus.27336] [Medline: 36043002]
8. Deb B. Personal health literacy. Agency for Healthcare Research and Quality (AHRQ). Aug 30, 2023. URL: https://psnet.ahrq.gov/primer/personal-health-literacy [accessed 2025-11-18]
9. Weis B. Health Literacy: A Manual for Clinicians. Chicago, IL. American Medical Association Foundation and American Medical Association; 2003.
10. Ghanem D, Covarrubias O, Harris AB, Shafiq B. Readability of the Orthopaedic Trauma Association patient education tool. J Orthop Trauma. Aug 01, 2023;37(8):e307-e311. [doi: 10.1097/BOT.0000000000002593] [Medline: 36862992]
11. Cheng BT, Kim AB, Tanna AP. Readability of online patient education materials for glaucoma. J Glaucoma. Jun 01, 2022;31(6):438-442. [doi: 10.1097/IJG.0000000000002012] [Medline: 35283441]
12. Kamath D, Mcintyre S, Peskin E, Stratman S, Agarwal N, Kamath PD, et al. Readability of online patient education materials for interventional pain procedures. Cureus. Sep 27, 2020;12(9):e10684. [FREE Full text] [doi: 10.7759/cureus.10684] [Medline: 33133850]
13. Armache M, Assi S, Wu R, Jain A, Lu J, Gordon L, et al. Readability of patient education materials in head and neck cancer: a systematic review. JAMA Otolaryngol Head Neck Surg. Aug 01, 2024;150(8):713-724. [doi: 10.1001/jamaoto.2024.1569] [Medline: 38900443]
14. Khan S, Moon J, Martin CA, Bowden E, Chen J, Tsui E, et al. Readability and suitability of online uveitis patient education materials. Ocul Immunol Inflamm. Sep 2024;32(7):1175-1179. [doi: 10.1080/09273948.2023.2203759] [Medline: 37145033]
15. Hartnett DA, Philips AP, Daniels AH, Blankenhorn BD. Readability of online foot and ankle surgery patient education materials. Foot Ankle Spec. Feb 2025;18(1):9-18. [doi: 10.1177/19386400221116463] [Medline: 35934974]
16. Okuhara T, Furukawa E, Okada H, Yokota R, Kiuchi T. Readability of written information for patients across 30 years: a systematic review of systematic reviews. Patient Educ Couns. Jun 2025;135:108656. [FREE Full text] [doi: 10.1016/j.pec.2025.108656] [Medline: 40068244]
17. Caverly TJ, Hayward RA. Dealing with the lack of time for detailed shared decision-making in primary care: everyday shared decision-making. J Gen Intern Med. Oct 2020;35(10):3045-3049. [FREE Full text] [doi: 10.1007/s11606-020-06043-2] [Medline: 32779137]
18. Hemsley B, Balandin S, Worrall L. Nursing the patient with complex communication needs: time as a barrier and a facilitator to successful communication in hospital. J Adv Nurs. Jan 2012;68(1):116-126. [doi: 10.1111/j.1365-2648.2011.05722.x] [Medline: 21831131]
19. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. ACM Trans Intell Syst Technol. Aug 18, 2025;16(5):1-72. [doi: 10.48550/arXiv.2307.06435]
20. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med (Lond). Jan 21, 2025;5(1):26. [FREE Full text] [doi: 10.1038/s43856-024-00717-2] [Medline: 39838160]
21. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Qureshi F, Cheungpasitporn W. Innovating personalized nephrology care: exploring the potential utilization of ChatGPT. J Pers Med. Dec 04, 2023;13(12):1-21. [FREE Full text] [doi: 10.3390/jpm13121681] [Medline: 38138908]
22. Eid K, Eid A, Wang D, Raiker RS, Chen S, Nguyen J. Optimizing ophthalmology patient education via chatbot-generated materials: readability analysis of AI-generated patient education materials and The American Society of Ophthalmic Plastic and Reconstructive Surgery patient brochures. Ophthalmic Plast Reconstr Surg. 2024;40(2):212-216. [doi: 10.1097/IOP.0000000000002549] [Medline: 37972974]
23. Alessandri-Bonetti M, Liu HY, Palmesano M, Nguyen VT, Egro FM. Online patient education in body contouring: a comparison between Google and ChatGPT. J Plast Reconstr Aesthet Surg. Dec 2023;87:390-402. [doi: 10.1016/j.bjps.2023.10.091] [Medline: 37939643]

24. Espinosa-Zaragoza I, Abreu-Salas J, Lloret E, Pozo P, Palomar M. A review of research-based automatic text simplification tools. 2023. Presented at: The 14th International Conference on Recent Advances in Natural Language Processing; September 4-6, 2023; Shoumen, Bulgaria. [doi: 10.26615/978-954-452-092-2_036]
25. Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J. A survey of large language models in medicine: progress, application, and challenge. arXiv. Preprint posted online November 9, 2023. [FREE Full text] [doi: 10.48550/arXiv.2311.05112]
26. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA, Devleesschauwer B, et al. Correction: World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. PLoS Med. Dec 2015;12(12):e1001940. [FREE Full text] [doi: 10.1371/journal.pmed.1001940] [Medline: 26701262]
27. Aromataris EL, Porritt K, Pilla B, Jordan Z. JBI Manual for Evidence Synthesis - 2024 Edition. JBI. URL: https://synthesismanual.jbi.global [accessed 2025-09-08]
28. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Tunçalp, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 02, 2018;169(7):467-473. [doi: 10.7326/m18-0850]
29. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S Group. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev. Jan 26, 2021;10(1):39. [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
30. Krenn C, Wilfling D, Loder C, Jeitler K, Siebenhofer-Kroitzsch A, Semlitsch T. Automated approaches to simplifying complex medical texts for laypersons: a scoping review protocol. Open Science Framework (OSF). URL: https://osf.io/wrb38/ [accessed 2025-11-18]
31. Schreck T. Nutzerzentrierte adaptive ansätze zur gesundheitsinformation (human-centered adaptive approaches for health information). Austrian Science Fund (FWF). Vienna. URL: https://www.fwf.ac.at/forschungsradar/10.55776/FG11 [accessed 2025-10-14]
32. Koster J. PubReMiner. PubMed. 2014. URL: https://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi [accessed 2025-05-14]
33. Better synonyms. TopAI. URL: https://topai.tools/t/better-synonyms [accessed 2025-05-14]
34. Zhang C, Liu S, Zhou X, Zhou S, Tian Y, Wang S, et al. Examining the role of large language models in orthopedics: systematic review. J Med Internet Res. Nov 15, 2024;26:e59607. [FREE Full text] [doi: 10.2196/59607] [Medline: 39546795]
35. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. J Med Internet Res. Nov 05, 2024;26:e56532. [FREE Full text] [doi: 10.2196/56532] [Medline: 39499913]
36. Spina A, Fereydouni P, Tang J, Andalib S, Picton B, Fox A. Tailoring glaucoma education using large language models: addressing health disparities in patient comprehension. Medicine (Baltimore). Jan 10, 2025;104(2):e41059. [FREE Full text] [doi: 10.1097/MD.0000000000041059] [Medline: 39792725]
37. Reaver C, Pereira D, Carrillo E, Marcos C, Goldfarb C. Evaluating the performance of artificial intelligence for improving readability of online English- and Spanish-language orthopaedic patient educational material: challenges in bridging the digital divide. The Journal of bone and joint surgery. 2025;107(8):e36. [doi: 10.2106/jbjs.24.01078]
38. Picton B, Andalib S, Spina A, Camp B, Solomon SS, Liang J, et al. Assessing AI simplification of medical texts: readability and content fidelity. Int J Med Inform. Mar 2025;195:105743. [FREE Full text] [doi: 10.1016/j.ijmedinf.2024.105743] [Medline: 39667051]
39. Chandra K, Ghilzai U, Lawand J, Ghali A, Fiedler B, Ahmed AS. Improving readability of shoulder and elbow surgery online patient education material with Chat GPT (Chat Generative Pretrained Transformer) 4. J Shoulder Elbow Surg. Nov 2025;34(11):e1119-e1124. [doi: 10.1016/j.jse.2025.02.025] [Medline: 40118438]
40. Busigó Torres R, Restrepo M, Stern BZ, Yahuaca BI, Buerba RA, García IA, et al. Artificial intelligence shows limited success in improving readability levels of Spanish-language orthopaedic patient education materials. Clin Orthop Relat Res. Feb 11, 2025;483(7):1185-1192. [doi: 10.1097/CORR.0000000000003413] [Medline: 39937452]
41. Zaki HA, Mai M, Abdel-Megid H, Liew SQR, Kidanemariam S, Omar AS, et al. Using ChatGPT to improve readability of interventional radiology procedure descriptions. Cardiovasc Intervent Radiol. Aug 2024;47(8):1134-1141. [doi: 10.1007/s00270-024-03803-z] [Medline: 38981939]
42. Shehab AA, Shedd KE, Alamah W, Mardini S, Bite U, Gibreel W. Bridging gaps in health literacy for cleft lip and palate: the role of artificial intelligence and interactive educational materials. Cleft Palate Craniofac J. Jan 2026;63(1):65-71. [doi: 10.1177/10556656241289653] [Medline: 39380385]
43. Patel EA, Fleischer L, Filip P, Eggerstedt M, Hutz M, Michaelides E, et al. The use of artificial intelligence to improve readability of otolaryngology patient education materials. Otolaryngol Head Neck Surg. Aug 2024;171(2):603-608. [doi: 10.1002/ohn.816] [Medline: 38751109]
44. Li G, Lin MX, Cui D, Mathews PM, Akpek EK. Enhancing online cataract surgery patient education materials through artificial intelligence. Can J Ophthalmol. Apr 2025;60(2):e197-e202. [doi: 10.1016/j.jcjo.2024.07.018] [Medline: 39163992]
45. Vallurupalli M, Shah ND, Vyas RM. Optimizing readability of patient-facing hand surgery education materials using Chat Generative Pretrained Transformer (ChatGPT) 3.5. J Hand Surg Am. Oct 2024;49(10):986-991. [doi: 10.1016/j.jhsa.2024.05.007] [Medline: 38970600]

46. Oliva AD, Pasick LJ, Hoffer ME, Rosow DE. Improving readability and comprehension levels of otolaryngology patient education materials using ChatGPT. Am J Otolaryngol. 2024;45(6):104502. [doi: 10.1016/j.amjoto.2024.104502] [Medline: 39197330]
47. Sudharshan R, Shen A, Gupta S, Zhang-Nunes S. Assessing the utility of ChatGPT in simplifying text complexity of patient educational materials. Cureus. Mar 2024;16(3):e55304. [FREE Full text] [doi: 10.7759/cureus.55304] [Medline: 38559518]
48. Ayre J, Mac O, McCaffery K, McKay BR, Liu M, Shi Y, et al. New frontiers in health literacy: using ChatGPT to simplify health information for people in the community. J Gen Intern Med. Mar 08, 2024;39(4):573-577. [FREE Full text] [doi: 10.1007/s11606-023-08469-w] [Medline: 37940756]
49. Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? Clin Orthop Relat Res. Nov 01, 2023;481(11):2260-2267. [doi: 10.1097/CORR.0000000000002668] [Medline: 37116006]
50. Baldwin AJ. An artificial intelligence language model improves readability of burns first aid information. Burns. Jun 2024;50(5):1122-1127. [doi: 10.1016/j.burns.2024.03.005] [Medline: 38492982]
51. Dihan QA, Brown AD, Chauhan MZ, Alzein AF, Abdelnaem SE, Kelso SD, et al. Leveraging large language models to improve patient education on dry eye disease. Eye (Lond). Apr 2025;39(6):1115-1122. [doi: 10.1038/s41433-024-03476-5] [Medline: 39681711]
52. Gupta M, Gupta P, Ho C, Wood J, Guleria S, Virostko J. Can generative AI improve the readability of patient education materials at a radiology practice? Clin Radiol. Nov 2024;79(11):e1366-e1371. [doi: 10.1016/j.crad.2024.08.019] [Medline: 39266371]
53. Kianian R, Sun D, Crowell EL, Tsui E. The use of large language models to generate education materials about uveitis. Ophthalmol Retina. Feb 2024;8(2):195-201. [FREE Full text] [doi: 10.1016/j.oret.2023.09.008] [Medline: 37716431]
54. Dihan QA, Brown AD, Zaldivar AT, Chauhan MZ, Eleiwa TK, Hassan AK, et al. Advancing patient education in idiopathic intracranial hypertension: the promise of large language models. Neurol Clin Pract. Feb 2025;15(1):e200366. [doi: 10.1212/CPJ.0000000000200366] [Medline: 39399571]
55. Dihan Q, Chauhan MZ, Eleiwa TK, Hassan AK, Sallam AB, Khouri AS, et al. Using large language models to generate educational materials on childhood glaucoma. Am J Ophthalmol. Sep 2024;265:28-38. [doi: 10.1016/j.ajo.2024.04.004] [Medline: 38614196]
56. Dihan Q, Chauhan MZ, Eleiwa TK, Brown AD, Hassan AK, Khodeiry MM, et al. Large language models: a new frontier in paediatric cataract patient education. Br J Ophthalmol. Sep 20, 2024;108(10):1470-1476. [doi: 10.1136/bjo-2024-325252] [Medline: 39174290]
57. Andalib S, Solomon S, Picton B, Spina A, Scolaro J, Nelson A. Source characteristics influence AI-enabled orthopaedic text simplification: recommendations for the future. JB JS Open Access. 2025;10(1):1-14. [FREE Full text] [doi: 10.2106/JBJS.OA.24.00007] [Medline: 39781102]
58. Garcia Valencia OA, Thongprayoon C, Miao J, Suppadungsuk S, Krisanapan P, Craici IM, et al. Empowering inclusivity: improving readability of living kidney donation information with ChatGPT. Front Digit Health. Apr 10, 2024;6:1366967. [FREE Full text] [doi: 10.3389/fdgth.2024.1366967] [Medline: 38659656]
59. Fanning JE, Escobar-Domingo MJ, Foppiani J, Lee D, Miller AS, Janis JE, et al. Improving readability and automating content analysis of plastic surgery webpages with ChatGPT. J Surg Res. Jul 2024;299:103-111. [doi: 10.1016/j.jss.2024.04.006] [Medline: 38749313]
60. Manasyan A, Lasky S, Jolibois M, Moshal T, Roohani I, Munabi N, et al. Expanding accessibility in cleft care: the role of artificial intelligence in improving literacy of alveolar bone grafting information. Cleft Palate Craniofac J. Nov 2025;62(11):1873-1880. [doi: 10.1177/10556656241281453] [Medline: 39246230]
61. Vallurupalli M, Shah ND, Vyas RM. Validation of ChatGPT 3.5 as a tool to optimize readability of patient-facing craniofacial education materials. Plast Reconstr Surg Glob Open. Feb 2024;12(2):e5575. [FREE Full text] [doi: 10.1097/GOX.0000000000005575] [Medline: 38313589]
62. Abreu AA, Murimwa GZ, Farah E, Stewart JW, Zhang L, Rodriguez J, et al. Enhancing readability of online patient-facing content: the role of AI chatbots in improving cancer information accessibility. J Natl Compr Canc Netw. May 15, 2024;22(2 D):e237334. [doi: 10.6004/jnccn.2023.7334] [Medline: 38749478]
63. Rouhi AD, Ghanem YK, Yolchieva L, Saleh Z, Joshi H, Moccia MC, et al. Can artificial intelligence improve the readability of patient education materials on aortic stenosis? A pilot study. Cardiol Ther. Mar 2024;13(1):137-147. [FREE Full text] [doi: 10.1007/s40119-023-00347-0] [Medline: 38194058]
64. Singh A, Gupta N, Chien DL, Singh R, Sachdeva A, Danasekaran K, et al. ChatGPT-4 in neurosurgery: improving patient education materials. Neurosurgery. Jan 01, 2026;98(1):147-160. [doi: 10.1227/neu.0000000000003606] [Medline: 40704789]
65. Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: cross-sectional study. J Med Internet Res. Jun 04, 2025;27:e69955. [FREE Full text] [doi: 10.2196/69955] [Medline: 40465378]
66. Naghdi M, Cao P, Essers R, Heijligers M, Paulussen ADC, van der Lugt A, et al. Artificial intelligence-simplified information to advance reproductive genetic literacy and health equity. Hum Reprod. Sep 01, 2025;40(9):1681-1688. [doi: 10.1093/humrep/deaf135] [Medline: 40692125]

67. Scott B. Readability scoring system plus. Readability Formulas. URL: https://readabilityformulas.com/readability-scoring-system.php [accessed 2026-02-03]
68. Hansberry DR, Agarwal N, Gonzales SF, Baker SR. Are we effectively informing patients? A quantitative analysis of on-line patient education resources from the American Society of Neuroradiology. AJNR Am J Neuroradiol. Jul 2014;35(7):1270-1275. [FREE Full text] [doi: 10.3174/ajnr.A3854] [Medline: 24763420]
69. Nasra M, Jaffri R, Pavlin-Premrl D, Kok HK, Khabaza A, Barras C, et al. Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. Intern Med J. Jan 25, 2025;55(1):20-34. [doi: 10.1111/imj.16607] [Medline: 39720869]
70. Zhao F, He H, Liang J, Cen J, Wang Y, Lin H, et al. Benchmarking the performance of large language models in uveitis: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, Google Gemini, and Anthropic Claude3. Eye (Lond). Apr 2025;39(6):1132-1137. [doi: 10.1038/s41433-024-03545-9] [Medline: 39690303]
71. Zhang Q, Wu Z, Song J, Luo S, Chai Z. Comprehensiveness of large language models in patient queries on gingival and endodontic health. Int Dent J. Feb 2025;75(1):151-157. [FREE Full text] [doi: 10.1016/j.identj.2024.06.022] [Medline: 39147663]
72. Zhan Y, Chen X, Ye F, Wu Z, Usman M, Yuan Z, et al. Evaluating AI chatbot responses to postkidney transplant inquiries. Transplant Proc. Mar 2025;57(2):394-405. [doi: 10.1016/j.transproceed.2024.12.028] [Medline: 39814625]
73. Zeljkovic I, Novak M, Jordan A, Lisicic A, Nemeth-Blažić T, Pavlovic N, et al. Evaluating ChatGPT-4's correctness in patient-focused informing and awareness for atrial fibrillation. Heart Rhythm O2. Jan 2025;6(1):58-63. [FREE Full text] [doi: 10.1016/j.hroo.2024.10.005] [Medline: 40224268]
74. Yoo M, Jang CW. Presentation suitability and readability of ChatGPT's medical responses to patient questions about on knee osteoarthritis. Health Informatics J. Jan 19, 2025;31(1):14604582251315587. [FREE Full text] [doi: 10.1177/14604582251315587] [Medline: 39828887]
75. Yıldız HA, Söğütdelen E. AI chatbots as sources of STD information: a study on reliability and readability. J Med Syst. Apr 03, 2025;49(1):43. [doi: 10.1007/s10916-025-02178-z] [Medline: 40178771]
76. Yılmaz IBE, Doğan L. Talking technology: exploring chatbots as a tool for cataract patient education. Clin Exp Optom. Jan 09, 2025;108(1):56-64. [doi: 10.1080/08164622.2023.2298812] [Medline: 38194585]
77. Xu Q, Wang J, Chen X, Wang J, Li H, Wang Z, et al. Assessing the efficacy of ChatGPT prompting strategies in enhancing thyroid cancer patient education: a prospective study. J Med Syst. Jan 17, 2025;49(1):11. [doi: 10.1007/s10916-024-02129-0] [Medline: 39820814]
78. Schwartzman JD, Shaath MK, Kerr MS, Green CC, Haidukewych GJ. ChatGPT is an unreliable source of peer-reviewed information for common total knee and hip arthroplasty patient questions. Adv Orthop. Jan 06, 2025;2025(1):5534704. [FREE Full text] [doi: 10.1155/aort/5534704] [Medline: 39817149]
79. Yang X, Xiao Y, Liu D, Shi H, Deng H, Huang J, et al. Enhancing physician-patient communication in oncology using GPT-4 through simplified radiology reports: multicenter quantitative study. J Med Internet Res. Apr 17, 2025;27:e63786. [FREE Full text] [doi: 10.2196/63786] [Medline: 40245397]
80. Eisinger F, Holderried F, Mahling M, Stegemann-Philipps C, Herrmann-Werner A, Nazarenus E, et al. What's going on with me and how can i better manage my health? The potential of GPT-4 to transform discharge letters into patient-centered letters to enhance patient safety: prospective, exploratory study. J Med Internet Res. Jan 21, 2025;27:e67143. [FREE Full text] [doi: 10.2196/67143] [Medline: 39836954]
81. Can E, Uller W, Vogt K, Doppler MC, Busch F, Bayerl N, et al. Large language models for simplified interventional radiology reports: a comparative analysis. Acad Radiol. Feb 2025;32(2):888-898. [FREE Full text] [doi: 10.1016/j.acra.2024.09.041] [Medline: 39353826]
82. Li HH, Moon JT, Kumar S, Ricci J, Sim N, Bercu ZL, et al. Evaluation of multilingual simplifications of IR procedural reports using GPT-4. J Vasc Interv Radiol. Apr 2025;36(4):696-703.e1. [doi: 10.1016/j.jvir.2025.01.002] [Medline: 39793700]
83. Nguyen AT, Li RA, Gosain AK, Galiano RD. Readability of online patient education materials for cleft care: a systematic review and meta-analysis. Cleft Palate Craniofac J. May 2026;63(5):1223-1232. [doi: 10.1177/10556656251327803] [Medline: 40095963]
84. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898. [FREE Full text] [doi: 10.3389/fmed.2024.1477898] [Medline: 39534227]
85. Zheng J, Yu H. Readability formulas and user perceptions of electronic health records difficulty: a corpus study. J Med Internet Res. Mar 02, 2017;19(3):e59. [FREE Full text] [doi: 10.2196/jmir.6962] [Medline: 28254738]
86. Kauchak D, Leroy G. Moving beyond readability metrics for health-related text simplification. IT Prof. 2016;18(3):45-51. [FREE Full text] [doi: 10.1109/MITP.2016.50] [Medline: 27698611]
87. Leroy G, Kauchak D. The effect of word familiarity on actual and perceived text difficulty. J Am Med Inform Assoc. Feb 2014;21(e1):e169-e172. [FREE Full text] [doi: 10.1136/amiajnl-2013-002172] [Medline: 24100710]
88. Kandula S, Curtis D, Zeng-Treitler Q. A semantic and syntactic text simplification tool for health content. AMIA Annu Symp Proc. Nov 13, 2010;2010:366-370. [FREE Full text] [Medline: 21347002]

89. Ondov B, Attal K, Demner-Fushman D. A survey of automated methods for biomedical text simplification. J Am Med Inform Assoc. Oct 07, 2022;29(11):1976-1988. [FREE Full text] [doi: 10.1093/jamia/ocac149] [Medline: 36083212]
90. Zhu D, Xiong Y, Zhang J, Xie X, Xia C. Understanding before reasoning: enhancing chain-of-thought with iterative summarization pre-prompting. arXiv. Preprint posted online January 8, 2025. [FREE Full text] [doi: 10.48550/arXiv.2501.04341]
91. Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. J Med Internet Res. Sep 15, 2025;27:e72644-e72644. [FREE Full text] [doi: 10.2196/72644] [Medline: 40955776]
92. Horvitz E. The power of prompting Microsoft. Microsoft Corporation. URL: https://www.microsoft.com/en-us/research/blog/the-power-of-prompting/ [accessed 2026-02-27]
93. Benning A, Madadian MA, Seehra J, Fan K. Patients understanding of terminology commonly used during combined orthodontic-orthognathic treatment. Surgeon. Oct 2021;19(5):e193-e198. [doi: 10.1016/j.surge.2020.09.012] [Medline: 33423926]
94. Liu J, Song S, Zhang Y. Linguistic features and consumer credibility judgment of online health information. ResearchGate. 2021. URL: https://www.researchgate.net/publication/350511487_Linguistic_features_and_consumer_credibility_judgment_of_online_health_information [accessed 2026-04-24]
95. Witchel HJ, Thompson GA, Jones CI, Westling CEI, Romero J, Nicotra A, et al. Spelling errors and shouting capitalization lead to additive penalties to trustworthiness of online health information: randomized experiment with laypersons. J Med Internet Res. Jun 10, 2020;22(6):e15171. [FREE Full text] [doi: 10.2196/15171] [Medline: 32519676]
96. Aich A, Liu T, Giorgi S, Isman K, Bobojonova R, Ungar L, et al. Language models in digital psychiatry: challenges with simplification of healthcare materials. NPP Digit Psychiatry Neurosci. May 22, 2025;3(1):10. [doi: 10.1038/s44277-025-00029-w] [Medline: 41361013]
97. Qin L, Chen Q, Zhou Y, Chen Z, Li Y, Liao L, et al. A survey of multilingual large language models. Patterns (N Y). Jan 10, 2025;6(1):101118. [FREE Full text] [doi: 10.1016/j.patter.2024.101118] [Medline: 39896256]

Abbreviations

AI: artificial intelligence
AMA: American Medical Association
ARI: Automated Readability Index
BLEU: Bilingual Evaluation Understudy
CLI: Coleman-Liau Index
FKGL: Flesch-Kincaid Grade Level
FRE: Flesch Reading Ease
GFI: Gunning Fog Index
INFLESZ: Índice Flesch-Szigriszt
JBI: Joanna Briggs Institute
LIX: Läsbarhetsindex Index
LLM: large language model
LWF: Linsear Write Formula
NLP: natural language processing
PCC: Population-Concept-Context
PEM: patient education material
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: PRISMA Statement for Reporting Literature Searches in Systematic Reviews
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RGL: reading grade level
RIX: Rate Index
ROUGE: Recall-Oriented Understudy for Gisting Evaluation
SMOG: Simple Measure of Gobbledygook
SOL: Spanish Orthographic Length

Edited by S Brini; submitted 24.Nov.2025; peer-reviewed by K Chen, Z Liu; comments to author 22.Dec.2025; revised version received 27.Feb.2026; accepted 02.Mar.2026; published 07.May.2026

Please cite as:
Krenn C, Loder C, Berger N, Jeitler K, Semlitsch T, Siebenhofer A, Wilfling D
Automated Approaches of Text Simplification of Patient Education Materials: Scoping Review
J Med Internet Res 2026;28:e88365
URL: https://www.jmir.org/2026/1/e88365
doi: 10.2196/88365
PMID: