

Utility of Generative Artificial Intelligence in Orthopaedic Surgery Patient Education Materials: A Scoping Review

Review began 09/29/2026
Review ended 10/21/2026
Published 10/26/2026

© Copyright 2026

Jones et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.118623

Hugh Jones ¹, Kayden Chahal ¹, Charles Picton ², Taha Sherief Benamer ³, Ahmed Elnewishy ¹

1. Department of Trauma and Orthopaedic Surgery, Maidstone and Tunbridge Wells NHS Trust, Kent, GBR 2. Department of Otorhinolaryngology, Maidstone and Tunbridge Wells NHS Trust, Kent, GBR 3. College of Medicine, King's College London, Kent, GBR

Corresponding author: Hugh Jones, h.jones77@nhs.net

Abstract

Large language models (LLMs) are increasingly used to generate patient education materials (PEMs) and answer frequently asked questions (FAQs), though their accuracy and quality remain uncertain. This scoping review describes the accuracy, readability and quality of responses generated by LLMs to orthopaedic FAQs and PEMs, characterising the tools and prompting strategies used to assess them.

A Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist-aligned scoping review of PubMed (NLM), Ovid MEDLINE® ALL and Embase databases (up to January 22, 2025) was conducted. Eligible studies used validated tools to assess LLM-generated or modified PEMs or FAQ responses. Accuracy was assessed using qualitative ordinal scales; readability using validated indices including the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease Score (FRES); and quality using the DISCERN tool. Studies available only as abstracts, non-English publications, and studies not evaluating LLM outputs were excluded. Results were charted as reported, without pooling or formal risk-of-bias assessment.

A total of 32 of 95 studies met inclusion criteria. Accuracy, readability and quality were reported in 16, 16 and 12 studies, respectively. Inaccuracies, omissions or potentially harmful content were reported in some studies; hallucination was explicitly assessed in 2/32 studies. Prompting was classifiable in 28/32 studies: 13/28 used zero-shot prompting alone, and 15/28 used additional prompting techniques. With zero-shot prompting alone, the mean FKGL was 12 to 15. A custom ChatGPT-4.0 chat using zero-shot prompting, constraint prompting and reinforcement learning from human feedback had a lower FKGL (8.20 vs 13.99, $P<0.001$), higher FRES (60 vs 32, $P<0.001$) and improved DISCERN score (45.95 vs 38.41, $P<0.001$) than ChatGPT-4.0 prompted by zero-shot prompting alone.

Reported accuracy varied, and inaccuracies, omissions and potentially harmful content were reported. Readability exceeded recommended levels with zero-shot prompting alone. Hallucination was minimally assessed. Prompting was often basic or incompletely reported. Expert oversight remains essential; future research should use standardised evaluation methods and report prompts in full.

Categories: Other, Orthopedics, Healthcare Technology

Keywords: chatgpt, generative ai, generative artificial intelligence, google gemini, large language models, online health information, orthopaedics, orthopaedic surgery, patient education, patient education materials

Introduction And Background

Generative artificial intelligence (AI) has exponentially risen in popularity, research and public conversation over the past decade [1,2]. As these technologies become increasingly integrated into everyday life, their impact on healthcare has garnered attention. Within orthopaedics, AI research has focused on image-based fracture detection, implant identification and outcome prediction [3-5].

Large language models (LLMs) are a type of generative AI that interpret and generate human-like text [6]. Text is divided into tokens (tokenisation); during pre-training, the model learns statistical patterns from large text datasets by predicting the next token, and developers may then refine it through fine-tuning, including reinforcement learning from human feedback (RLHF) [7,8]. The effectiveness of LLMs hinges on the quality of the prompts they receive. Prompt engineering, applied by the user, is the structured design of prompts to optimise LLM output [9]. Examples include zero-shot prompting, prompting an LLM without an example, relying solely on its pre-trained knowledge; one-shot and few-shot prompting, incorporating one or several examples to guide output structure; constraint prompting, defining rules and formats to guide LLM output; chain-of-thought prompting, encouraging stepwise reasoning by instructing the LLM to articulate its reasoning steps; and more advanced strategies such as iterative refinement [9]. Iterative refinement is the stepwise improvement of LLM output through many processes, one of which includes

How to cite this article

Jones H, Chahal K, Picton C, et al. (October 26, 2026) Utility of Generative Artificial Intelligence in Orthopaedic Surgery Patient Education Materials: A Scoping Review. Cureus 18(10): e118623. DOI 10.7759/cureus.118623

repeated re-prompting [10]. These prompting strategies are distinct from retrieval-augmented generation (RAG), which supplies verifiable external sources retrieved at query time to the LLM [11]. RLHF refines the LLM response as the user provides feedback on the response to then provide a further, more-refined response [10].

LLMs have a learning cut-off date; the data on which they are trained is up to a certain date, beyond which LLMs have no knowledge of new developments or research [7,12]. For ChatGPT-3.5, this is reported as September 2021 [13,14]. For ChatGPT-4.0, this is reported as April 2023 [15]. LLMs may also hallucinate [13,16], which we define as content that is fabricated or unsupported by evidence but presented fluently and confidently. We distinguish this from factual inaccuracy and from omission, although studies will often use these terms interchangeably [12].

Traditionally, common sources of health-related information for patients lay in the form of in-person consultations with healthcare providers or via patient education materials (PEMs) [10,17]. Patients commonly turn to the internet to seek information about medical procedures and conditions [18,19]. Effective patient education is central to informed decision-making, yet studies have consistently shown that most online health content remains too complex for the average reader [10,20].

LLMs have emerged as contenders for patient education due to their potential ability to synthesise and simplify complex medical information into accessible, human-like responses to improve health literacy [1,17,21]. However, given the lack of validation of LLMs and reports of misleading information, their accuracy and potential for use in patient education remain debated [22,23]. As LLMs become further integrated into society, it is imperative to critically evaluate their accuracy, usability, and overall reliability [24,25].

The relevance of this scoping review is underscored by several timely factors: the increasing accessibility of LLMs, a lack of evidence on their clinical quality and safety, and the vital role they could play in patient education to bridge a gap between healthcare providers and patients within a highly specialised field.

This scoping review aims to evaluate the accuracy, readability and quality of responses provided by LLMs when responding to frequently asked questions (FAQs), generating PEMs and rewriting pre-existing PEMs within the field of orthopaedic surgery. We further examine the prompt engineering strategies employed across studies. A broad approach was adopted to characterise the available evidence underpinning this rapidly evolving field and to examine how research is currently being conducted, with the aim of identifying existing knowledge gaps and informing future research in this field.

Review

Methods

Search Strategy

A scoping review was conducted in line with the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist (see Appendix A) [26]. One author (HJ) searched three databases on January 22, 2025: PubMed (NLM), Ovid MEDLINE® ALL and Embase, the latter two accessed through the Ovid platform. Platforms, dates, applied limits, fields and records retrieved are summarised in Table 1. The complete search strings are detailed in Appendix B. Reference lists of similar review papers were screened by HJ and KC to identify any studies that may have been missed by our search. This review has been registered with the Open Science Framework (OSF) [27].

	PubMed (NLM)	Ovid MEDLINE® ALL	Embase
Platform	PubMed (NLM)	Ovid	Ovid
Coverage	From inception to January 22, 2025	1946 to January 22, 2025	1974 to January 22, 2025
Date searched	January 22, 2025	January 22, 2025	January 22, 2025
Limits	None	None	None
Fields	[tiab], [MeSH]	.mp.	.mp.
Records	35	30	30

TABLE 1: A summary of database searches: platforms, coverage, search dates, limits and records retrieved

Complete search strategies are provided in Appendix B.

Selection Process

Records from all three databases were imported into Covidence® systematic review software (Veritas Health Innovation, Melbourne, Australia). Two reviewers (HJ, KC) independently screened all titles and abstracts, and then independently assessed all full texts against the eligibility criteria. Disagreements were resolved by discussion between the two reviewers (HJ, KC) and a third reviewer (CP), acting as adjudicator. Inter-reviewer agreement was not calculated. The inclusion criteria were papers in which LLMs were used to generate, rewrite or modify PEMs or respond to FAQs in orthopaedics; papers in which outputs were evaluated using any validated quality, accuracy or readability assessment tool; papers that compared the output between different LLMs; and papers involving patients presenting with orthopaedic conditions managed both surgically and conservatively. The exclusion criteria were papers available only as abstracts, non-English-language papers and papers that did not assess LLMs.

As a scoping review aims to map evidence, no formal risk-of-bias tool was applied, and no study was excluded on quality [26,28]. Given the heterogeneity of the included studies, we recorded key methodological features as reported: prompt reproducibility, number and expertise of assessors, assessor blinding, repeated-query testing, use of validated outcome measures, and sample size. These features are present in the Results.

Data Collection and Outcomes

Data were collected by all authors, each responsible for a subset of papers, and verified against the source papers by one author (HJ) before collation in the final data sheet. The primary outcomes assessed were the accuracy, readability and quality of LLM-generated output. Accuracy was assessed through varying qualitative ordinal scales: Likert scales, scales from excellent-unsatisfactory or appropriate-unreliable and percentages of overall accuracy. Readability was assessed through the Flesch-Kincaid Grade Level (FKGL) formula and the Flesch Reading Ease Score (FRES) formula. Quality was assessed through the DISCERN tool. Secondary outcomes included formal readability scores, descriptions of accuracy and comparison of LLM responses against expert opinion. Prompting strategy was classified from the prompt wording and method of entry reported in each paper, irrespective of the authors' own terminology, using the categories zero-shot (the question entered without added instruction), constraint (an added instruction restricting the format, length, reading level or content of the response), role (an assigned persona or context) and iterative (follow-up prompts in the same conversation) [9]. Studies were classified only where the paper showed what was entered; studies that did not report sufficient detail (for example, the question alone with no statement that nothing further was entered) were classified as unclear.

Results

A total of 95 papers were identified in the literature search (Figure 1). Of these, 42 were excluded owing to duplication. Fifty-three papers were screened by title and abstract. Twenty papers were excluded owing to irrelevance or being available only in abstract form. Thirty-three papers underwent full-text review. One paper was excluded as it evaluated MedCAT, a machine learning algorithm, rather than assessing an LLM. This left 32 studies to be included in this review; their summary can be seen in Table 2.

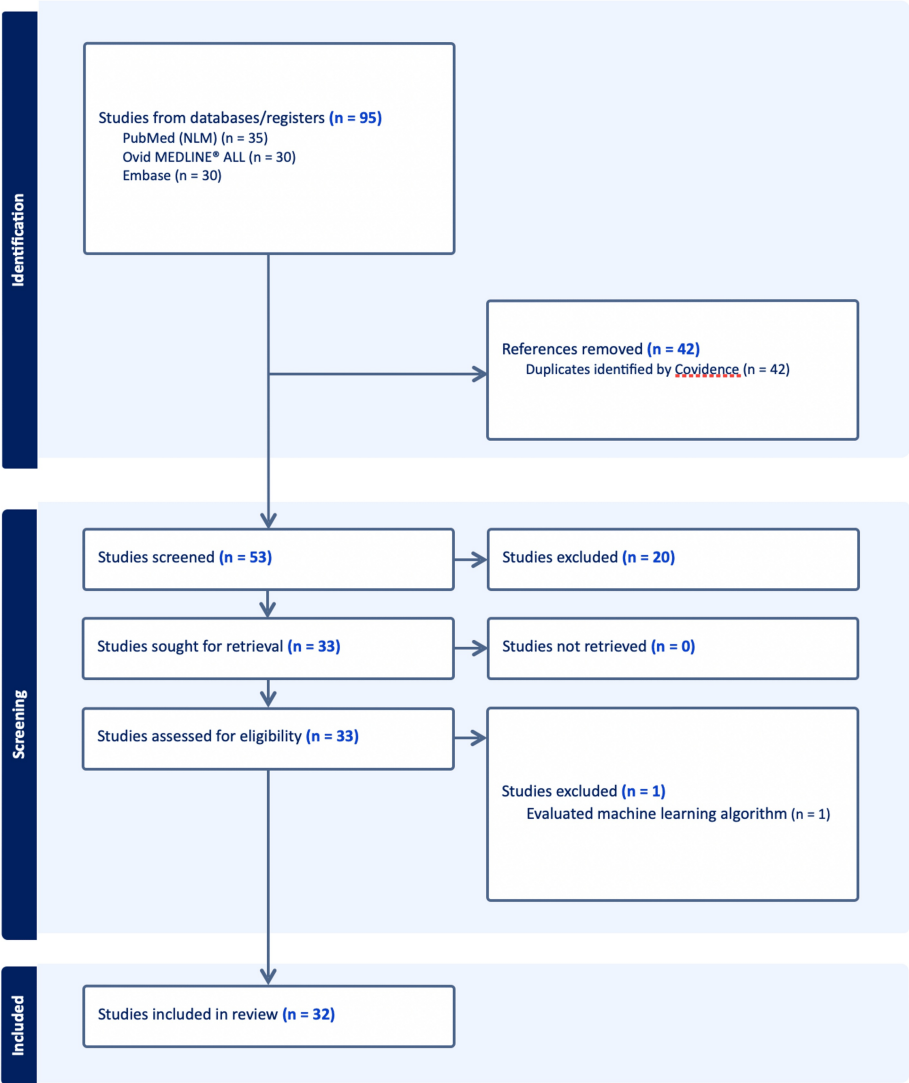


FIGURE 1: A PRISMA-ScR flowchart of study selection

PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews

Author, year, country	LLM(s) assessed, date accessed	Prompt source, area of orthopaedics, assessors, n and assessor characteristics	Prompt strategy	Sample size, n	Assessor blinding	Repeated query testing	Validated scale used	Accuracy (Acc), readability (Read) and quality (Qual) scores
Schwartzmann, 2025, USA [1]	ChatGPT-3.5, November 2023	Expert opinion. TKA, THA. No. of assessors not reported. 'Fellowship-trained adult reconstructive surgeons'.	Zero-shot prompting, Constraint prompting.	33 FAQs	No	No	Qualitative ordinal scale (Likert)	Acc: 4.4±1.4/6, Read, Qual: not assessed.
Anastasio, 2023, USA [2]	ChatGPT (unspecified iteration), March 2023	Expert opinion. Foot, Ankle. 2 assessors. 'Orthopaedic research team'.	Unclear (question only reported)	5 FAQs	Not reported	No	DISCERN AI response metric (AIRM)	Acc, Read: not assessed. Qual: Mean DISCERN: 47.75 (range 18-75). Mean AIRM: 18.33 (range 13-25).
Smith, 2024, USA [6]	ChatGPT (unspecified iteration), 2024*	Expert opinion. ACL reconstruction, Microdiscectomy, THA, Rotator cuff repair, Carpal tunnel release, Ankle fracture ORIF. No. of assessors not reported. 'Panel of experienced orthopaedic surgeons'.	Role prompting.	60 FAQs (10 FAQs per procedure)	Not reported	No	Qualitative ordinal scale (Likert)	Acc: Likert scale: Mean score 2.43 ± 0.59, 2.25±0.5 (ACL repair), 3±0 (microdiscectomy), 2.25±0.5 (THA), 2.20±0.837 (Carpal tunnel release), 2.25±0.5 (rotator cuff repair), 3±0 (Ankle

								fracture ORIF). Read, Qual: Not assessed.
White, 2024, USA [8]	ChatGPT-3.5, December 2023	Website of institution or orthopaedic society, expert opinion. TSA. 3 assessors. 2 'board-certified and fellowship-trained shoulder and elbow surgeons', 1 x shoulder and elbow fellow.	Constraint prompting.	15 FAQs	Not reported	No	Alphabetical grading system (A-F)	Acc: Overall: B- (adequate response with only minor clarification needed). Read, Qual: Not assessed.
Fahy, 2024, Germany [10]	ChatGPT-4.0, May 2024	Google search, "People Also Asked" section. HTO, knee osteotomy. Quality - 3 assessors, 'orthopaedic surgeons specialized in the field of knee surgery'. Readability - no assessors, assessment automated using the Readability Studio Professional Edition Programme.	Native: Zero-shot prompting. 'Knee-guide' Zero-shot prompting, Constraint prompting, RLHF.	50 FAQs (25 FAQs per model)	Quality - Yes. Readability - N/A (Automatic)	No	DISCERN, Readability - Readability Studio Professional Edition Programme (including FRES, FKGL)	Acc: Not assessed. Read: FRES; Mean FRES: 32 (ChatGPT4.0), 60 ('The Knee Guide'), FKGL; Mean FKGL: 13.99 (ChatGPT4.0), 8.20 ('The Knee Guide') (p ≤ 0.001). Qual: DISCERN; Mean DISCERN: 38.41 (ChatGPT4.0), 45.95 ('The Knee Guide') (p<0.001).
Zhang, 2024, Singapore [13]	ChatGPT-3.5, 2024*	Google search, expert opinion. TKR. 3 Assessors. 'Consultant orthopaedic surgeons'.	Zero-shot prompting.	50 FAQs	Not reported	No	Qualitative ordinal scale (Likert)	Acc: 4.6/5, 44/50 (88% accurate). Read, Qual: not assessed
Yeramosu, 2024, USA [14]	ChatGPT-3.5, August 2023	Website of institution or orthopaedic society, Expert opinion. TSA. 2 assessors. 'Fellowship-trained orthopaedic surgeons'.	Constraint prompting.	10 FAQs	Not reported	No	DISCERN, FRES, FKGL	Acc: not assessed. Read: Mean FRES: 25.59±12.64. Mean FKGL: 15.35±3.08, Qual: Mean DISCERN: 46.60.
Wright, 2024, USA [15]	ChatGPT (unspecified iteration), March 2023	Google search. TKA, THA. 5 assessors. 'Orthopaedic Surgery Residents'.	Iterative prompting.	60 FAQs (30 FAQs per procedure)	Not reported	Yes (each question twice)	FKGL, Qualitative ordinal scale (Likert), Dichotomous nominal scale (Yes/No)	Acc: Mean Likert: Accuracy: 4.26±0.07, Comprehensiveness: 3.79±0.1, Dichotomous nominal scale; 59.2% (71/120) acceptable, Read: Mean FKGL: 11.65 (95% CI 11.19-12.10). Qual: not assessed.
Lack, 2024, USA [16]	ChatGPT-3.5, November 2023	Website of institution or orthopaedic society. rTSA. 2 assessors. 'Orthopaedic surgeons'.	Zero-shot prompting.	10 FAQs	Not reported	No	DISCERN, FKGL	Acc: not assessed. Read: Mean FKGL: 14.35 (range 9.0-19.3). Qual: Mean DISCERN; 44 (range 38-51).
Tharakan, 2024, USA [29]	ChatGPT-3.0, 2024*	Google search. TSA, TEA. 2 assessors. 'Authors'.	Zero-shot prompting.	20 FAQs recorded (10 FAQs per LLM)	Not reported	No	Modified Rothwell Classification System	Acc: Not assessed. Most common source of information, 'Academic references' 65% (26/40); ChatGPT-3.0 citing 80% academic references; most common academic resource: AAOS 30% (6/20). Read, Qual: Not assessed.
Stroop, 2024, Germany [30]	ChatGPT (unspecified iteration), February 2023	Expert opinion. LDH. 24 assessors. 'Spinal surgeons'.	Unclear ('free-text, user-generated questions')	139 FAQs	N/A (assessors rated responses)	No	Qualitative ordinal scale (Percentage)	Acc: 97% as 'completely understandable', 52% as 'completely correct', 55% as 'medically complete and comprehensive', 1.3% statements incorrect. Read: Not reported. Qual: 86% satisfactory.
Li, 2023, China [31]	ChatGPT-4.0, 2023*	Clinical guidelines. Knee OA. 2 assessors. Grade and Specialty not reported.	Zero-shot prompting, Iterative prompting.	15 FAQs	Not reported	No	Qualitative ordinal scale (Likert)	Acc: AAOS, mean accuracy (5 points) 4.3±1.6, mean completeness (3 points) 2.8 ±0.6. Chinese orthopaedic association guidelines, mean accuracy (5-points) 4.0±0.6, 4.5±0.6, and 4.5±0.7, mean completeness (3-points) 2.9±0.3, 2.3±0.9, and 2.1±1.0 for A-, B- and C-levels respectively. Read, Qual: Not assessed.
Subramaniam, 2024, USA [17]	ChatGPT-3.5, 2024*	Expert opinion. Cervical spine surgery. 2 assessors. 'Spinal surgeons'.	Zero-shot prompting, Role prompting.	9 FAQs	Not reported	No	FKGL, Qualitative ordinal scale (Likert)	Acc: Mean Likert - 7.11/10. Read: FKGL - prompted mean 11.2 (original mean 13.5), Qual: not assessed.
Gupta, 2024,	ChatGPT (unspecified	Google search. TSA, rTSA. 2 assessors.	Unclear (question	60 FAQs (30 FAQs			DISCERN,	Acc: Not assessed, Read: Mean FRES - 15.15, 11.14, 10.95 (fact, policy, value for aTSA), 48.02, 12.51, 17.22 (fact, policy, value for rTSA), Mean FKGL - 17.48, 17.72,

USA [18]	iteration), 2024*	'Authors'.	only reported)	per procedure)	Not reported	No	FRES, FKGL	17.96 (fact, policy, value for aTSA), 8.10, 17.27, 16.56 (fact, policy, value for rTSA), Qual: Mean DISCERN - 57, 61, 58 (fact, policy, value),
Yüce, 2024, Turkey [19]	ChatGPT-4.0, August 2023	Google search, Expert opinion. THA. 2 assessors. 'Consultant orthopaedic surgeons'.	Zero-shot prompting, Iterative prompting	10 FAQs	Not reported	No	Hip information scoring system (HISS)	Acc, Read: Not assessed. Qual: mean 11.57 (range 10-14.5). On identifying missing criteria for THA, mean 12.05 (range 5-21).
Kirchner, 2023, USA [20]	ChatGPT (unspecified iteration), 2023*	Website of institution or orthopaedic society. LDH, Scoliosis, Spinal stenosis, TKA, THA. 4 assessors. 'Authors'.	Constraint prompting.	20 PEMs	Not reported	Yes (4 sequential conversions)	FKGL, FRES	Acc: no factual errors. Read: Median FKGL: 6.3 (from 12.0, for TKA), 6.1 (from 11.6, for THA), 5.0 (from 9.5, for LDH), 5.6 (from 12.6, for scoliosis), 6.9 (from 10.9, for spinal stenosis). Median FRES: 76.0 (from 44.3, for TKA), 77.1 (from 54.1, for THA), 80.4 (from 59.7, for LDH), 79 (from 45.1, for scoliosis), 75.5 (from 51.8, for spinal stenosis). Qual: not assessed.
Kasapovic, 2024, Germany [21]	ChatGPT-3.5-Turbo, 2024*	Expert opinion. 'Common Trauma and Orthopaedic surgeries'. No. of assessors not reported. Grade and speciality not reported.	Unclear (entry not described)	20 FAQs	Not reported	No	mDISCERN	Acc, Read: Not assessed. Qual: Mean mDISCERN, 2.4/5 (range 1.3-3.4), 0% (0/20) responses were rated as "high - minimal deficiencies", 45% (9/20) as "moderate - potentially important but not significant deficiencies", 55% (11/20) as "low - significant deficiencies"
Sparks, 2024, USA [22]	ChatGPT-3.5, February, March 2023	Website of institution or orthopaedic society. 'Common orthopaedic conditions'. 2 assessors. 1 'sports medicine fellowship-trained attending orthopaedic surgeon', 1 'orthopaedic surgery sports medicine fellow'.	Zero-shot prompting.	40 prompts	Not reported	No	Qualitative ordinal scale (Accuracy - 100%, 75%-99%, 50%-74%, <50%)	Acc: 65% (26/40) responses were 75%-99% accurate; 35% (14/40) responses were 50%-74% accurate. Read, Qual: Not assessed
Artioli, 2024, Italy [23]	ChatGPT-3.5, March 2024	Website of institution or orthopaedic society, expert opinion. TAA. 2 assessors. 'Orthopaedic surgeons'.	Zero-shot prompting.	10 FAQs	Not reported	No	WordCalc software package (including FKGL, Gunning FOG, DISCERN)	Acc: 30% (3/10) 'Excellent responses not requiring clarification', 70% (7/10), 'Satisfactory response requiring minimal clarification'. Read: FKGL 12, Gunning Fog 18.3. Qual: DISCERN 51.1
Marder, 2024, USA [24]	ChatGPT-3.5, ChatGPT-4.0, April 2024	Google search. 'Common upper- and lower-extremity orthopaedic conditions'. No assessors (automated through Python Library Textstat).	Question 1: Zero-shot prompting, Question 2: One-shot prompting, Constraint.	100 FAQs	N/A (automated assessment)	No	Readability scores calculated through Python library Textstat (including FKGL, FRES)	Acc: Not assessed. Read: (prompt two (prompt one)) Upper extremity results: FRES: 46.37 (60.37, ChatGPT-3.5), 80.33 (79.63, ChatGPT-4.0), FKGL: 10.80 (9.09, ChatGPT-3.5), 5.98 (6.94, ChatGPT-4.0). Lower extremity results: FRES: 49.46 (64.00, ChatGPT-3.5), 83.08 (81.49, ChatGPT-4.0), FKGL: 10.42 (8.48, ChatGPT-3.5), 5.46 (6.69, ChatGPT-4.0). Qual: not assessed.
Mika, 2023, USA [25]	ChatGPT (unspecified iteration), February 2023	Website of institution or orthopaedic society. THA. 2 assessors. 'Authors'.	Zero-shot prompting.	10 FAQs	Not reported	No	Qualitative ordinal scale (Excellent-Unsatisfactory)	Acc: 20% (2/10) 'Excellent response not requiring clarification', 70% (7/10) 'Satisfactory response requiring minimal clarification', 10% (1/10) 'Unsatisfactory requiring substantial clarification'. Read, Qual: not assessed.
Crook, 2023, USA [32]	ChatGPT 3.0, April 2023	Website of institution or orthopaedic society. Carpal tunnel, cubital tunnel and trigger finger release and distal radius fracture fixation. No. of assessors not reported. Grade and speciality not reported.	Zero-shot prompting.	66 FAQs	Not reported	No	DISCERN, FRES, FKGL	Acc: not assessed. Read: Mean FRES: 34 (ChatGPT 3.0), 60 (ASSH HandCare). Mean FKGL: 15 (ChatGPT 3.0), 9 (ASSH HandCare). Qual: Mean DISCERN: 58 (ChatGPT 3.0), 71 (ASSH HandCare).
							FKGL,	Acc: 46% (23/50) responses were excellent ('no clarification'), 44% (22/50)

Amaral, 2024, USA [33]	ChatGPT-3.5, January 2024	Website of institution or orthopaedic society, expert opinion. Paediatric in-toeing. 2 assessors. 'Fellowship-trained paediatric orthopaedic surgeons'.	Zero-shot prompting.	25 FAQs	Not reported	Yes (each question twice)	Qualitative ordinal scale (excellent-unsatisfactory)	satisfactory ('minimal clarification'), 8% (4/50) satisfactory ('moderate clarification'), 2% (1/50) unsatisfactory ('substantial clarification'). Read: FKGL average, 12.7±1.4. Qual: not assessed.
Nian, 2025, USA [34]	ChatGPT-3.5, ChatGPT-4.0, Google Gemini, July 2024	Clinical guidelines, expert opinion. Paediatric supracondylar humerus and diaphyseal femur fractures. 2 assessors. 'Senior-author fellowship-trained attending paediatric orthopaedic surgeons'.	Zero-shot prompting.	39 FAQs (13 FAQs per LLM)	Yes	No	FKGL, FRES, Gunning-Fog index	Acc: ChatGPT-4.0, -3.5 and Gemini; 11/13 (84.6%), 9/13 (69.2%), 11/13 (84.6%). Read: ChatGPT 4.0; Mean FKGL: 14.5±1.4, Mean FRES: 28.4±9.3. ChatGPT 3.5; Mean FKGL: 14.9±2.2, Mean FRES: 22.3±12.0. Google Gemini; Mean FKGL: 12.4±1.8, Mean FRES: 38.5±8.8. Qual: not assessed.
Giorgino, 2024, Italy [35]	ChatGPT-3.5, Google Bard 2024*	Google search "People Also Ask" section, expert opinion. ACL, Flat foot. 4 assessors. 'Orthopaedic surgeons'.	Zero-shot prompting.	10 FAQs	Yes	No	Global quality score	Acc, Read: not assessed. Qual: ChatGPT 3.5; Mean: 4.1±0.7 (sports medicine), mean: 3.8±0.83 (paediatric orthopaedics). Google Bard; Mean: 4±0.78 (sports medicine), mean: 3.5±1 (paediatric orthopaedics)
Miskiewicz, 2024, USA [36]	ChatGPT-4.0, 2024*	Website of institution or orthopaedic society. Rotator cuff injuries. 1 assessor. 'PGY-5 orthopaedic surgery resident'.	Zero-shot prompting, Constraint prompting, Role prompting	25 PEMs	Not reported	No	Readability score from the Readability website (including FKGL)	Acc: not assessed. Read: Mean reading grade level for re-written PEMs 5.67±1.77 (original PEM, 10.02±1.86). Mean word count for re-written PEMs 312.2±65.73 (original PEM, 755.0±379.76 words). Qual: not assessed.
Miskiewicz, 2024, USA [37]	ChatGPT-4.0, 2024*	Website of institution or orthopaedic society. Lateral Epicondylitis. 1 assessor. 'PGY-5 orthopaedic surgery resident'.	Zero-shot prompting, Constraint prompting, Role prompting	25 PEMs	Not reported	No	Readability score from the Readability website (including FKGL)	Acc: not assessed. Read: Average reading grade level and words for re-written PEMs: 6.12±0.97 (original PEM 9.81±1.76), 253.96±100.76 (original PEM 600.68±409.44), respectively. Qual: not assessed.
Vallurupalli, 2024, USA [38]	ChatGPT-3.5, 2024*	Website of institution or orthopaedic society. Hand surgery. No assessors (automated through Readability Scoring System v2.0).	Constraint prompting.	18 PEMs	N/A (automated assessment)	No	'Readability Scoring System 2.0' (including FKGL)	Acc: not assessed. Read: From the 18 excerpts; mean reading grade level, original: 10.44 (simplified: 7.28). From ChatGPT-3.5; mean reading grade level, original: 10.22 (simplified: 7.06). Qual: not assessed.
Andalib, 2025, USA [39]	ChatGPT-3.5, ChatGPT-4.0, Claude 2, Llama 2 July, August 2023	Website of institution or orthopaedic society. Common orthopaedic conditions. No assessors (automated through Textstat, NumPy Python Packages).	Constraint prompting.	48 PEMs	N/A (automated through Textstat, NumPy Python Packages)	No	Textstat, NumPy Python Packages (including FKGL, FRES)	Acc: not assessed. Read: ChatGPT 4.0; Mean FKGL: 6.76 (Original PEM: 10.84 ± 1.49) (p<0.001), mean FRES: 76.80 (Original PEM: 53.40±8.00) (p<0.001). ChatGPT 3.5; Mean FKGL: 7.69 (Original PEM: 10.84 ± 1.49) (p<0.001), mean FRES: 69.40 (Original PEM: 53.40±8.00) (p<0.001). Qual: not assessed.
Kienzle, 2024, Germany [40]	ChatGPT-4.0, June 2023	Expert opinion. TKA. 3 assessors. 1 'senior resident orthopaedic surgeon', 2 'senior orthopaedic specialists'.	Zero-shot prompting.	50 FAQs	Yes	No	DISCERN	Acc: not assessed. Qual: Mean DISCERN, for the following questions: 3.7±0.6 (diagnosis/non-surgical treatment), 4.0±0.0 (surgery-related), 3.7±0.6 (daily-life related), 3.3±0.6 (complications-related). Read: not assessed.
Hu, 2024, Germany [41]	ChatGPT-4.0, August 2023	Website of institution or orthopaedic society, expert opinion. PJI post-TKA and THA. 3 assessors, 1 'senior infectiologist specialist', 1 'senior orthopaedic surgeon', 1 'resident infectiologist'.	Zero-shot prompting.	12 FAQs	Not reported	No	Qualitative ordinal scale (excellent-unsatisfactory)	Acc: 8.3% (1/12) responses were excellent ('no clarification'), 58% (7/12) satisfactory ('minimal clarification'), 33% (4/12) satisfactory ('moderate clarification'), 0% (0/12) unsatisfactory ('substantial clarification'). Read, Qual: not assessed.
Dubin, 2024,	ChatGPT-3.5 March,	Institutional source. THA. 3 assessors.	Zero-shot	60 FAQs	Not reported	No	Qualitative ordinal scale	Acc: 93.3% (56/60) responses 'appropriate', 6.7% (4/60) graded

TABLE 2: Study level characteristics, methodological features and results of the included studies

* Where the month of LLM access was left unspecified, the year of publication has been used.

Prompting strategy was assigned from reported wording and entry method. Studies without sufficient detail are classed as unclear.

LLM: large language model; FAQs: frequently asked questions; TKA: total knee arthroplasty; THA: total hip arthroplasty; ACL: anterior cruciate ligament; ORIF: open reduction and internal fixation; aTSA: anatomic total shoulder arthroplasty; RLHF: reinforcement learning from human feedback; rTSA: reverse total shoulder arthroplasty; LDH: lumbar disc herniation; AAOS: American Academy of Orthopaedic Surgeons; HISS: Hip Information Scoring System; PEMs: patient education materials; AANS: American Association of Neurological Surgeons; AAHKS: American Association of Hip and Knee Surgeons; ASSH: American Society for Surgery of the Hand; FKGL: Flesch-Kincaid Grade Level; FRES: Flesch Reading Ease Score; Acc: accuracy; Read: readability; Qual: quality

Accuracy was assessed in 16/32 studies, readability in 16/32 and quality in 12/32 (Table 2). Across the 32 included studies, accuracy and quality were reported using a range of scales. On Likert scales, a mean accuracy score of 4.6/5 was reported for total knee replacement (TKR) FAQs [13], 4.4±1.4 for pre-operative total knee arthroplasty (TKA) and total hip arthroplasty (THA) FAQs [1], and a combined average of 7.11/10 for cervical spine surgery FAQs [17]. DISCERN scores ranged from 38.41 [10] to 61 [18]. Inaccurate responses were reported in some studies; 12% (6/50) of TKR responses contained factual errors [13] and 20% (5/25) of paediatric in-toeing responses inaccurately emphasised early intervention [33]. Kasapovic et al. found 35% (7/20) of generated PEMs 'not useful' and 20% (4/20) potentially 'dangerous to patients' [21].

Nian et al. compared three LLMs; ChatGPT-4.0 and Google Gemini each achieved 84.6% (11/13) accuracy versus 69.2% (9/13) for ChatGPT-3.5 [34]. Giorgino et al. compared ChatGPT-3.5 with Google Gemini (formerly Google Bard) using the Global Quality Score across both sports medicine (4.1±0.7 vs 4.0±0.78) and paediatric orthopaedics (3.8±0.83 vs 3.5±1.0) [35].

Hallucination was addressed in 2/32 studies, which quantified fabricated references: 64.2% of 109 cited sources were not traceable to an existing publication [1], and 37% (10/27) of cited references were fabricated or non-existent [40]. Hallucination was mentioned in 5/32 without assessing it as an outcome [13,16,19,29,35].

In studies using zero-shot prompting alone, mean FKGL values included 15 [32], 12.7±1.4 [33] and 14.35 [16]. In a comparative study, Nian et al. found that mean FKGL and FRES scores were 14/5±1.4 and 28.4±9.3 for ChatGPT-4.0, 14.9±2.2 and 22.3±12.0 for ChatGPT-3.5 and 12.4±1.8 and 38.5±8.8 for Google Gemini [34]. For studies that prompted LLMs to re-write existing PEMs, ChatGPT-4.0 reduced FKGL from 10.02±1.86 to 5.67±1.77 (P<0.001) for rotator cuff PEMs [36] and from 9.81±1.76 to 6.12±0.97 (P<0.001) for lateral epicondylitis PEMs [37], with significant reductions in word count in both studies (P<0.001). ChatGPT-3.5 reduced mean reading grade level from 10.44 to 7.28 for hand surgery PEMs [38]. Fahy et al. compared 'The Knee Guide', a separate ChatGPT-4.0 chat prompted with zero-shot prompting, constraint prompting and RLHF, with ChatGPT-4.0: FKGL (8.20 vs 13.99, P<0.001), FRES (60 vs 32, P<0.001), DISCERN score (45.95 vs 38.41, P<0.001) [10].

The September 2021 learning cut-off date for ChatGPT-3.5 was attributed to deficiencies in its responses due to its limited ability to incorporate up-to-date information [8,22,35].

Table 2 shows the source of the prompts. Of these, 7/32 studies used expert opinion alone, 5/32 used a Google search alone, 3/32 used a Google search combined with expert opinion and 1/32 used FAQs from their institutional source. A total of 9/32 papers used FAQs or PEMs from a website of an established institution or orthopaedic society, and 5/32 papers used a website of an established institution or orthopaedic society and expert opinion. About 1/32 papers used FAQs from clinical guidelines, and 1/32 papers used FAQs from clinical guidelines and expert opinion (Table 2).

Four of the 32 studies did not report sufficient prompt detail and were classified as unclear [2,18,21,30]. Zero-shot prompting alone was used in 13/28 studies. A total of 15/28 studies used an additional prompting technique; 5/28 used constraint prompting alone, 1/28 role prompting alone, 1/28 zero-shot with constraint prompting, 1/28 zero-shot with constraint prompting and RLHF, 1/28 zero-shot with role prompting, 2/28 zero-shot with constraint and role prompting and 3/28 zero-shot with iterative prompting (Table 2). One of the 28 studies used zero-shot prompting alone for one prompt and zero-shot prompting with constraint prompting for a second prompt [24].

Blinding of assessors was not reported in 23/32 studies. Assessors were blinded in 4/32 studies [10,34,35,40]; 1/32 reported not blinding assessors [1]. It was not applicable in 5/32 studies, as a result of automation or non-assessor-rated outcomes [10,24,30,38,39]. In one study, assessors were blinded for

quality assessment, but the readability assessment was automated [10]. The same prompt was entered more than once in 3/32 studies [15,20,33]; the remaining studies did not report repeated entry (Table 2).

Discussion

Inaccurate, incomplete or potentially harmful content was reported in some studies. Hallucinations were explicitly assessed in only 2/32 studies [1,40]. Readability commonly exceeded recommended levels when only zero-shot prompting was used. Prompting strategies were often basic or incompletely reported.

Accuracy and Hallucinations

Orthopaedics is a complex speciality, one that takes years to learn, so it is a notable idea that an LLM, without formal medical training, can produce accurate responses in a seemingly instantaneous manner [13]. However, inaccuracies were reported in some studies. ChatGPT-3.5 reported that a 'femoral fracture may be a fracture of the femur or pelvis' [22], and ChatGPT-4.0 omitted diagnostic tests required for confirming periprosthetic joint infection [41]. LLMs do not 'reason' or 'think' in a similar way to a trained clinician; they generate responses based on recognised patterns and structures within the text that they are trained on [13]. When this is paired with the learning cut-off dates that exist for the studied LLMs, responses may continue to be inaccurate [8,22,35]. Hallucinations, as we have defined earlier, were assessed in 2/32 studies, both of which noted the inclusion of fabricated references [1,40]. It was a term that was used inconsistently and so, although remaining studies did report inaccuracies or omissions, this review is unable to establish how often LLMs hallucinate in orthopaedic patient education.

Clinical Safety

Clinical safety was not directly evaluated in this review. However, potentially harmful or dangerous responses were reported in a few studies [16,21]. Erroneous information could influence patients' decisions around their health. Expert supervision remains essential whenever LLMs are used in a healthcare setting.

Readability

When prompted with zero-shot prompting alone, responses were commonly generated that require a college-level reading ability to comprehend [18,32,33], above the regularly recommended sixth- to eighth-grade reading level [20]. However, when zero-shot prompting, constraint prompting and RLHF were used, ChatGPT-4.0 reduced the PEM reading level from a college level to an eighth-grade reading level [10]. LLMs could provide a tool that tailors to the health literacy of patients, enhancing shared decision-making between patients and healthcare providers, when delivered at an appropriate reading level [40]. One study did, however, report that simplifications in responses led to generalisations, with failures to include relevant information noted [17].

Prompting Strategies

Prompting could be classified in 28/32 studies; 13/28 used zero-shot prompting alone and 15/28 used an additional prompting technique, of which constraint prompting was most often used (10/28). Zero-shot prompting is the quickest form of prompting. Its use was justified in two studies as best reflecting a patient's interaction with LLMs [40,41]. It is rather basic and leads to general outputs with variable quality and limited guidance for complex tasks [9]. In 4/32 studies, the prompt was not reported in enough detail to classify. This limits prompt reproducibility, and we recommend that prompts and the method of entry be consistently reported in full.

In the included studies of this review, prompting strategies were largely basic. This may reflect the wider landscape, in which Kunze et al. found that only 6.1% of studies evaluated different LLM prompting strategies to obtain desired outputs [43]. No paper used chain-of-thought prompting or RAG. Recent research has pointed towards the combination of RAG and LLMs in assisting in planning, decision making and complex tasks, as supposed to simple information retrieval [11]. This notion is reiterated by Tripathi et al., who demonstrated that prompt specificity, contextualisation, chain-of-thought prompting and iterative refinement substantially influence LLM output quality [44].

Orthopaedic subspecialties often rely on well-established clinical rules. Only 2/32 studies sourced prompts from clinical guidelines [31,34], and 15/32 studies sourced prompts from a website of an institutional source (Table 2). Prompts could be designed to align LLM-generated outputs with the latest clinical guidelines and evidence-based practices to ensure scientific validity [9].

Comparisons Between LLMs

Few studies compared LLM output directly [24,34,35,39]. Within these studies, ChatGPT-4.0 had higher reported accuracy than ChatGPT-3.5 in one [34] and lower FKGL in three [24,34,39]. Giorgino et al. found no

significant difference in Global Quality Score between ChatGPT-3.5 and Google Gemini (formerly Google Bard) [35]. These are indirect and heterogeneous studies with different methodological approaches and, therefore, a model-version comparison cannot be established in this review.

The use of LLMs within orthopaedics is novel, and its full potential is currently unknown. To become integrated into clinical practice, LLM output will need to be reliably accurate and safe. Research that continues to use basic prompts may underestimate what LLMs can achieve.

Limitations

We searched three databases, PubMed, Ovid MEDLINE® ALL and Embase (via Ovid), with the search conducted by one author. Given the novelty of this research topic, only searching these databases could have restricted our view of the research into this topic. A formal measure of inter-reviewer agreement was not calculated in the screening stage, limiting the quantitative assessment of the reliability of study selection. Our broad inclusion criteria have meant that direct comparability between studies has been limited; however, this is consistent with the scoping review nature of this study. In retrospect, a more extensive exploration of grey literature, the use of citation tracking techniques and the inclusion of non-English studies would have increased our evidence base and reduced the risk of publication bias. Since the search cutoff date of January 22, 2025, LLMs have progressed considerably, with newer model iterations likely to demonstrate improved accuracy, readability and quality beyond those captured in the included studies. Assessor blinding was not reported in 23/32 studies, and repeated entry of prompts was reported in 3/32 studies, so assessor bias and variability between responses could not be evaluated. Prompting strategy was classified from the reported wording and method of entry and so may not reflect the authors' intended approach.

Conclusions

LLMs can generate and rewrite orthopaedic PEMs and offer an easily accessible source of information, so their use in orthopaedic patient education is likely to grow. Across 32 heterogeneous studies, reported accuracy varied; inaccuracies, omissions of information and potentially harmful content were reported, and hallucinations were minimally assessed. With zero-shot prompting alone, readability exceeded recommended average reading levels. Lower reading grade levels were reported where LLMs were prompted to simplify existing PEMs. Prompting techniques were basic and sometimes incompletely reported. In any context where LLM outputs are used, expert oversight and validation remain essential. Future research should use standardised, transparent evaluation of accuracy, hallucination, readability and report prompts in full.

Appendices

Appendix A

Section	Item	PRISMA-ScR checklist item	Reported on page
Title			
Title	1	Identify the report as a scoping review.	1
Abstract			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	1
Introduction			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	2
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	2
Methods			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	3
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	3

Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	3
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	3
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	3
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	3
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	3
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	N/A
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	3
Results			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	4
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	04-Dec
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	N/A
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	04-Dec
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	05-Nov
Discussion			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	Dec-14
Limitations	20	Discuss the limitations of the scoping review process.	14
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	14
Funding			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	15

TABLE 3: PRISMA-ScR checklist

JB: Joanna Briggs Institute; PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews

* Where sources of evidence (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and websites.

† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research,

expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with information sources (see first footnote).

‡ The frameworks by Arksey and O'Malley [6] and Levac and colleagues [7] and the JBI guidance [4,5] refer to the process of data extraction in a scoping review as data charting.

§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of 'risk of bias' (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy documents).

Appendix B

PubMed (NLM)	From inception to January 22, 2025					
Search number	Query	Sort By	Filters	Search Details	Results	Time
7	#1 AND #5 AND #6	-	-	("Artificial Intelligence"[MeSH Terms] OR "artificial intel"[Title/Abstract] OR "language model"[Title/Abstract] OR "llm"[Title/Abstract] OR "generative ai"[Title/Abstract] OR "generative artificial intel"[Title/Abstract]) AND ("Patient Education"[Title/Abstract] OR "patient educ"[Title/Abstract] OR "patient train"[Title/Abstract] OR "patient involv"[Title/Abstract] OR "patient quest"[Title/Abstract] OR "patient learn"[Title/Abstract] OR "patient instruction"[Title/Abstract] OR "patient communicat"[Title/Abstract]) AND ("orthopaedic surgery"[Title/Abstract] OR "orthopedic surgery"[Title/Abstract] OR "orthopaedics"[Title/Abstract] OR "orthopedics"[Title/Abstract] OR "bone surgery"[Title/Abstract] OR "joint surgery"[Title/Abstract] OR "musculoskeletal surgery"[Title/Abstract] OR "spinal surgery"[Title/Abstract] OR "arthroplasty"[Title/Abstract] OR "fracture fixation"[Title/Abstract] OR "orthopedic procedure"[Title/Abstract] OR "orthopedic procedure"[Title/Abstract])	35	11:08:16
6	("orthopaedic surgery"[tiab] OR "orthopedic surgery"[tiab] OR "orthopaedics"[tiab] OR "orthopedics"[tiab] OR "bone surgery"[tiab] OR "joint surgery"[tiab] OR "musculoskeletal surgery"[tiab] OR "spinal surgery"[tiab] OR "arthroplasty"[tiab] OR "fracture fixation"[tiab] OR "orthopedic procedure"[tiab] OR "orthopedic procedure"[tiab])	-	-	orthopaedic surgery[Title/Abstract] OR "orthopedic surgery"[Title/Abstract] OR "orthopaedics"[Title/Abstract] OR "orthopedics"[Title/Abstract] OR "bone surgery"[Title/Abstract] OR "joint surgery"[Title/Abstract] OR "musculoskeletal surgery"[Title/Abstract] OR "spinal surgery"[Title/Abstract] OR "arthroplasty"[Title/Abstract] OR "fracture fixation"[Title/Abstract] OR "orthopedic procedure"[Title/Abstract] OR "orthopedic procedure"[Title/Abstract]	149956	11:07:42
5	("Patient Education"[tiab] OR "patient educ"[tiab] OR "patient train"[tiab] OR "patient involv"[tiab] OR "patient quest"[tiab] OR "patient learn"[tiab] OR "patient instruction"[tiab] OR "patient communicat"[tiab])	-	-	Patient Education[Title/Abstract] OR "patient educ"[Title/Abstract] OR "patient train"[Title/Abstract] OR "patient involv"[Title/Abstract] OR "patient quest"[Title/Abstract] OR "patient learn"[Title/Abstract] OR "patient instruction"[Title/Abstract] OR "patient communicat"[Title/Abstract]	43295	11:06:54
	("Artificial Intelligence"[MeSH] OR "artificial intel"[tiab] OR			Artificial Intelligence[MeSH Terms] OR "artificial intel"		

1	"language model"[tiab] OR llm[tiab] OR "generative ai"[tiab] OR "generative artificial intel"[tiab])	-	-	[Title/Abstract] OR "language model"[Title/Abstract] OR "llm"[Title/Abstract] OR "generative ai"[Title/Abstract] OR "generative artificial intel"[Title/Abstract]	255542	11:03:11
Ovid MEDLINE® 1946 to January 22, 2025 ALL						
Search number	Query	-	-	-	Results	-
1	(Artificial Intelligence or artificial intel or language model or llm or generative ai or generative artificial intel).mp.	-	-	-	92,388	-
2	(Patient Education or patient educ or patient train or patient involv or patient quest or patient learn or patient instruction or patient communicat).mp.	-	-	-	1,13,525	-
3	(orthopaedic surgery or orthopedic surgery or orthopaedics or orthopedics or bone surgery or joint surgery or musculoskeletal surgery or spinal surgery or arthroplasty or fracture fixation or orthopedic procedure or orthopedic procedure).mp.	-	-	-	2,51,437	-
4	1 and 2 and 3	-	-	-	30	-
Embase (via Ovid) 1974 to January 22, 2025						
Search number	Query	-	-	-	Results	-
1	(Artificial Intelligence or artificial intel or language model or llm or generative ai or generative artificial intel).mp.	-	-	-	1,11,821	-
2	(Patient Education or patient educ or patient train or patient involv or patient quest or patient learn or patient instruction or patient communicat).mp.	-	-	-	1,42,126	-
	(orthopaedic surgery or orthopedic surgery or orthopaedics or orthopedics or bone surgery or joint surgery or musculoskeletal					

3	surgery or spinal surgery or arthroplasty or fracture fixation or orthopedic procedure or orthopedic procedure).mp.	-	-	-	2,76,151	-
4	1 and 2 and 3				30	-

TABLE 4: Line-by-line search strategies and results for PubMed, Ovid MEDLINE® ALL and Embase

Fields searched by .mp. in MEDLINE® = title, book title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism, protocol, rare disease, population and anatomy supplementary concept words, unique identifier and synonyms.

Fields searched by .mp. in Embase = title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword heading word, floating subheading word, candidate term word.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Hugh Jones, Kayden Chahal, Charles Picton, Taha Sherief Benamer, Ahmed Elnewishy

Acquisition, analysis, or interpretation of data: Hugh Jones, Kayden Chahal, Charles Picton

Drafting of the manuscript: Hugh Jones, Kayden Chahal, Charles Picton, Taha Sherief Benamer, Ahmed Elnewishy

Critical review of the manuscript for important intellectual content: Hugh Jones, Kayden Chahal, Charles Picton, Taha Sherief Benamer, Ahmed Elnewishy

Supervision: Hugh Jones, Kayden Chahal, Charles Picton, Ahmed Elnewishy

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: Payment/services info: All authors have declared that no financial support was received from any organization for the submitted work. Financial relationships: All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. Other relationships: All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. 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. 2025, 2025:5534704. 10.1155/aort/5534704

2. Anastasio AT, Mills FB 4th, Karavan MP Jr, Adams SB Jr: Evaluating the quality and usability of artificial intelligence-generated responses to common patient questions in foot and ankle surgery. Foot Ankle Orthop. 2023, 8:24730114231209919. 10.1177/24730114231209919

3. Kuo RY, Harrison C, Curran TA, et al.: Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology. 2022, 304:50-62. 10.1148/radiol.211785

4. Ren M, Yi PH: Artificial intelligence in orthopedic implant model classification: a systematic review . Skeletal Radiol. 2022, 51:407-16. 10.1007/s00256-021-03884-8

5. Lisacek-Kiosoglous AB, Powling AS, Fontalis A, Gabr A, Mazomenos E, Haddad FS: Artificial intelligence in orthopaedic surgery. Bone Joint Res. 2023, 12:447-54. 10.1302/2046-3758.127.BJR-2023-0111.R1

6. Smith AM, Jacquez EA, Argintar EH: Assessing the efficacy of an AI-powered chatbot (ChatGPT) in providing information on orthopedic surgeries: a comparative study with expert opinion. Cureus. 2024, 16:e63287. 10.7759/cureus.63287

7. Currie G, Robbie S, Tually P: ChatGPT and patient information in nuclear medicine: GPT 3.5 versus GPT-4 . J Nucl Med Technol. 2023, 51:307-13. 10.2967/jnmt.123.266151

8. White CA, Masturov YA, Haunschild E, Michaelson E, Shukla DR, Cagle PJ: Can ChatGPT reliably answer the most common patient questions regarding total shoulder arthroplasty?. J Shoulder Elbow Surg. 2025,

- 34:e254-64. [10.1016/j.jse.2024.08.025](https://doi.org/10.1016/j.jse.2024.08.025)
9. Liu J, Liu F, Wang C, Liu S: Prompt engineering in clinical practice: tutorial for clinicians . J Med Internet Res. 2025, 27:e72644. [10.2196/72644](https://doi.org/10.2196/72644)
10. Fahy S, Oehme S, Milinkovic DD, Bartek B: Enhancing patient education on the role of tibial osteotomy in the management of knee osteoarthritis using a customized ChatGPT: a readability and quality assessment. Front Digit Health. 2024, 6:1480381. [10.3389/fdgh.2024.1480381](https://doi.org/10.3389/fdgh.2024.1480381)
11. Liu S, McCoy AB, Wright A: Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025, 32:605-15. [10.1093/jamia/ocaf008](https://doi.org/10.1093/jamia/ocaf008)
12. Roberts RH, Ali SR, Dobbs TD, Whitaker IS: Can large language models generate outpatient clinic letters at first consultation that incorporate complication profiles from UK and USA Aesthetic Plastic Surgery Associations?. Aesthet Surg J Open Forum. 2024, 6:ojad109. [10.1093/asjof/ojad109](https://doi.org/10.1093/asjof/ojad109)
13. Zhang S, Liau ZQ, Tan KL, Chua WL: Evaluating the accuracy and relevance of ChatGPT responses to frequently asked questions regarding total knee replacement. Knee Surg Relat Res. 2024, 36:15. [10.1186/s43019-024-00218-5](https://doi.org/10.1186/s43019-024-00218-5)
14. Yerasomu T, Johns WL, Onor G, Menendez ME, Namdari S, Hammoud S: ChatGPT is capable of providing satisfactory responses to frequently asked questions regarding total shoulder arthroplasty. Shoulder Elbow. 2024, 16:407-12. [10.1177/17585732241246560](https://doi.org/10.1177/17585732241246560)
15. Wright BM, Bodnar MS, Moore AD, et al.: Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients?. Bone Jt Open. 2024, 5:139-46. [10.1302/2633-1462.52.BJO-2023-0113.R1](https://doi.org/10.1302/2633-1462.52.BJO-2023-0113.R1)
16. Lack BT, Mouhawas E, Childers JT, et al.: Can ChatGPT answer patient questions regarding reverse shoulder arthroplasty?. J ISAKOS. 2024, 9:100323. [10.1016/j.jisako.2024.100323](https://doi.org/10.1016/j.jisako.2024.100323)
17. Subramanian T, Araghi K, Amen TB, et al.: Chat generative pretraining transformer answers patient-focused questions in cervical spine surgery. Clin Spine Surg. 2024, 37:E278-81. [10.1097/BSD.0000000000001600](https://doi.org/10.1097/BSD.0000000000001600)
18. Gupta S, Haislup BD, Vaughan AK, Hoffman RA, Murthi AM: Accessing information provided via artificial intelligence regarding reverse and anatomic total shoulder arthroplasty. Semin Arthroplasty JSES. 2025, 35:56-61. [10.1053/j.sart.2024.09.001](https://doi.org/10.1053/j.sart.2024.09.001)
19. Yüce A, Yerli M, Misir A, Çakar M: Enhancing patient information texts in orthopaedics: how OpenAI's 'ChatGPT' can help. J Exp Orthop. 2024, 11:e70019. [10.1002/jeo2.70019](https://doi.org/10.1002/jeo2.70019)
20. Kirchner GJ, Kim RY, Weddle JB, Bible JE: Can artificial intelligence improve the readability of patient education materials?. Clin Orthop Relat Res. 2023, 481:2260-7. [10.1097/CORR0000000000002668](https://doi.org/10.1097/CORR0000000000002668)
21. Kasapovic A, Ali T, Babasiz M, Bojko J, Gathen M, Kaczmarczyk R, Roos J: Does the information quality of ChatGPT meet the requirements of orthopaedics and trauma surgery?. Cureus. 2024, 16:e60318. [10.7759/cureus.60318](https://doi.org/10.7759/cureus.60318)
22. Sparks CA, Fasulo SM, Windsor JT, Bankauskas V, Contrada EV, Kraeutler MJ, Scillia AJ: ChatGPT is moderately accurate in providing a general overview of orthopaedic conditions. JB JS Open Access. 2024, 9:e23.00129. [10.2106/JBJS.OA.23.00129](https://doi.org/10.2106/JBJS.OA.23.00129)
23. Artioli E, Veronesi F, Mazzotti A, Brogini S, Zielli SO, Giavaresi G, Faldini C: Assessing ChatGPT responses to common patient questions regarding total ankle arthroplasty. J Exp Orthop. 2025, 12:e70138. [10.1002/jeo2.70138](https://doi.org/10.1002/jeo2.70138)
24. Marder RS, Abdelmalek G, Richards SM, et al.: ChatGPT-3.5 and -4.0 do not reliably create readable patient education materials for common orthopaedic upper- and lower-extremity conditions. Arthrosc Sports Med Rehabil. 2025, 7:101027. [10.1016/j.asmr.2024.101027](https://doi.org/10.1016/j.asmr.2024.101027)
25. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM: Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. J Bone Joint Surg Am. 2023, 105:1519-26. [10.2106/JBJS.23.00209](https://doi.org/10.2106/JBJS.23.00209)
26. Tricco AC, Lillie E, Zarin W, et al.: PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018, 169:467-73. [10.7326/M18-0850](https://doi.org/10.7326/M18-0850)
27. Jones H, Chahal K, Picton C, Benamer TS, Elnewishy A: A systematic review of the utility of generative artificial intelligence in orthopaedic surgery patient education material [PREPRINT]. JMIR. 2026, [10.2196/preprints.81429](https://doi.org/10.2196/preprints.81429)
28. Peters MD, Marnie C, Tricco AC, et al.: Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020, 18:2119-26. [10.11124/JBIES-20-00167](https://doi.org/10.11124/JBIES-20-00167)
29. Tharakan S, Klein B, Bartlett L, Atlas A, Parada SA, Cohn RM: Do ChatGPT and Google differ in answers to commonly asked patient questions regarding total shoulder and total elbow arthroplasty?. J Shoulder Elbow Surg. 2024, 33:e429-37. [10.1016/j.jse.2023.11.014](https://doi.org/10.1016/j.jse.2023.11.014)
30. Stroop A, Stroop T, Zawy Alsofy S, Nakamura M, Möllmann F, Greiner C, Stroop R: Large language models: are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery?. Eur Spine J. 2024, 33:4135-43. [10.1007/s00586-023-07975-z](https://doi.org/10.1007/s00586-023-07975-z)
31. Li J, Gao X, Dou T, Gao Y, Li X, Zhu W: Quantitative evaluation of GPT-4's performance on US and Chinese osteoarthritis treatment guideline interpretation and orthopaedic case consultation. BMJ Open. 2024, 14:e082344. [10.1136/bmjopen-2023-082344](https://doi.org/10.1136/bmjopen-2023-082344)
32. Crook BS, Park CN, Hurley ET, Richard MJ, Pidgeon TS: Evaluation of online artificial-intelligence-generated information on common hand procedures. J Hand Surg Am. 2023, 48:1122-7. [10.1016/j.jhsa.2023.08.003](https://doi.org/10.1016/j.jhsa.2023.08.003)
33. Amaral JZ, Schultz RJ, Martin BM, et al.: Evaluating chat generative pre-trained transformer responses to common pediatric in-toeing questions. J Pediatr Orthop. 2024, 44:e592-7. [10.1097/BPO.0000000000002695](https://doi.org/10.1097/BPO.0000000000002695)
34. Nian PP, Umesh A, Simpson SK, et al.: Pediatric supracondylar humerus and diaphyseal femur fractures: a comparative analysis of chat generative pretrained transformer and Google Gemini recommendations versus American Academy of Orthopaedic Surgeons clinical practice guidelines. J Pediatr Orthop. 2025, 45:e338-44. [10.1097/BPO.0000000000002890](https://doi.org/10.1097/BPO.0000000000002890)
35. Giorgino R, Alessandri-Bonetti M, Del Re M, Verdoni F, Peretti GM, Mangiavini L: Google Bard and ChatGPT in orthopedics: which is the better doctor in sports medicine and pediatric orthopedics? The role of AI in patient education. Diagnostics (Basel). 2024, 14:1253. [10.3390/diagnostics14121253](https://doi.org/10.3390/diagnostics14121253)

36. Miskiewicz MJ, Perez M, Capotosto S, Ling K, Hance F, Komatsu D, Wang ED: Enhancing access to orthopedic education: exploring the potential of generative artificial intelligence (AI) in improving health literacy on rotator cuff injuries. *Cureus*. 2024, 16:e72833. [10.7759/cureus.72833](https://doi.org/10.7759/cureus.72833)
37. Miskiewicz MJ, Leonardo C, Capotosto S, Ling K, Cohen D, Komatsu D, Wang ED: Can generative artificial intelligence enhance health literacy about lateral epicondylitis?. *Cureus*. 2024, 16:e61384. [10.7759/cureus.61384](https://doi.org/10.7759/cureus.61384)
38. 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*. 2024, 49:986-91. [10.1016/j.jhsa.2024.05.007](https://doi.org/10.1016/j.jhsa.2024.05.007)
39. Andalib S, Solomon SS, Picton BG, Spina AC, Scolaro JA, Nelson AM: Source characteristics influence AI-enabled orthopaedic text simplification: recommendations for the future. *JB JS Open Access*. 2025, 10:e24.00007. [10.2106/jbjs.OA.24.00007](https://doi.org/10.2106/jbjs.OA.24.00007)
40. Kienzle A, Niemann M, Meller S, Gwinner C: ChatGPT may offer an adequate substitute for informed consent in patients prior to total knee arthroplasty - yet caution is needed. *J Pers Med*. 2024, 14:69. [10.3390/jpm14010069](https://doi.org/10.3390/jpm14010069)
41. Hu X, Niemann M, Kienzle A, et al.: Evaluating ChatGPT responses to frequently asked patient questions regarding periprosthetic joint infection after total hip and knee arthroplasty. *Digit Health*. 2024, 10:20552076241272620. [10.1177/20552076241272620](https://doi.org/10.1177/20552076241272620)
42. Dubin JA, Bains SS, DeRogatis MJ, et al.: Appropriateness of frequently asked patient questions following total hip arthroplasty from ChatGPT compared to arthroplasty-trained nurses. *J Arthroplasty*. 2024, 39:S306-11. [10.1016/j.arth.2024.04.020](https://doi.org/10.1016/j.arth.2024.04.020)
43. Kunze KN, Gerhold C, Dave U, et al.: Large language model use cases in health care research are redundant and often lack appropriate methodological conduct: a scoping review and call for improved practices. *Arthroscopy*. 2025, 41:4928-45.e2. [10.1016/j.arthro.2025.03.066](https://doi.org/10.1016/j.arthro.2025.03.066)
44. Tripathi S, Alkhulaifat D, Lyo S, et al.: A Hitchhiker's guide to good prompting practices for large language models in radiology. *J Am Coll Radiol*. 2025, 22:841-7. [10.1016/j.jacr.2025.02.051](https://doi.org/10.1016/j.jacr.2025.02.051)