Review

Examining the Role of Large Language Models in Orthopedics: Systematic Review

Cheng Zhang[1,2,3*], MS; Shanshan Liu[1,2,3*], MD; Xingyu Zhou[4], BS; Siyu Zhou[1,2,3], MD; Yinglun Tian[1,2,3], MD; Shenglin Wang[1,2,3], MD; Nanfang Xu[1,2,3*], MD, PhD; Weishi Li[1,2,3*], MD

[1]Department of Orthopaedics, Peking University Third Hospital, Beijing, China

[2]Engineering Research Center of Bone and Joint Precision Medicine, Ministry of Education, Beijing, China

[3]Beijing Key Laboratory of Spinal Disease Research, Beijing, China

[4]Peking University Health Science Center, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Weishi Li, MD
Department of Orthopaedics
Peking University Third Hospital
49 North Garden Road
Beijing, 100191
China
Phone: 86 01082267360
Email: puh3liweishi@163.com

Abstract

Background: Large language models (LLMs) can understand natural language and generate corresponding text, images, and even videos based on prompts, which holds great potential in medical scenarios. Orthopedics is a significant branch of medicine, and orthopedic diseases contribute to a significant socioeconomic burden, which could be alleviated by the application of LLMs. Several pioneers in orthopedics have conducted research on LLMs across various subspecialties to explore their performance in addressing different issues. However, there are currently few reviews and summaries of these studies, and a systematic summary of existing research is absent.

Objective: The objective of this review was to comprehensively summarize research findings on the application of LLMs in the field of orthopedics and explore the potential opportunities and challenges.

Methods: PubMed, Embase, and Cochrane Library databases were searched from January 1, 2014, to February 22, 2024, with the language limited to English. The terms, which included variants of "large language model," "generative artificial intelligence," "ChatGPT," and "orthopaedics," were divided into 2 categories: *large language model* and *orthopedics*. After completing the search, the study selection process was conducted according to the inclusion and exclusion criteria. The quality of the included studies was assessed using the revised Cochrane risk-of-bias tool for randomized trials and CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) guidance. Data extraction and synthesis were conducted after the quality assessment.

Results: A total of 68 studies were selected. The application of LLMs in orthopedics involved the fields of clinical practice, education, research, and management. Of these 68 studies, 47 (69%) focused on clinical practice, 12 (18%) addressed orthopedic education, 8 (12%) were related to scientific research, and 1 (1%) pertained to the field of management. Of the 68 studies, only 8 (12%) recruited patients, and only 1 (1%) was a high-quality randomized controlled trial. ChatGPT was the most commonly mentioned LLM tool. There was considerable heterogeneity in the definition, measurement, and evaluation of the LLMs' performance across the different studies. For diagnostic tasks alone, the accuracy ranged from 55% to 93%. When performing disease classification tasks, ChatGPT with GPT-4's accuracy ranged from 2% to 100%. With regard to answering questions in orthopedic examinations, the scores ranged from 45% to 73.6% due to differences in models and test selections.

Conclusions: LLMs cannot replace orthopedic professionals in the short term. However, using LLMs as copilots could be a potential approach to effectively enhance work efficiency at present. More high-quality clinical trials are needed in the future, aiming to identify optimal applications of LLMs and advance orthopedics toward higher efficiency and precision.

(J Med Internet Res 2024;26:e59607) doi: 10.2196/59607

KEYWORDS

large language model; LLM; orthopedics; generative pretrained transformer; GPT; ChatGPT; digital health; clinical practice; artificial intelligence; AI; generative AI; Bard

Introduction

Background

Large language models (LLMs) typically refer to pretrained language models (PLMs) that have a large number of parameters and are trained on massive amounts of data. In recent years, this area has emerged as one of the most prominent areas of research in artificial intelligence (AI) innovation [1,2]. What makes LLMs different from smaller-scale PLMs is their remarkable emergent abilities to solve complex tasks. Studies have found that LLMs, such as generative pretrained transformer (GPT)-3 with approximately 175 billion parameters, exhibit a significant leap in natural language processing (NLP) capabilities compared to PLMs with fewer parameters, such as GPT-2 with approximately 1.5 billion parameters [2,3]. Generative AI applications developed based on LLMs not only possess the ability to understand natural language but can also generate corresponding text, images, and even videos based on input sources. This human-machine interaction mode holds great potential in medical scenarios.

LLMs have undergone significant advancements in recent years; currently, the most prevalent web-based LLM service is ChatGPT (OpenAI). Launched in November 2022, ChatGPT is a chatbot application developed based on GPT-3.5 or GPT-4 after fine-tuning, and it can quickly respond to questions posed by users. In addition to ChatGPT, applications include Bard (upgraded to Gemini in December 2023) based on Language Model for Dialogue Applications (Google LLC); Med-PaLM 2 (Google LLC); ERNIE Bot (Baidu); and MOSS (Fudan University). GPT-4 can approach or achieve human-level performance in cognitive tasks across various fields, including medical domains [4]. When answering the 2022 United States Medical Licensing Examination questions, without further training or reinforcement, ChatGPT reached or approached a passing level in all 3 examinations [5]. However, answering examination questions does not directly reflect the performance of LLMs in clinical applications. The value and safety of a chatbot that is already in use are still not fully understood, making clinical research both essential and imperative. Published narrative reviews and editorials have explored the medical applications of LLM technology from 3 perspectives: clinical practice, education, and research [1,6-8]. These publications also provide a preliminary assessment of the value and safety of LLMs, offering guidance for exploring their use in specialized medical fields.

Orthopedics is a significant branch of medicine, typically encompassing disciplines such as trauma, spine surgery, joint surgery, sports medicine, hand surgery, and bone oncology. Orthopedic diseases have a broad impact on populations and pose a major global health threat. Low back pain, a common symptom in orthopedics or spine surgery, has been identified as the leading cause of global productivity loss, as measured in years, according to a large-scale epidemiological study covering 195 countries and regions; in 126 countries, low back pain ranks first among the causes of years lived with disability [9]. In traditional health care systems, the annual medical expenditure for low back pain in the United States is estimated to exceed US $100 billion, contributing to a significant socioeconomic burden [10]. Similarly, osteoarthritis is also a critical global health issue. The global prevalence of knee osteoarthritis in adults aged >40 years is 23%, with approximately 61% of adults aged >45 years showing radiographic evidence of knee osteoarthritis [11]. Therefore, applying LLMs in orthopedics holds the potential to alleviate the current heavy socioeconomic burden.

It is worth noting that several pioneers in orthopedics have conducted studies on LLMs across various subspecialties to explore their performance in addressing different issues. However, there are currently few reviews and summaries of these studies. The published reviews primarily focus on introducing and popularizing the basic concepts of LLMs in orthopedics [12,13], or they offer forward-looking perspectives by categorizing LLM applications in clinical practice, education, and research [14]. A systematic summary of existing research is absent. To the best of our knowledge, this review is the first to systematically summarize existing research findings. In contrast to prior works, we place greater emphasis on the quantitative evaluation methods and results of these studies because we believe that these methods and outcomes can help orthopedic and computer science researchers better understand the current state of LLM research and the performance of LLMs. Regarding application categorization, we consider tasks involving NLP in management as another important application area for LLMs in orthopedics. Therefore, this review adds a category for orthopedic management applications to the existing classification framework.

Objectives

The objective of this review was to comprehensively summarize the research findings on the application of LLMs in orthopedics and outline the advantages, limitations, and methodological evaluations, while also exploring the potential opportunities and challenges emerging in this era, for facilitating interdisciplinary collaboration and advancement among researchers in computer science and orthopedics. The ultimate goal is to contribute to improved efficiency and quality of orthopedic care as well as a reduction in medical costs and the associated socioeconomic burden.

Methods

Search Strategy

The protocol for this systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (checklist can be found in the Multimedia Appendix 1) [15]. PubMed, Embase, and Cochrane Library databases were searched, with the language limited to

English. The time frame was set from January 1, 2014, to February 22, 2024. Search terms were divided into 2 categories, with the first category including LLM-related terms and the second containing words related to orthopedics and its subspecialties (Textbox 1). Terms within each category were connected using "OR," while terms within different categories were connected using "AND." The full search strategy can be found in Multimedia Appendix 2.

Textbox 1. Categories and terms applied in the search queries.

Category 1

- "large language model," "LLM," "generative artificial intelligence," "generative AI," "ChatGPT," and "Generative Pre-Trained Transformer"

Category 2

- "orthopedics," "bone," "musculoskeletal," "injury," "wound," "trauma," "articular," "joint," "sports medicine," "hand surgery," "spine," "spinal," "cervical vertebrae," "thoracic vertebrae," "lumbar vertebrae," "sacrum," "coccyx," "spinal canal," "vertebral body," and "intervertebral disc"

Study Selection

The records were downloaded from the databases and imported into EndNote (version 21.2; Clarivate) for article management. The study selection process was conducted independently by 2 investigators (CZ and SL). The inclusion and exclusion criteria are listed in Textbox 2. The results were cross-checked, and discrepancies were resolved through discussion, with the final determination made by a third investigator (YT).

Textbox 2. Inclusion and exclusion criteria.

Inclusion criteria

- Article type
  - Original research
- Language
  - Articles written in English
- Content
  - Studies that use at least 1 large language model (LLM)
  - Studies that are relevant to the field of orthopedics

Exclusion criteria

- Article type
  - Reviews, editorials, letters, and study protocols
- Language
  - Articles written in a language other than English
- Content
  - Studies that do not involve LLMs
  - Studies that use LLMs for tasks such as code generation, debugging, or text generation without any performance evaluation of the model

Quality Assessment of Studies

Quality assessment was conducted by 2 investigators (CZ and SL) independently. First, the study designs were identified. Studies that involved only posing questions to LLMs, did not recruit participants, and did not report a study design were classified as surveys. Given the diverse nature of the survey types included in the review, quality assessments were conducted only for studies that recruited participants. The revised Cochrane risk-of-bias tool for randomized trials [16] was used to assess randomized controlled trials (RCTs), and the CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) guidance [17] was used to evaluate prospective or retrospective observational studies. The revised Cochrane risk-of-bias tool (version of August 22, 2019) is designed for assessing RCTs and contains 5 domains: bias arising from the randomization process, bias due to deviations from the intended interventions, bias due to missing outcome data, bias in measurement of the outcome, and bias in selection of the reported result. The CONSORT-AI guidance is a new reporting guideline specifically designed for clinical trials that assess interventions with an AI component. The quality assessment domains under this guidance include a statement of the AI algorithm used, details of how the AI intervention fits

within the clinical pathway, inclusion and exclusion criteria for input data, a description of the approaches used to handle unavailable input data, a description of the input data acquisition process for the AI intervention, specifications of human-AI interaction in the collection of input data, the output of the AI algorithm, and explanations of how the AI intervention's outputs contribute to health behavior changes. The results were cross-checked, and discrepancies were resolved through discussion, with the final determination made by another investigator (YT).

Data Extraction and Synthesis

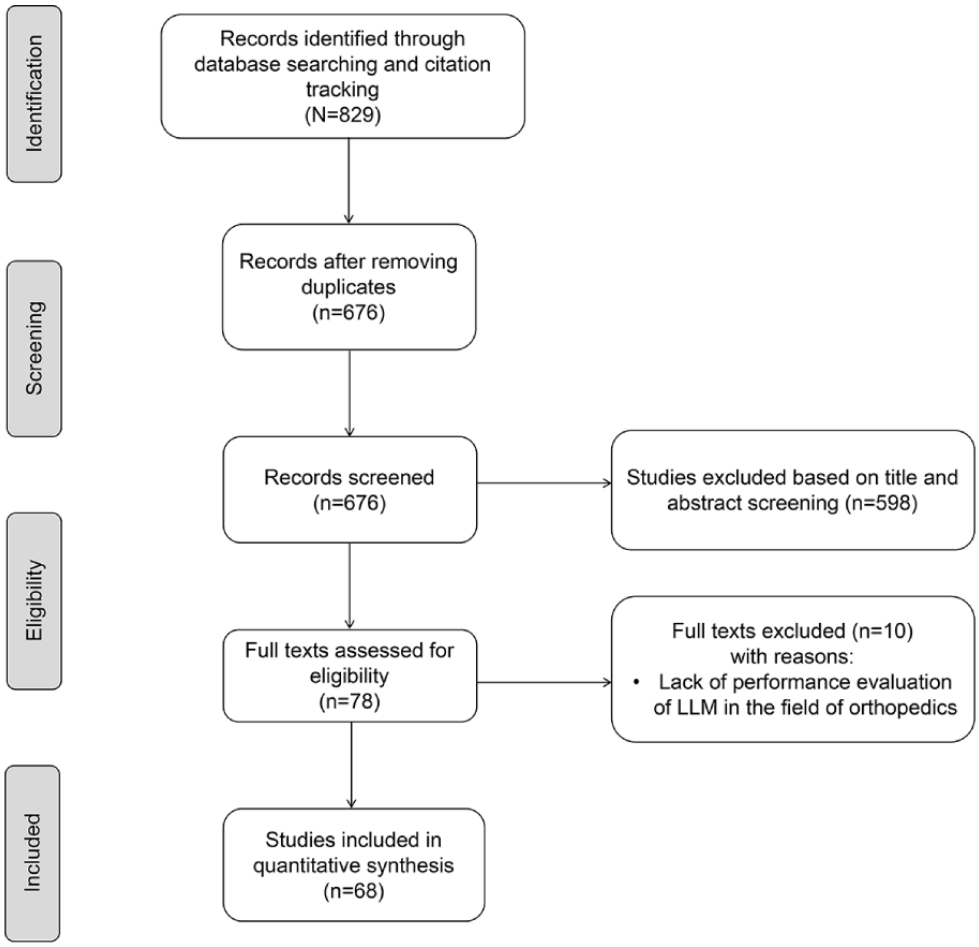
The studies were categorized into 4 groups based on their application areas: clinical practice, education, research, and management. Data extraction and synthesis were conducted by 2 investigators (CZ and SL) independently. In addition to general characteristics, the composition of extracted data varied depending on the specific category. Details of the data extraction strategy can be found in Multimedia Appendix 3. In cases where there were inconsistencies in the process, a third investigator (XZ) participated in the discussion and made the final decision. For studies with high heterogeneity, we did not synthesize the parameters for model performance evaluation and instead focused on providing a descriptive analysis of the data. Microsoft Excel 2019 was used for data collection, analysis, and visualization.

Results

Characteristics of Included Studies

A total of 829 studies were identified; after removing duplicates and screening, 68 (8.2%) studies were selected in the literature review. The inclusion process is shown in Figure 1.

Figure 1. Flowchart of literature screening based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement. LLM: large language model.



The application of LLMs in orthopedics involves the fields of clinical practice, education, research, and management. Of the 68 included studies, 47 (69%) focused on clinical practice (Table 1) [18-64], 12 (18%) addressed orthopedic education (Table 2) [65-76], 8 (12%) were related to scientific research (Table 3) [77-84], and 1 (1%) pertained to the field of management (Table 3) [85]. Of the 68 studies, 55 (81%) were classified as surveys; furthermore, only 8 (12%) recruited patients, only 1 (1%) was a high-quality study (RCT), and only 1 (1%) was a prospective study. Since June 2023, research on the application of LLMs in orthopedics has increased month by month (Figure 2).

Table 1. Characteristics of the included studies focused on clinical practice.

Study, year	Study design	Task	LLM[a] tools	Main evaluation metrics for model performance and their values	Enrolled participants, n	Subjective or objective assessment of the model's performance
Agharia et al [18], 2024	Survey	Formulate clinical decisions	GPT-3.5; GPT-4; Bard	Proportion of most popular response: 68% (GPT-4); 40.2% (GPT-3.5); 45.4% (Bard)	—[b]	Subjective
Anastasio et al [19], 2023	Survey	Generate answers to clinical questions	GPT-3.5	Ratio of responses in different quality grades: bottom-tier rating 4.5%; middle-tier rating 27.3%; top-tier rating 68.2%	—	Subjective
Baker et al [20], 2024	RCT[c]	Assist with writing patient histories	GPT-4	Mean time: 69.8 (SD 26.2) s; mean word count: 135.8 (SD 40.3); mean PDQI-9[d] score: 35.6 (SD 3.1); mean overall rating: 3.8 (SD 0.6); ratio of erroneous documents: 36%	11	Subjective
Christy et al [21], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Ratio of appropriate responses in total responses: 78%; intraclass correlation coefficient: 0.12	—	Subjective
Coraci et al [22], 2023	Cross-sectional study	Create questionnaire for assessment	GPT-3.5	Correlation: acceptable correlation with ODI[e] and QBPDS[f]; no statistical correlation with RMDQ[g] or NRS[h]	20	Subjective
Crook et al [23], 2023	Survey	Generate answers to clinical questions	GPT-3	DISCERN score: 58; JAMA[i] benchmark score: 0/4; FRE[j] score: 34; FKGL[k] score: 15	—	Subjective
Daher et al [24], 2023	Prospective study	Diagnose and manage patients	GPT-3	Accuracy of diagnosis: 93%; accuracy of management: 83%	29	Objective
Decker et al [25], 2023	Cross-sectional study	Generate informed consent documentation	GPT-3.5	Mean readability, accuracy, and completeness scores (surgeons vs LLMs): readability= 15.7 vs 12.9; risks=1.7 vs 1.7; benefits=1.4 vs 2.3; alternatives=1.4 vs 2.7; overall impression=1.9 vs 2.3; composite: 1.6 vs 2.2	—	Subjective
Draschl et al [26], 2023	Survey	Generate answers to clinical questions	GPT-3.5	5-point Likert scores, mean: completeness=3.80 (SD 0.63); misleading=4.04 (SD 0.67); errors=4.14 (SD 0.58); up-to-dateness=3.90 (SD 0.45); suitability for patients=3.69 (SD 0.64); suitability for surgeons=3.63 (SD 0.95)	—	Subjective
Dubin et al [27], 2023	Survey	Generate answers to clinical questions	GPT-3	25% of the questions were similar when performing a Google web search and a search of ChatGPT for all search terms; 75% of the questions were answered by government websites; 55% of the answers were different between Google web search and ChatGPT in terms of numerical questions	—	Subjective
Duey et al [28], 2023	Survey	Generate answers to clinical questions	GPT-3.5; GPT-4	Accuracy: 33% (GPT-3.5); 92% (GPT-4)	—	Subjective
Fabijan et al [29], 2023	Cross-sectional study	Classify cases of single-curve scoliosis	GPT-4; Microsoft Bing with GPT[l]; Scholar AI Premium	GPT-4 and Scholar AI Premium excelled in classifying single-curve scoliosis with perfect sensitivity (100%) and specificity (100%)	56	Objective
Fahy et al [30], 2024	Survey	Generate answers to clinical questions	GPT-3.5; GPT-4	GPT-3.5 vs GPT-4 mean DISCERN score: 55.4 vs 62.09; mean reading grade level score: 18.08 vs 17.90	—	Subjective

Study, year	Study design	Task	LLM[a] tools	Main evaluation metrics for model performance and their values	Enrolled participants, n	Subjective or objective assessment of the model's performance
Gianola et al [31], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Internal consistency: 49%; accuracy: 33%	—	Subjective
Hurley et al [32], 2024	Survey	Generate answers to clinical questions	ChatGPT	DISCERN score: 60; JAMA benchmark score: 0; FRE score: 26.2; FKGL score: considered to be that of a college graduate	—	Subjective
Johns et al [33], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Satisfaction rate: 60%	—	Subjective
Johns et al [34], 2024	Survey	Generate answers to clinical questions	GPT-3.5	DISCERN score: 41; FKGL score: 13.4; satisfaction rate: 40%	—	Subjective
Kaarre et al [35], 2023	Survey	Generate answers to clinical questions	GPT-4	Average correctness of responses for patients and physicians: 1.69 and 1.66, respectively (on a scale ranging from 0=incorrect, 1=partially correct, and 2=correct)	—	Subjective
Kasthuri et al [36], 2024	Survey	Generate answers to clinical questions	Microsoft Bing with GPT-4	Mean completeness score: 2.03; mean accuracy score: 4.49	—	Subjective
Kienzle et al [37], 2024	Survey	Generate answers to clinical questions	GPT-4	Mean DISCERN score in overall quality: 3.675	—	Subjective
Kirchner et al [38], 2023	Survey	Rewrite patient education materials	GPT-3.5	Mean FKGL score in patient education materials related to herniated lumbar disk, scoliosis, stenosis, TKA[m], and THA[n]: before rewrite=9.5, 12.6, 10.9, 12.0, and 6.3, respectively; after rewrite=5.0, 5.6, 6.9, 11.6, and 6.1, respectively	—	Subjective
Kuroiwa et al [39], 2023	Survey	Generate answers to clinical questions	GPT-3.5	Ratios of correct answers: 25/25, 1/25, 24/25, 16/25, and 17/25 for carpal tunnel syndrome, cervical myelopathy, lumbar spinal stenosis, knee osteoarthritis, and hip osteoarthritis, respectively	—	Objective
Li et al [40], 2023	Survey	Generate answers to clinical questions	GPT-4	Mean accuracy score (out of 5): 4.3; mean completeness score (out of 3): 2.8	—	Subjective
Li et al [41], 2024	Survey	Generate answers to clinical questions	GPT-3.5	1 response was excellent, requiring no clarification; 4 responses were satisfactory, requiring minimal clarification; 3 responses were satisfactory, requiring moderate clarification; 2 responses were unsatisfactory	—	Subjective
Lower et al [42], 2023	Survey	Deliver safe and coherent medical advice	GPT-4	Mean Likert scale score: 3.2	—	Subjective
Magruder et al [43], 2024	Survey	Generate answers to clinical questions	ChatGPT	Answer grades (from 1 to 5), mean: relevance=4.43 (SD 0.77); clarity=4.22 (SD 0.86); accuracy=4.10 (SD 0.90); evidence based=3.92 (SD 1.01); completeness=3.91 (SD 0.88); consistency=3.54 (SD 1.10)	—	Subjective

Study, year	Study design	Task	LLM[a] tools	Main evaluation metrics for model performance and their values	Enrolled participants, n	Subjective or objective assessment of the model's performance
Mika et al [44], 2023	Survey	Generate answers to clinical questions	GPT-3.5	2 responses were excellent, requiring no clarification; 4 responses were satisfactory, requiring minimal clarification; 3 responses were satisfactory, requiring moderate clarification; 1 response was unsatisfactory	—	Subjective
Pagano et al [45], 2023	Retrospective observational study	Formulate diagnosis and potential treatment suggestions	GPT-4	Diagnostic accuracy: 100% for the total cases; concordance in therapeutic recommendations: 83% for the total cases	100	Objective
Mejia et al [46], 2024	Survey	Generate answers to clinical questions	GPT-3.5; GPT-4	Accuracy: 52% (GPT-3.5); 59% (GPT-4); overconclusiveness: 48% (GPT-3.5); 45% (GPT-4)	—	Subjective
Russe et al [47], 2023	Retrospective observational study	Provide accurate fracture classification based on radiology reports	FraCChat; GPT-3.5-Turbo; GPT-4	Accuracy: GPT 3.5=3%; GPT 4=2%; FraCChat 3.5=48%; FraCChat 4=71%	—	Objective
Schonfeld et al [48], 2024	Retrospective cohort study	Predict outcome of adult spinal deformities	Gatortron	AUC[o] scores: 0.565 (pulmonary complication); 0.559 (neurological complication); 0.557 (sepsis); 0.508 (delirium); F_1-scores: 0.545 (pulmonary complication); 0.250 (neurological complication); 0.383 (sepsis); 0.156 (delirium)	209	Objective
Seth et al [49], 2023	Survey	Generate answers to clinical questions	ChatGPT	Mean Likert scale score: 3.1	—	Subjective
Shrestha et al [50], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Accuracy: 44%-65% for different guideline variations	—	Subjective
Sosa et al [51], 2024	Survey	Generate answers to clinical questions	GPT-4; Bard; Bing AI	Ratios of appropriate answers to questions related to bone physiology: 83.3% (GPT-4); 23.3% (Bing AI); 16.7% (Bard)	—	Subjective
Stroop et al [52], 2023	Survey	Generate answers to clinical questions	ChatGPT	Ratio of medically complete correct answers: 52%; ratio of medically complete and comprehensive answers: 55%	—	Subjective
Suthar et al [53], 2023	Retrospective observational study	Generate diagnosis	GPT-4	Accuracy rate in spine cases: 55%	—	Objective
Taylor et al [54], 2024	Survey	Generate answers to clinical questions	ChatGPT	Ratio of surgeons who reported that the questions had been appropriately answered: 91%	—	Subjective
Temel et al [55], 2024	Survey	Generate answers to clinical questions	GPT-4	Ensuring Quality Information for Patients score: mean 43.02 (SD 6.37); FRE score: mean 26.24 (SD 13.81); FKGL score: mean 14.84 (SD 1.79)	—	Subjective
Tharakan et al [56], 2024	Survey	Generate answers to clinical questions	GPT-3	Answers provided by ChatGPT cited more academic references than those provided by a Google search (80% vs 50%)	—	Subjective
Truhn et al [57], 2023	Retrospective observational study	Prioritize treatment recommendations	GPT-4	The overall quality of the treatment recommendations was rated as good or better	20	Subjective

Study, year	Study design	Task	LLM[a] tools	Main evaluation metrics for model performance and their values	Enrolled participants, n	Subjective or objective assessment of the model's performance
Warren et al [58], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Answers to fact, policy, and value questions (mean scores): DISCERN=51, 53, and 55, respectively; JAMA benchmark=0, 0, and 0, respectively; FRE=48.3, 42.0, and 38.4, respectively; FKGL=10.3, 10.9, and 11.6, respectively	—	Subjective
Wilhelm et al [59], 2023	Survey	Generate treatment recommendations	Claude-instant-v1.0; GPT 3.5-Turbo; Command-xlarge-nightly; Bloomz	Mean DISCERN quality scores: 3.4 (Claude-instant-v1.0); 2.8 (GPT 3.5-Turbo); 2.2 (Command-xlarge-nightly); 1.1 (Bloomz)	—	Subjective
Wright et al [60], 2024	Survey	Generate answers to clinical questions	GPT-3.5	Mean accuracy score: 4.26; mean comprehensiveness score: 3.79	—	Subjective
Yang et al [61], 2024	Retrospective observational study	Generate diagnosis	GPT-3.5	Accuracy: 0.87; sensitivity: 0.99; specificity: 0.73	1366	Objective
Yang et al [62], 2024	Survey	Generate answers to clinical questions	ChatGPT; Bard	Concordance with the AAOS[p] Clinical Practice Guidelines: 80% (ChatGPT); 60% (Bard)	—	Subjective
Yapar et al [63], 2024	Survey	Generate answers to clinical questions	GPT-4	Accuracy: 79.8%; applicability: 75.2%; comprehensiveness: 70.6%; communication clarity: 75.6%	—	Subjective
Zhou et al [64], 2024	Case study	Generate answers to clinical questions related to the case	GPT-3.5	No statistical results	—	Subjective

[a]LLM: large language model.

[b]Not applicable.

[c]RCT: randomized controlled trial.

[d]PDQI-9: Physician Documentation Quality Instrument-9.

[e]ODI: Oswestry Disability Index.

[f]QBPDS: Quebec Back Pain Disability Scale.

[g]RMDQ: Roland-Morris Disability Questionnaire.

[h]NRS: numerical rating scale.

[i]JAMA: Journal of the American Medical Association.

[j]FRE: Flesch reading ease.

[k]FKGL: Flesch-Kincaid grade level.

[l]GPT: generative pretrained transformer.

[m]TKA: total knee arthroplasty.

[n]THA: total hip arthroplasty.

[o]AUC: area under the curve.

[p]AAOS: American Academy of Orthopaedic Surgeons.

Table 2. Characteristics of the included studies focused on orthopedic education.

Study, year	Study design	Task	LLM[a] tools	Source	Questions, n	Scores or accuracy (%)
Cuthbert and Simpson [65], 2023	Survey	Examination	GPT-3.5	UKITE[b]	134	35.8
Ghanem et al [66], 2023	Survey	Examination	GPT-4	OITE[c]	201	61.2
Han et al [67], 2024	Survey	Examination	GPT-3.5	ASSH[d]	1583	36.2
Hofmann et al [68], 2023	Survey	Examination	GPT-3.5; GPT-4	OITE	410 (GPT-3.5); 396 (GPT-4)	GPT-3.5: 46.3; GPT-4: 63.4
Jain et al [69], 2024	Survey	Examination	GPT-3.5	OITE	360	52.8
Kung et al [70], 2023	Survey	Examination	GPT-3.5; GPT-4	OITE	360	GPT-3.5: 54.3; GPT-4: 73.6
Lum [71], 2023	Survey	Examination	GPT-3.5	OITE	207	47
Massey et al [72], 2023	Survey	Examination	GPT-3.5; GPT-4	ResStudy Orthopaedic Examination Question Bank	180	GPT-3.5: 29.4; GPT-4: 47.2
Ozdag et al [73], 2023	Survey	Examination	GPT-3.5	OITE	102	45
Rizzo et al [74], 2024	Survey	Examination	GPT-3.5-Turbo; GPT-4	OITE	2022: 207; 2021: 213; 2020: 215	2022: GPT-4=67.63; GPT 3.5-Turbo=50.24; 2021: GPT-4=58.69; GPT 3.5-Turbo=47.42; 2020: GPT-4=59.53; GPT 3.5-Turbo=46.51
Saad et al [75], 2023	Survey	Examination	GPT-4	Mock FRCS Orth[e] Part A	240	67.5
Traoré et al [76], 2023	Survey	Examination	GPT-3.5	EBHS[f] diploma examination	18	0

[a]LLM: large language model.

[b]UKITE: United Kingdom and Ireland In-Training Examination.

[c]OITE: Orthopaedic Surgery In-Training Examination.

[d]ASSH: American Society for Surgery of the Hand.

[e]FRCS Orth: Orthopaedic Fellow of the Royal College of Surgeons.

[f]EBHS: European Board of Hand Surgery.

Table 3. Characteristics of the included studies focused on orthopedic research and management.

Study, year	Study design	Task	LLM[a] tools	Input	Key findings
Gill et al [77], 2024	Survey	Improve readability	GPT-3.5	IRB[b]-approved orthopedic surgery research consent forms	ChatGPT can significantly improve the readability of orthopedic clinical research consent forms; 63.2% of the post-ChatGPT consent forms had at least 1 error
Hakam et al [78], 2024	Survey	AI[c]-Generated scientific literature	GPT-3.4; You.com	Five abstracts about meniscal injuries	The AI-generated texts could not be successfully identified
Kacena et al [79], 2024	Survey	Write scientific review articles	GPT-4	Prompts	AI reduced the time for writing but had significant inaccuracies
Lawrence et al [80], 2024	Survey	Generate abstract	GPT-3	A standard set of input commands	Interrater reliability for abstract quality scores was moderate
Lotz et al [81], 2023	Survey	Assist new research hypothesis exploration	Toolkit based on GPT-3.5	Prior studies	LLMs may be useful for analyzing and distinguishing publications, as well as determining the degree to which the literature supports or contradicts emergent hypotheses
Methnani et al [82], 2023	Survey	Calculate sample size	GPT-3.5	All necessary data, such as mean, percentage SD, normal deviations, and study design	In 1 (25%) of the 4 trials, the sample size was correctly calculated
Nazzal et al [83], 2024	Survey	Write a review article	GPT-4	Prompts	The AI-only paper was the most inaccurate, with inappropriate reference use, and the AI-assisted paper had the greatest incidence of plagiarism
Sanii et al [84], 2023	Survey	Perform an orthopedic surgery literature review	GPT-3; Perplexity	Standard prompts	The current iteration of ChatGPT cannot perform a reliable literature review, and Perplexity is only able to perform a limited review of the medical literature
Zaidat et al [85], 2023	Retrospective cohort study	Predict CPT[d] codes	GPT-4	Surgical operative notes	The AUROC[e] score was 0.87, and the AUPRC[f] score was 0.67

[a]LLM: large language model.

[b]IRB: institutional review board.

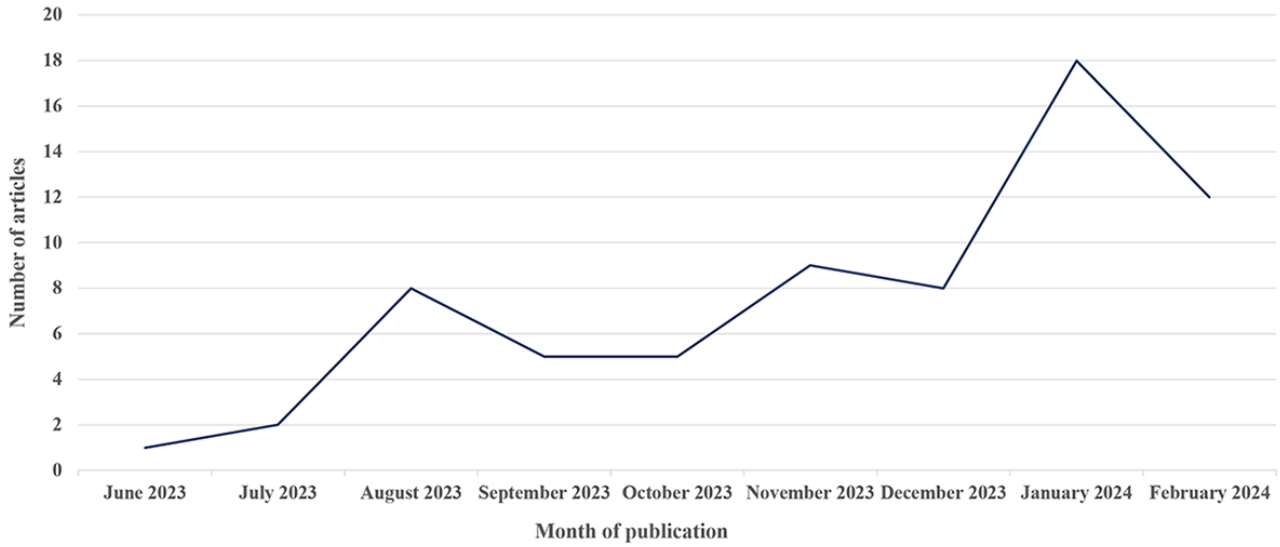
[c]AI: artificial intelligence.

[d]CPT: current procedural terminology.

[e]AUROC: area under the receiver operating characteristic curve.

[f]AUPRC: area under the precision-recall curve.

Figure 2. Trends in the number of publications.



Quality Assessment of Studies

We conducted quality assessments for the 8 studies that recruited participants (Multimedia Appendices 4 and 5). The RCT study was evaluated using the revised Cochrane risk-of-bias tool for randomized trials, and it was found to have a low risk of bias in all 5 domains—bias arising from the randomization process, bias due to deviations from the intended interventions, bias due to missing outcome data, bias in measurement of the outcome, and bias in selection of the reported result—indicating high study quality. The remaining studies (7/8, 88%; observational studies) were evaluated using the CONSORT-AI guidance and were found to be of good quality.

Distribution of LLM Tools

Among all the LLM tools applied, ChatGPT was the most commonly mentioned. Other LLM tools included Bard, Microsoft Bing, Scholar AI, Perplexity, Gatortron, Claude, Command-xlarge-nightly, and Bloomz, as well as software developed by researchers based on commonly used LLM kernels. Currently, there are 2 main versions of ChatGPT available: GPT-3 (including GPT-3.5 and GPT-3.5-Turbo) and GPT-4. Most of the studies (48/68, 71%) specified the version of the tool used. The majority of the studies (25/48, 52%) only used GPT-3 or 3.5, likely due to the publication lag because these studies were conducted before the release of GPT-4. Given that GPT-4 outperforms GPT-3 in most tasks, future research should primarily use GPT-4.

Model Performance Evaluation

As shown in Tables 1-3, there is considerable heterogeneity in the definition, measurement, and evaluation of LLM performance across the included studies. Currently, there is no unified research paradigm for the application of LLMs in medicine. Therefore, this review focused on different model performance evaluation metrics according to various application categories. For clinical applications of LLMs, we were particularly concerned with the accuracy of model reasoning and the readability of the generated text; unfortunately, the majority of the studies (39/47, 83%) relied on subjective assessments of the model's performance. In studies with objective evaluations, the heterogeneity in the subtasks performed by the LLMs (including diagnosis, classification, clinical case analysis, and case text generation) prevented us from pooling the data. For diagnostic tasks alone, the accuracy ranged from 55% to 93% [24,53]. When performing disease classification tasks, GPT-4's accuracy ranged from 2% to 100% [29,47] (Table 1). In studies on readability, the most commonly used metrics are the Flesch Reading Ease and the Flesch-Kincaid Grade Level (FKGL) scores. The FKGL metric correlates reading difficulty with years of education, providing a straightforward reflection of the readability of generated materials. In these studies, the generated texts had FKGL scores ranging from a minimum of 5.0 [38], indicating primary school reading difficulty, to the maximum required years of education [32], showing significant variability. This variability is likely due to differences in the research questions, methodologies, prompts, and evaluators. In the educational applications (eg, answering questions from examination papers), the most frequently used test source (7/12, 58%) was the Orthopaedic Surgery In-Training Examination (OITE). The test scores are widely recognized as performance evaluation metrics for the models, with final scores ranging from 45% to 73.6% due to differences in models and test selections (Table 2). For applications of LLMs in research and management, the flexible and varied nature of the tasks led to substantial differences in performance measurement and evaluation. Therefore, we collected the model inputs and major findings for a descriptive presentation (Table 3).

Discussion

Overview

Despite the relatively short time since their introduction and the absence of rigorous and comprehensive performance evaluation in highly specialized fields such as orthopedics, it is an undeniable fact that LLMs have already been made accessible to the public. Given the increasing acceptance and widespread adoption of LLMs, it is imperative for orthopedic surgeons to possess a comprehensive understanding of their operational mechanisms and limitations. Users should also delineate the safe application boundaries while harnessing the benefits offered by LLMs, all while mitigating potential risks in their daily clinical practice. This section presents a comprehensive overview of application examples and model performance across diverse fields, while providing strategic approaches to address LLMs based on our findings. In addition, in this section, we critically evaluate research methodologies and offer potential recommendations for future investigations.

Application of LLMs in the Field of Orthopedic Education

LLMs can not only provide answers but also offer explanations and even engage in further discussions on a given topic, demonstrating potential value in orthopedic education. Several studies have evaluated the performance of ChatGPT in answering questions related to orthopedics and further discussed its value in the field of orthopedic education. The source of questions includes the OITE [66,68-71,73,74], the ResStudy Orthopaedic Examination Question Bank [72], the Fellowship of the Royal College of Surgeons (Trauma and Orthopaedic Surgery) examination [65,75], and hand surgery examinations in the United States and Europe [67,76]. Accuracy (scores) and whether the answers meet the standard are important evaluation criteria for LLM performance. Another educational indicator is the correctness and reasonableness of answer explanations. Studies evaluating OITE questions usually convert accuracy into postgraduate year (PGY) levels for evaluation. Due to differences in software applications and question selection, different studies have reported varying performances of ChatGPT. ChatGPT with GPT-4 performed at an average level ranging from PGY-2 to PGY-5 [66,68,70,74], while ChatGPT with GPT-3.5 performed slightly better than PGY-1 or below the average level of PGY-1 [68-71,73,74]. For correct answers, ChatGPT can provide explanations and reasoning processes consistent with those of examiners, which helps students understand the questions and general orthopedic principles [66,69]. However, ChatGPT failed to pass the Fellowship of the Royal College of Surgeons (Trauma and Orthopaedic

Surgery) examination and hand surgery examinations in the United States and Europe [65,67,75,76]. In addition, as a language model, ChatGPT cannot analyze medical images correctly [70], limiting its role in orthopedic imaging education.

Although LLMs currently cannot fully replace orthopedic instructors, they can still serve as a valuable supplementary tool for learning. Integrating their responses with authoritative resources for verification and using appropriate prompts can optimize their capacity to offer logical explanations and foster critical thinking.

Application of LLMs in Clinical Practice

Medical Consultation and Physician-Patient Communication

One challenge faced by orthopedic physicians is that, unlike in the case of other clinical interventions, LLMs have already been integrated as medical consultation tools in the diagnosis and treatment process of numerous diseases without sufficient clinical evidence and regulatory review from authorities such as the US Food and Drug Administration. LLMs can be considered an alternative approach for patients who have sustained injuries or experience discomfort before seeking guidance from primary care physicians or specialists. When confronted with medical issues, individuals who rely heavily on the internet for problem-solving in their personal and professional lives often exhibit a tendency to seek treatment decisions on the web [30]. Compared to traditional search engines or Wikipedia, LLMs could potentially become a significant source of medical consultation information, especially in cases of nonacute diseases such as lower back pain or joint pain. Meanwhile, many physicians also hope that LLMs can help alleviate their burden of simple medical consultations and repetitive paperwork related to physician-patient communication (such as preoperative consent forms), which is considered 1 of the important factors contributing to physician burnout [86]. Although LLMs can provide concise, clarified, or simplified responses related to the given topic and deliver high-quality and empathetic answers [66,87], given their imperfect performance in addressing questions related to orthopedics [65-76], caution should be exercised regarding their reliability in orthopedic consultation scenarios.

Studies have evaluated the performance of LLMs in answering questions related to hand surgery [23], spinal cord injuries [55], joint and sports medicine [19,27,35,41,56,58], and preoperative physician-patient communication for lumbar disk herniation [52] and hip replacement surgery [44]. In these studies, the evaluation criteria of interest typically encompass the model's answer accuracy, readability, completeness, and information sources. Evaluation methods often encompass scale assessments or subjective ratings conducted by researchers. The DISCERN score is commonly used to evaluate answer quality [19,23,58,88], while FKGL and Flesch Reading Ease scores are commonly used to measure readability [23,55,58]. The accuracy of LLMs' responses is closely correlated with the specific topic. Questions in the field of joint and sports medicine often receive high-quality responses, while there are serious issues with the quality of answers regarding spinal cord injuries. There are also significant differences in the evaluation of the readability or comprehensibility of LLMs' answers, with some researchers considering them to be easily understood [44,52], while studies using Flesch-related scales suggest that LLMs' answers require a reading level of at least 10 years of education or even university level for full comprehension [23,55,58]. The underlying factors contributing to this phenomenon can be attributed to variations in question topics, prompts, and evaluation methodologies used for answer assessment. Consequently, orthopedic surgeons should exercise caution when interpreting the findings of these studies.

Although LLMs can offer more scholarly health information in comparison to search engines [27,56], they still cannot replace orthopedic physicians in medical consultation and physician-patient communication. Using LLMs as a guiding tool and maintaining communication with physicians during further diagnosis and treatment decisions may be a safer and more effective strategy.

Clinical Workflow

The performance of LLMs in orthopedic examinations suggests that they cannot handle complex tasks independently, but they hold potential to serve as valuable assistants for orthopedic physicians. One possible application is using LLMs to automate simple, repetitive tasks such as writing medical records for common orthopedic diseases [20]. In the context of complex disease management tasks, LLMs can possess a more extensive and specialized knowledge base than less experienced newly graduated physicians and assist them in various aspects of disease management. Some researchers have tested the performance of LLMs using specific clinical questions or guidelines [21,26,28,50], while others have directly inputted clinical case data into the model, allowing it to summarize and provide corresponding diagnostic or treatment decisions autonomously [18,24,45,57,64]. Currently, there is no further research on introducing LLMs into orthopedic operations, likely because of the limited availability of intelligent terminals and digital scenarios that may combine operative procedures with LLMs. Potential docking scenarios for the LLM model could include intelligent surgical applications such as mixed reality operating rooms [89,90] and autonomous laminectomy robots [91,92].

In the context of clinical practice, apart from the fundamental requirement of accurate response, time consumption and work efficiency also serve as crucial reference indicators for evaluating LLMs' performance. Despite variations in the assessment of model accuracy across the included studies, potentially attributed to differences in research objectives, prompt design, evaluation criteria, and assessment tools, no study has presented evidence indicating that LLMs can independently perform clinical work. Therefore, the current models still require rigorous supervision during their use. An RCT study evaluating the performance of ChatGPT in assisting with orthopedic clinical documentation found that there was no significant efficiency advantage in using ChatGPT: the time taken to complete medical history writing was not superior to voice input, and instances of fabricated content were observed within the ChatGPT-generated medical histories [20].

Although LLMs currently have limitations, they remain valuable tools for orthopedic surgeons in their daily practice. It is important to approach cautiously the responses provided by LLMs and seek additional evidence and explanations from the model used when faced with unclear answers. By incorporating evidence-based medicine tools, we can ultimately achieve superior clinical diagnoses and treatment plans, thereby elevating the quality of care delivered by physicians.

Application of LLMs in the Field of Research

Research is generally considered a creative endeavor, and introducing LLMs into the field of research may offer more flexibility. Currently, there are limited attempts to use LLMs in orthopedic research. A study found that lowering the reading threshold of professional texts through LLMs can assist in improving the readability of informed consent forms for orthopedic clinical research, but the forms did not meet the recommended sixth-grade reading level set by the American Medical Association [77]. In addition, the literature summarization and generation capabilities of LLMs can contribute to independent or assisted writing of literature reviews in the orthopedic field [79,83]. On the other side of the coin, concerns about integrity arise when scholars find that the model's output can be deceptively realistic. The abstracts generated by LLMs for studies on *meniscal injuries* and *joint replacement* were indistinguishable from those written by human researchers [78,80]. However, web-based LLMs do not perform well in tasks such as literature review or sample size estimation in sports medicine research [82,84]. Possible reasons may include the potential limitations of LLMs in meeting logical reasoning requirements and the inappropriate use of prompts. For more complex tasks, an optimization approach could involve developing task-specific toolkits based on the fundamental architecture of LLMs. The feasibility of this approach has been validated in interdisciplinary research on the management of back pain [81].

Application of LLMs in Management

Trained NLP models can convert natural language into structured data and have demonstrated superior performance in tasks involving the current procedural terminology for identifying spinal surgery records [93]. However, ChatGPT, with its larger parameters, performs weaker than NLP models in the task of identifying spinal surgery current procedural terminology codes [85]. One possible reason is that traditional NLP models have been trained on more targeted datasets, whereas researchers cannot fine-tune the backend model of ChatGPT using these data. Despite the current model's performance limitations hindering its further application in this field, the potential advancements in "fine-tuning" techniques may enable LLMs to assume a more influential role in orthopedic management in the future.

Current Advantages and Limitations of LLMs in Orthopedic Applications

Overview

In contrast to conventional pretrained machine learning models, LLMs exhibit the advantage of versatility by accurately addressing problems across various domains without necessitating additional training on specific samples. In the field of orthopedics, another advantage of LLMs is their user-friendly and convenient nature. Users do not need to go through the long process of waiting and referral from general practitioners to specialists. By simply accessing apps equipped with LLMs, users can inquire about diverse subspecialties in orthopedics at any time and from anywhere, receiving answers promptly at a minimal cost or even free of charge. This service surpasses the capabilities of current health care systems and is unlikely to be replicated in the foreseeable future.

However, as mentioned previously, these advantages are based on unverified answers and unpredictable risks. The answers provided by LLMs for questions related to orthopedics are less robust than those for everyday common knowledge and have significant limitations in terms of accuracy, readability, reliability, and timeliness, as detailed in the following subsections.

Accuracy

Almost all studies (66/68, 97%) found errors in LLMs' responses, with more noticeable inaccuracies in specialized areas such as hip and knee joints and hand surgery [62,67,76]. Some answers even contradicted fundamental orthopedic knowledge [52]. Therefore, some researchers argue that current expectations for guidance provided by AI platforms should be tempered by both physicians and patients [62]. Possible reasons include the limited availability of publicly accessible orthopedic data for training, especially for specialized diseases, as well as privacy concerns that restrict public access to a large amount of data. In the future, besides waiting for more powerful next-generation LLMs, using existing LLMs to learn orthopedic cases and fine-tuning them may be a potential solution to improve accuracy.

Readability

Some of the included studies (3/68, 4%) suggest that the content generated by LLMs is not satisfactory in terms of readability for the general population [23,55,58]. The potential reasons for the lack of readability may include not only the limited training data but also the quality of the trained data. By incorporating more popular science content and common clinical responses, it may be possible to address the issue of readability through fine-tuning the model.

Reliability

Different ways of asking the same question may yield completely different answers [21]. This instability, particularly in response to specific prompts, not only affects users' experience and trust but also greatly interferes with researchers' homogenized evaluations. It is imperative to establish standardized questioning processes and prompt criteria.

Timeliness

Training LLMs from scratch is both costly and time consuming, leading to significant retraining expenses. However, unlike everyday common knowledge, orthopedics is constantly evolving with new diagnostic and treatment approaches as well as surgical techniques. Therefore, outdated information becomes

an important risk factor leading to inaccurate answers, necessitating caution in this context.

Methodological Limitations of the Selected Studies

Although there are 47 studies related to clinical issues, only 8 (17%) recruited patients [20,22,24,29,45,48,57,61]. Many of the studies (39/47, 83%) only focus on investigation and evaluation, lacking rigorous methods for clinical research, such as RCTs. Furthermore, there is a lack of research end points directly linked to patient outcomes, such as cure rates or improvements in quality of life, making it difficult to find direct evidence of prognosis. Most of the studies (46/68, 68%) rely on subjective methodologies, such as expert ratings, for model evaluation and lack objective criteria and approaches for assessment, leading to unreliable research results. Furthermore, the absence of standardized questioning paradigms has led to instability in LLM responses, posing challenges for reproducibility and limiting the reliability and clinical significance of the study findings.

Limitations of This Review

This systematic review has several limitations. First, only English-language articles were included, which may have led to the exclusion of relevant studies published in other languages. Second, due to significant heterogeneity in study designs, model tasks, and evaluation parameters among the included studies, we did not perform a comprehensive synthesis of most of the data, nor did we conduct a meta-analysis. Third, our search was restricted to commonly used medical research databases such as PubMed, Embase, and Cochrane Library, potentially overlooking relevant studies from other sources, including conference papers and gray literature. Finally, given the limited availability of rigorous clinical studies, we included a considerable number of subjective surveys. Although our objective was to provide a broad overview of LLM-related information, this may have introduced bias into the findings. These limitations are expected to be addressed as more standardized, high-quality clinical studies become available in future research.

Conclusions

Due to the current limitations of LLMs, they cannot replace orthopedic professionals in the short term. However, using LLMs as copilots could be a potential approach to effectively enhance work efficiency at present. In addition, developing task-specific downstream tools based on LLMs is also a potential solution to improve model performance for further use.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82272577). The funding organization had no involvement in the design of the study, data collection, data analysis, data interpretation, writing of the report, or the decision to publish. Its role was strictly limited to providing financial support. We used the generative AI tool ChatGPT with GPT-4 [94] for language refinement purposes. The AI tool was only used to perform grammar checks and assist with enhancing the clarity and fluency of the writing. All intellectual contributions, including research design, data collection, analysis, and interpretation, were made by the authors. The final manuscript has been reviewed and approved by the authors, who take full responsibility for its content.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Authors' Contributions

NX, WL, CZ, and SL made contributions to conception and design. XZ, SZ, and YT made contributions to the acquisition, analysis, and interpretation of data. CZ, SL, and XZ drafted the manuscript, and NX, WL, SZ, YT, and SW revised it critically for important intellectual content. All authors approved the version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 83 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search strategy.
[DOCX File , 25 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Data extraction strategy.
[DOCX File , 22 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Revised Cochrane risk-of-bias tool for randomized trials template for assessment completion.
[PDF File (Adobe PDF File), 232 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Quality assessment of studies of large language models based on CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) guidance.
[DOCX File , 53 KB-Multimedia Appendix 5]

References

1. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. Aug 17, 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
2. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv. Preprint posted online on March 31, 2023. [FREE Full text]
3. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv. Preprint posted online on June 15, 2022. [FREE Full text]
4. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv. Preprint posted online on March 15, 2023. [FREE Full text]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. Feb 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA. Sep 05, 2023;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
7. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls. Ann Intern Med. Feb 2024;177(2):210-220. [doi: 10.7326/m23-2772]
8. Tang YD, Dong ED, Gao W. LLMs in medicine: the need for advanced evaluation systems for disruptive technologies. Innovation (Camb). May 06, 2024;5(3):100622. [FREE Full text] [doi: 10.1016/j.xinn.2024.100622] [Medline: 38699776]
9. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. Nov 10, 2018;392(10159):1789-1858. [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
10. Knezevic NN, Candido KD, Vlaeyen JW, Van Zundert J, Cohen SP. Low back pain. Lancet. Jul 03, 2021;398(10294):78-92. [doi: 10.1016/S0140-6736(21)00733-9] [Medline: 34115979]
11. Duong V, Oo WM, Ding C, Culvenor AG, Hunter DJ. Evaluation and treatment of knee pain: a review. JAMA. Oct 24, 2023;330(16):1568-1580. [doi: 10.1001/jama.2023.19675] [Medline: 37874571]
12. Yao JJ, Aggarwal M, Lopez RD, Namdari S. Current concepts review: large language models in orthopaedics: definitions, uses, and limitations. J Bone Joint Surg Am. Jun 19, 2024. [doi: 10.2106/JBJS.23.01417] [Medline: 38896652]
13. Fayed AM, Mansur NS, de Carvalho KA, Behrens A, D'Hooghe P, de Cesar Netto C. Artificial intelligence and ChatGPT in orthopaedics and sports medicine. J Exp Orthop. Jul 26, 2023;10(1):74. [FREE Full text] [doi: 10.1186/s40634-023-00642-8] [Medline: 37493985]
14. Merrell LA, Fisher ND, Egol KA. Large language models in orthopaedic trauma: a cutting-edge technology to enhance the field. J Bone Joint Surg Am. Sep 06, 2023;105(17):1383-1387. [doi: 10.2106/JBJS.23.00395] [Medline: 37402227]
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
16. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. Aug 28, 2019;366:l4898. [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
17. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, SPIRIT-AICONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. BMJ. Sep 09, 2020;370:m3164. [FREE Full text] [doi: 10.1136/bmj.m3164] [Medline: 32909959]
18. Agharia S, Szatkowski J, Fraval A, Stevens J, Zhou Y. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: an analysis of ChatGPT 3.5, ChatGPT 4, and Bard. J Orthop. Apr 2024;50:1-7. [FREE Full text] [doi: 10.1016/j.jor.2023.11.063] [Medline: 38148925]
19. Anastasio AT, Mills FB4, Karavan MPJ, Adams SBJ. Evaluating the quality and usability of artificial intelligence-generated responses to common patient questions in foot and ankle surgery. Foot Ankle Orthop. Oct 22, 2023;8(4):24730114231209919. [FREE Full text] [doi: 10.1177/24730114231209919] [Medline: 38027458]

20. Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's ability to assist with clinical documentation: a randomized controlled trial. J Am Acad Orthop Surg. Feb 01, 2024;32(3):123-129. [doi: 10.5435/JAAOS-D-23-00474] [Medline: 37976385]
21. Christy M, Morris MT, Goldfarb CA, Dy CJ. Appropriateness and reliability of an online artificial intelligence platform's responses to common questions regarding distal radius fractures. J Hand Surg Am. Feb 2024;49(2):91-98. [doi: 10.1016/j.jhsa.2023.10.019] [Medline: 38069953]
22. Coraci D, Maccarone MC, Regazzo G, Accordi G, Papathanasiou JV, Masiero S. ChatGPT in the development of medical questionnaires. The example of the low back pain. Eur J Transl Myol. Dec 15, 2023;33(4):12114. [FREE Full text] [doi: 10.4081/ejtm.2023.12114] [Medline: 38112605]
23. 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. Nov 2023;48(11):1122-1127. [doi: 10.1016/j.jhsa.2023.08.003] [Medline: 37690015]
24. Daher M, Koa J, Boufadel P, Singh J, Fares MY, Abboud JA. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? JSES Int. Nov 2023;7(6):2534-2541. [FREE Full text] [doi: 10.1016/j.jseint.2023.07.018] [Medline: 37969495]
25. Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open. Oct 02, 2023;6(10):e2336997. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.36997] [Medline: 37812419]
26. Draschl A, Hauer G, Fischerauer SF, Kogler A, Leitner L, Andreou D, et al. Are ChatGPT's free-text responses on periprosthetic joint infections of the hip and knee reliable and useful? J Clin Med. Oct 20, 2023;12(20):6655. [FREE Full text] [doi: 10.3390/jcm12206655] [Medline: 37892793]
27. Dubin JA, Bains SS, Chen Z, Hameed D, Nace J, Mont MA, et al. Using a Google web search analysis to assess the utility of ChatGPT in total joint arthroplasty. J Arthroplasty. Jul 2023;38(7):1195-1202. [doi: 10.1016/j.arth.2023.04.007] [Medline: 37040823]
28. Duey AH, Nietsch KS, Zaidat B, Ren R, Ndjonko LC, Shrestha N, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. Spine J. Nov 2023;23(11):1684-1691. [doi: 10.1016/j.spinee.2023.07.015] [Medline: 37499880]
29. Fabijan A, Polis B, Fabijan R, Zakrzewski K, Nowosławska E, Zawadzka-Fabijan A. Artificial intelligence in scoliosis classification: an investigation of language-based models. J Pers Med. Dec 09, 2023;13(12):1695. [FREE Full text] [doi: 10.3390/jpm13121695] [Medline: 38138922]
30. Fahy S, Oehme S, Milinkovic D, Jung T, Bartek B. Assessment of quality and readability of information provided by ChatGPT in relation to anterior cruciate ligament injury. J Pers Med. Jan 18, 2024;14(1):104. [FREE Full text] [doi: 10.3390/jpm14010104] [Medline: 38248805]
31. Gianola S, Bargeri S, Castellini G, Cook C, Palese A, Pillastrini P, et al. Performance of ChatGPT compared to clinical practice guidelines in making informed decisions for lumbosacral radicular pain: a cross-sectional study. J Orthop Sports Phys Ther. Mar 2024;54(3):222-228. [doi: 10.2519/jospt.2024.12151] [Medline: 38284363]
32. Hurley ET, Crook BS, Lorentz SG, Danilkowicz RM, Lau BC, Taylor DC, et al. Evaluation high-quality of information from ChatGPT (artificial intelligence-large language model) artificial intelligence on shoulder stabilization surgery. Arthroscopy. Mar 2024;40(3):726-31.e6. [doi: 10.1016/j.arthro.2023.07.048] [Medline: 37567487]
33. Johns WL, Kellish A, Farronato D, Ciccotti MG, Hammoud S. ChatGPT can offer satisfactory responses to common patient questions regarding elbow ulnar collateral ligament reconstruction. Arthrosc Sports Med Rehabil. Apr 2024;6(2):100893. [FREE Full text] [doi: 10.1016/j.asmr.2024.100893] [Medline: 38375341]
34. Johns WL, Martinazzi BJ, Miltenberg B, Nam HH, Hammoud S. ChatGPT provides unsatisfactory responses to frequently asked questions regarding anterior cruciate ligament reconstruction. Arthroscopy. Jul 2024;40(7):2067-79.e1. [doi: 10.1016/j.arthro.2024.01.017] [Medline: 38311261]
35. Kaarre J, Feldt R, Keeling LE, Dadoo S, Zsidai B, Hughes JD, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports Traumatol Arthrosc. Nov 2023;31(11):5190-5198. [FREE Full text] [doi: 10.1007/s00167-023-07529-2] [Medline: 37553552]
36. Kasthuri VS, Glueck J, Pham H, Daher M, Balmaceno-Criss M, McDonald CL, et al. Assessing the accuracy and reliability of AI-generated responses to patient questions regarding spine surgery. J Bone Joint Surg Am. Jun 19, 2024;106(12):1136-1142. [doi: 10.2106/JBJS.23.00914] [Medline: 38335266]
37. Kienzle A, Niemann M, Meller S, Gwinner C. ChatGPT may offer an adequate substitute for informed consent to patients prior to total knee arthroplasty-yet caution is needed. J Pers Med. Jan 05, 2024;14(1):69. [FREE Full text] [doi: 10.3390/jpm14010069] [Medline: 38248771]
38. 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]
39. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res. Sep 15, 2023;25:e47621. [FREE Full text] [doi: 10.2196/47621] [Medline: 37713254]

40. Li J, Gao X, Dou T, Gao Y, Zhu W. Assessing the performance of GPT-4 in the filed of osteoarthritis and orthopaedic case consultation. medRxiv. Preprint posted online on August 09, 2023. [FREE Full text] [doi: 10.1101/2023.08.06.23293735]
41. Li LT, Sinkler MA, Adelstein JM, Voos JE, Calcei JG. ChatGPT responses to common questions about anterior cruciate ligament reconstruction are frequently satisfactory. Arthroscopy. Jul 2024;40(7):2058-2066. [doi: 10.1016/j.arthro.2023.12.009] [Medline: 38171421]
42. Lower K, Seth I, Lim B, Seth N. ChatGPT-4: transforming medical education and addressing clinical exposure challenges in the post-pandemic era. Indian J Orthop. Sep 2023;57(9):1527-1544. [doi: 10.1007/s43465-023-00967-7] [Medline: 37609022]
43. Magruder ML, Rodriguez AN, Wong JCJ, Erez O, Piuzzi NS, Scuderi GR, et al. Assessing ability for ChatGPT to answer total knee arthroplasty-related questions. J Arthroplasty. Aug 2024;39(8):2022-2027. [doi: 10.1016/j.arth.2024.02.023] [Medline: 38364879]
44. 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. Oct 04, 2023;105(19):1519-1526. [doi: 10.2106/JBJS.23.00209] [Medline: 37459402]
45. Pagano S, Holzapfel S, Kappenschneider T, Meyer M, Maderbacher G, Grifka J, et al. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative AI model GPT-4. J Orthop Traumatol. Nov 28, 2023;24(1):61. [FREE Full text] [doi: 10.1186/s10195-023-00740-4] [Medline: 38015298]
46. Mejia MR, Arroyave JS, Saturno M, Ndjonko LC, Zaidat B, Rajjoub R, et al. Use of ChatGPT for determining clinical and surgical treatment of lumbar disc herniation with radiculopathy: a North American Spine Society guideline comparison. Neurospine. Mar 2024;21(1):149-158. [FREE Full text] [doi: 10.14245/ns.2347052.526] [Medline: 38291746]
47. Russe MF, Fink A, Ngo H, Tran H, Bamberg F, Reisert M, et al. Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports. Sci Rep. Aug 30, 2023;13(1):14215. [FREE Full text] [doi: 10.1038/s41598-023-41512-8] [Medline: 37648742]
48. Schonfeld E, Pant A, Shah A, Sadeghzadeh S, Pangal D, Rodrigues A, et al. Evaluating computer vision, large language, and genome-wide association models in a limited sized patient cohort for pre-operative risk stratification in adult spinal deformity surgery. J Clin Med. Jan 23, 2024;13(3):656. [FREE Full text] [doi: 10.3390/jcm13030656] [Medline: 38337352]
49. Seth I, Xie Y, Rodwell A, Gracias D, Bulloch G, Hunter-Smith DJ, et al. Exploring the role of a large language model on carpal tunnel syndrome management: an observation study of ChatGPT. J Hand Surg Am. Oct 2023;48(10):1025-1033. [doi: 10.1016/j.jhsa.2023.07.003] [Medline: 37530687]
50. Shrestha N, Shen Z, Zaidat B, Duey AH, Tang JE, Ahmed W, et al. Performance of ChatGPT on NASS clinical guidelines for the diagnosis and treatment of low back pain: a comparison study. Spine (Phila Pa 1976). May 01, 2024;49(9):640-651. [doi: 10.1097/BRS.0000000000004915] [Medline: 38213186]
51. Sosa BR, Cung M, Suhardi VJ, Morse K, Thomson A, Yang HS, et al. Capacity for large language model chatbots to aid in orthopedic management, research, and patient queries. J Orthop Res. Jun 21, 2024;42(6):1276-1282. [doi: 10.1002/jor.25782] [Medline: 38245845]
52. Stroop A, Stroop T, Zawy Alsofy S, Nakamura M, Möllmann F, Greiner C, et al. Large language models: are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery? Eur Spine J. Oct 11, 2023. (forthcoming). [doi: 10.1007/s00586-023-07975-z] [Medline: 37821602]
53. Suthar PP, Kounsal A, Chhetri L, Saini D, Dua SG. Artificial intelligence (AI) in radiology: a deep dive into ChatGPT 4.0's accuracy with the American Journal of Neuroradiology's (AJNR) "case of the month". Cureus. Aug 2023;15(8):e43958. [FREE Full text] [doi: 10.7759/cureus.43958] [Medline: 37746411]
54. Taylor WL4, Cheng R, Weinblatt AI, Bergstein V, Long WJ. An artificial intelligence chatbot is an accurate and useful online patient resource prior to total knee arthroplasty. J Arthroplasty. Aug 2024;39(8S1):S358-S362. [doi: 10.1016/j.arth.2024.02.005] [Medline: 38350517]
55. Temel MH, Erden Y, Bağcıer F. Information quality and readability: ChatGPT's responses to the most common questions about spinal cord injury. World Neurosurg. Jan 2024;181:e1138-e1144. [doi: 10.1016/j.wneu.2023.11.062] [Medline: 38000671]
56. 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. Aug 2024;33(8):e429-e437. [doi: 10.1016/j.jse.2023.11.014] [Medline: 38182023]
57. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep. Nov 17, 2023;13(1):20159. [FREE Full text] [doi: 10.1038/s41598-023-47500-2] [Medline: 37978240]
58. Warren EJ, Hurley ET, Park CN, Crook BS, Lorentz S, Levin JM, et al. Evaluation of information from artificial intelligence on rotator cuff repair surgery. JSES Int. Jan 2024;8(1):53-57. [FREE Full text] [doi: 10.1016/j.jseint.2023.09.009] [Medline: 38312282]
59. Wilhelm TI, Roos J, Kaczmarczyk R. Large language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res. Oct 30, 2023;25:e49324. [FREE Full text] [doi: 10.2196/49324] [Medline: 37902826]

60. Wright BM, Bodnar MS, Moore AD, Maseda MC, Kucharik MP, Diaz CC, et al. Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients? Bone Jt Open. Feb 15, 2024;5(2):139-146. [FREE Full text] [doi: 10.1302/2633-1462.52.BJO-2023-0113.R1] [Medline: 38354748]
61. Yang F, Yan D, Wang Z. Large-scale assessment of ChatGPT's performance in benign and malignant bone tumors imaging report diagnosis and its potential for clinical applications. J Bone Oncol. Feb 2024;44:100525. [FREE Full text] [doi: 10.1016/j.jbo.2024.100525] [Medline: 38314324]
62. Yang J, Ardavanis KS, Slack KE, Fernando ND, Della Valle CJ, Hernandez NM. Chat generative pretrained transformer (ChatGPT) and bard: artificial intelligence does not yet provide clinically supported answers for hip and knee osteoarthritis. J Arthroplasty. May 2024;39(5):1184-1190. [doi: 10.1016/j.arth.2024.01.029] [Medline: 38237878]
63. Yapar D, Demir Avcı Y, Tokur Sonuvar E, Eğerci Ö, Yapar A. ChatGPT's potential to support home care for patients in the early period after orthopedic interventions and enhance public health. Jt Dis Relat Surg. Jan 01, 2024;35(1):169-176. [FREE Full text] [doi: 10.52312/jdrs.2023.1402] [Medline: 38108178]
64. Zhou Y, Moon C, Szatkowski J, Moore D, Stevens J. Evaluating ChatGPT responses in the context of a 53-year-old male with a femoral neck fracture: a qualitative analysis. Eur J Orthop Surg Traumatol. Feb 2024;34(2):927-955. [FREE Full text] [doi: 10.1007/s00590-023-03742-4] [Medline: 37776392]
65. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass section 1 of the Fellowship of the Royal College of Surgeons (trauma and orthopaedics) examination? Postgrad Med J. Sep 21, 2023;99(1176):1110-1114. [doi: 10.1093/postmj/qgad053] [Medline: 37410674]
66. Ghanem D, Covarrubias O, Raad M, LaPorte D, Shafiq B. ChatGPT performs at the level of a third-year orthopaedic surgery resident on the orthopaedic in-training examination. JBJS Open Access. 2023;8(4):e23.00103. [FREE Full text] [doi: 10.2106/JBJS.OA.23.00103] [Medline: 38638869]
67. Han Y, Choudhry HS, Simon ME, Katt BM. ChatGPT's performance on the hand surgery self-assessment exam: a critical analysis. J Hand Surg Glob Online. Mar 2024;6(2):200-205. [FREE Full text] [doi: 10.1016/j.jhsg.2023.11.014] [Medline: 38903839]
68. Hofmann HL, Guerra GA, Le JL, Wong AM, Hofmann GH, Mayfield CK, et al. The rapid development of artificial intelligence: GPT-4's performance on orthopedic surgery board questions. Orthopedics. 2024;47(2):e85-e89. [doi: 10.3928/01477447-20230922-05] [Medline: 37757748]
69. Jain N, Gottlich C, Fisher J, Campano D, Winston T. Assessing ChatGPT's orthopedic in-service training exam performance and applicability in the field. J Orthop Surg Res. Jan 03, 2024;19(1):27. [FREE Full text] [doi: 10.1186/s13018-023-04467-0] [Medline: 38167093]
70. Kung J, Marshall C, Gauthier C, Gonzalez T, Jackson JB3. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access. 2023;8(3):e23.00056. [FREE Full text] [doi: 10.2106/JBJS.OA.23.00056] [Medline: 37693092]
71. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res. Aug 01, 2023;481(8):1623-1630. [doi: 10.1097/CORR.0000000000002704] [Medline: 37220190]
72. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg. Dec 01, 2023;31(23):1173-1179. [FREE Full text] [doi: 10.5435/JAAOS-D-23-00396] [Medline: 37671415]
73. Ozdag Y, Hayes DS, Makar GS, Manzar S, Foster BK, Shultz MJ, et al. Comparison of artificial intelligence to resident performance on upper-extremity orthopaedic in-training examination questions. J Hand Surg Glob Online. Mar 2024;6(2):164-168. [FREE Full text] [doi: 10.1016/j.jhsg.2023.10.013] [Medline: 38903829]
74. Rizzo MG, Cai N, Constantinescu D. The performance of ChatGPT on orthopaedic in-service training exams: a comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education. J Orthop. Apr 2024;50:70-75. [doi: 10.1016/j.jor.2023.11.056] [Medline: 38173829]
75. Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. Surgeon. Oct 2023;21(5):263-266. [doi: 10.1016/j.surge.2023.07.001] [Medline: 37517980]
76. Traoré SY, Goetsch T, Muller B, Dabbagh A, Liverneaux PA. Is ChatGPT able to pass the first part of the European Board of Hand Surgery diploma examination? Hand Surg Rehabil. Sep 2023;42(4):362-364. [doi: 10.1016/j.hansur.2023.06.005] [Medline: 37353199]
77. Gill B, Bonamer J, Kuechly H, Gupta R, Emmert S, Kurkowski S, et al. ChatGPT is a promising tool to increase readability of orthopedic research consents. J Orthop Trauma Rehab. Jan 22, 2024. [doi: 10.1177/22104917231208212]
78. Hakam HT, Prill R, Korte L, Lovreković B, Ostojić M, Ramadanov N, et al. Human-written vs AI-generated texts in orthopedic academic literature: comparative qualitative analysis. JMIR Form Res. Feb 16, 2024;8:e52164. [FREE Full text] [doi: 10.2196/52164] [Medline: 38363631]
79. Kacena MA, Plotkin LI, Fehrenbacher JC. The use of artificial intelligence in writing scientific review articles. Curr Osteoporos Rep. Feb 2024;22(1):115-121. [FREE Full text] [doi: 10.1007/s11914-023-00852-0] [Medline: 38227177]

80. Lawrence KW, Habibi AA, Ward SA, Lajam CM, Schwarzkopf R, Rozell JC. Human versus artificial intelligence-generated arthroplasty literature: a single-blinded analysis of perceived communication, quality, and authorship source. Int J Med Robot. Feb 13, 2024;20(1):e2621. [doi: 10.1002/rcs.2621] [Medline: 38348740]
81. Lotz JC, Ropella G, Anderson P, Yang Q, Hedderich MA, Bailey J, et al. An exploration of knowledge-organizing technologies to advance transdisciplinary back pain research. JOR Spine. Dec 2023;6(4):e1300. [FREE Full text] [doi: 10.1002/jsp2.1300] [Medline: 38156063]
82. Methnani J, Latiri I, Dergaa I, Chamari K, Ben Saad H. ChatGPT for sample-size calculation in sports medicine and exercise sciences: a cautionary note. Int J Sports Physiol Perform. Oct 01, 2023;18(10):1219-1223. [doi: 10.1123/ijspp.2023-0109] [Medline: 37536678]
83. Nazzal MK, Morris AJ, Parker RS, White FA, Natoli RM, Fehrenbacher JC, et al. Using AI to write a review article examining the role of the nervous system on skeletal homeostasis and fracture healing. Curr Osteoporos Rep. Feb 13, 2024;22(1):217-221. [FREE Full text] [doi: 10.1007/s11914-023-00854-y] [Medline: 38217755]
84. Sanii RY, Kasto JK, Wines WB, Mahylis JM, Muh SJ. Utility of artificial intelligence in orthopedic surgery literature review: a comparative pilot study. Orthopedics. 2024;47(3):e125-e130. [FREE Full text] [doi: 10.3928/01477447-20231220-02] [Medline: 38147494]
85. Zaidat B, Lahoti YS, Yu A, Mohamed KS, Cho SK, Kim JS. Artificially intelligent billing in spine surgery: an analysis of a large language model. Global Spine J. Dec 26, 2023:21925682231224753. [FREE Full text] [doi: 10.1177/21925682231224753] [Medline: 38147047]
86. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. J Med Internet Res. Jul 15, 2020;22(7):e19274. [FREE Full text] [doi: 10.2196/19274] [Medline: 32673234]
87. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. Jun 01, 2023;183(6):589-596. [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
88. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. Feb 1999;53(2):105-111. [FREE Full text] [doi: 10.1136/jech.53.2.105] [Medline: 10396471]
89. Dennler C, Bauer DE, Scheibler AG, Spirig J, Götschi T, Fürnstahl P, et al. Augmented reality in the operating room: a clinical feasibility study. BMC Musculoskelet Disord. May 18, 2021;22(1):451. [FREE Full text] [doi: 10.1186/s12891-021-04339-w] [Medline: 34006234]
90. Verhey JT, Haglin JM, Verhey EM, Hartigan DE. Virtual, augmented, and mixed reality applications in orthopedic surgery. Int J Med Robot. Apr 2020;16(2):e2067. [doi: 10.1002/rcs.2067] [Medline: 31867864]
91. Li Z, Jiang S, Song X, Liu S, Wang C, Hu L, et al. Collaborative spinal robot system for laminectomy: a preliminary study. Neurosurg Focus. Jan 2022;52(1):E11. [doi: 10.3171/2021.10.FOCUS21499] [Medline: 34973664]
92. Li Z, Wang C, Song X, Liu S, Zhang Y, Jiang S, et al. Accuracy evaluation of a novel spinal robotic system for autonomous laminectomy in thoracic and lumbar vertebrae: a cadaveric study. J Bone Joint Surg Am. Jun 21, 2023;105(12):943-950. [doi: 10.2106/JBJS.22.01320] [Medline: 36943914]
93. Kim JS, Vivas A, Arvind V, Lombardi J, Reidler J, Zuckerman SL, et al. Can natural language processing and artificial intelligence automate the generation of billing codes from operative note dictations? Global Spine J. Sep 2023;13(7):1946-1955. [FREE Full text] [doi: 10.1177/21925682211062831] [Medline: 35225694]
94. ChatGPT. URL: https://chatgpt.com [accessed 2024-04-29]

Abbreviations

AI: artificial intelligence
CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence
FKGL: Flesch-Kincaid grade level
GPT: generative pretrained transformer
LLM: large language model
NLP: natural language processing
OITE: Orthopaedic Surgery In-Training Examination
PGY: postgraduate year
PLM: pretrained language model
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by G Eysenbach, A Mavragani; submitted 17.04.24; peer-reviewed by J Wu, X Liu, D-G Chang; comments to author 20.06.24; revised version received 01.08.24; accepted 11.09.24; published 15.11.24

Please cite as:
Zhang C, Liu S, Zhou X, Zhou S, Tian Y, Wang S, Xu N, Li W
Examining the Role of Large Language Models in Orthopedics: Systematic Review
J Med Internet Res 2024;26:e59607
URL: https://www.jmir.org/2024/1/e59607
doi: 10.2196/59607
PMID: