

Review

Applications of Large Language Models in Medical Research: From Systematic Reviews to Clinical Studies

Eun Jeong Gong ^{1,2,3} , Chang Seok Bang ^{1,2,3,*}  and Yong Seok Shin ¹¹ Department of Internal Medicine, Hallym University College of Medicine, Chuncheon 24252, Republic of Korea² Institute for Liver and Digestive Diseases, Hallym University, Chuncheon 24252, Republic of Korea³ Institute of New Frontier Research, Hallym University College of Medicine, Chuncheon 24252, Republic of Korea

* Correspondence: csbang@hallym.ac.kr; Tel.: +82-33-240-5821; Fax: +82-33-241-8064

Abstract

Background: Large Language Models (LLMs) are reshaping medical research workflows. **Objective:** This narrative review synthesizes evidence on LLM applications across systematic reviews, scientific writing, and clinical research. **Methods:** We reviewed literature from 2023–2025 examining LLM applications in medical research, identified through PubMed, Scopus, Web of Science, arXiv, medRxiv, and Google Scholar. Studies reporting empirical findings, methodological evaluations, or systematic analyses of LLM applications were included; editorials and commentaries without empirical data were excluded. **Results:** In systematic reviews, LLMs achieve 80–94% data extraction accuracy and 40% reduction in screening workload, but show only slight-to-moderate agreement ($\kappa = 0.16$ – 0.43) in risk-of-bias assessment. In scientific writing, hallucination rates of 47–55% for fabricated references and over 90% prevalence of demographic bias require rigorous verification. For clinical research, LLMs assist with statistical coding and protocol development but require human validation. Critically, excessive reliance on automated tools may cause cognitive offloading that compromises analytical capabilities. **Conclusions:** LLMs are powerful but unstable tools requiring constant verification. Success depends on maintaining human-in-the-loop approaches that preserve critical thinking while leveraging AI efficiency.

Keywords: large language models; ChatGPT; GPT-4; systematic review; medical research; artificial intelligence; prompt engineering; evidence synthesis

1. Introduction

Since ChatGPT's public release in November 2022, large language models (LLMs)—transformer-based neural networks capable of understanding context, following complex instructions, and generating human-like text—have generated substantial interest within the medical research community [1–5]. These models, trained on massive datasets and containing billions of parameters, offer potential solutions to longstanding challenges in medical research, particularly the labor-intensive nature of evidence synthesis and the exponential growth of medical literature [6,7].

The systematic review (SR) process presents significant challenges. Reviews require an average of 67.3 weeks from protocol registration to publication, with research teams investing hundreds of person-hours in screening thousands of abstracts, extracting data, and synthesizing findings [8]. With PubMed adding over 1.5 million citations annually, the volume of literature has exceeded human capacity for full synthesis using traditional



Academic Editor: Ning Deng

Received: 10 February 2026

Revised: 11 March 2026

Accepted: 14 March 2026

Published: 20 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

methods [9]. Given these challenges, LLMs may help researchers manage the growing volume of literature, though human expertise remains essential for quality control [10].

Recent surveys of mental health researchers found that 69.5% have used LLMs, though fewer than 15% employ them for complex analytical tasks such as data analysis or study design [11]. This gap between interest and implementation stems from multiple factors, including uncertainty about best practices, concerns about accuracy, a lack of institutional guidance, and limited practical implementation frameworks [12]. Early implementations have shown mixed results, highlighting the need for evidence-based guidance on appropriate use cases, validation methods, and ethical considerations.

The phenomenon of hallucination, where LLMs generate plausible but entirely fabricated information, poses significant risks in medical contexts where accuracy directly affects patient care decisions [5,13]. Additionally, questions about transparency, reproducibility, and accountability challenge traditional notions of authorship and scientific responsibility [14]. These concerns have prompted major medical journals and professional organizations to develop guidelines for artificial intelligence (AI) use in research, though standards remain heterogeneous and evolving [15].

This narrative review assesses current evidence on LLM applications across three domains: SR methodology, narrative review composition, and clinical research applications, identifying strengths, limitations, and optimal integration strategies.

Although numerous reviews have addressed LLM applications in medicine between 2023 and 2025, existing publications tend to focus on individual domains: clinical applications [2,16], SR automation [17,18], scientific writing [19], or clinical trials [20]. No existing review integrates all three domains—SR methodology, scientific writing, and clinical research applications—under a unified, researcher-centric framework spanning the entire research lifecycle. Our review addresses this gap by providing an integrative synthesis that connects these domains and introduces several novel conceptual contributions, including the “cognitive offloading paradox” with supporting neuroscience evidence, the concepts of “never-skilling” and “mis-skilling” in research training, and the “paywall blind spot” as a systematic limitation of LLM training data for evidence synthesis.

We chose a narrative rather than a systematic review format for several methodological reasons. The LLM field evolves on a timescale of weeks, making systematic review protocols impractical for capturing current developments [21]. The literature we synthesize spans computational experiments, clinical pilots, theoretical analyses, and conference proceedings—sources that cannot be meaningfully combined under a single PICO framework. Furthermore, our integrative purpose—connecting disparate domains under a unified framework—requires the interpretive flexibility that narrative synthesis affords [22,23]. The SANRA (Scale for the Assessment of Narrative Review Articles) framework [24] provides the appropriate quality assessment standard for this format.

1.1. Search Strategy and Study Selection

This narrative review was conducted by searching PubMed, Scopus, Web of Science, arXiv, medRxiv, and Google Scholar for articles published between January 2023 and May 2025. The primary search terms included “large language model,” “ChatGPT,” “GPT-4,” “LLM,” and “generative AI,” combined with domain-specific terms including “systematic review,” “meta-analysis,” “medical research,” “clinical trial,” “scientific writing,” “data extraction,” and “evidence synthesis.” Studies were included if they reported original empirical findings, methodological evaluations, or systematic analyses of LLM applications in medical research contexts. Editorials, commentaries, and opinion pieces without empirical data were excluded unless they provided novel conceptual frameworks. Reference lists of included studies and relevant review articles were manually screened to identify

additional eligible publications. As a narrative review, we aimed for comprehensive but not exhaustive coverage, prioritizing studies with empirical performance data and those reporting validation metrics.

1.2. Large Language Models in Systematic Reviews

Literature Search Strategy Generation

The development of thorough search strategies represents the foundational step in SR methodology, requiring expertise in controlled vocabularies, Boolean operators, and database-specific syntax. Recent investigations have examined whether LLMs can assist in generating these complex queries, with mixed results that highlight both potential utility and significant limitations.

Wang et al. conducted an early systematic evaluation of ChatGPT's ability to generate Boolean queries for SR literature searches [25]. Their findings revealed a trade-off: LLM-generated queries demonstrated high precision but marked reduced recall compared to expert-crafted strategies. This precision-recall imbalance poses particular concerns for SRs, where comprehensive retrieval is paramount to minimizing selection bias. The application of guided chain-of-thought (CoT) prompting improved F1 scores dramatically from 0.077 to 0.517, suggesting that sophisticated prompt engineering can partially mitigate these limitations [25].

Subsequent research has further characterized the strengths and weaknesses of LLM-generated search strategies. Yu and colleagues, applying PRESS (Peer Review of Electronic Search Strategies) guidelines for evaluation, found that GPT-4 significantly outperformed GPT-3.5 in search strategy generation, particularly in the appropriate inclusion of Medical Subject Headings (MeSH) terms [26]. However, systematic deficiencies persist across models. An evidence summary by Parisi and Sutton identified that LLM-generated strategies frequently fail to incorporate synonymous entry terms, miss clinical practice jargon, incorrectly group acronyms, insert unjustified date limitations, and, critically, omit validated study design filters for identifying randomized controlled trials (RCTs) [27].

A critical limitation is that LLMs cannot directly execute searches across bibliographic databases such as PubMed, Embase, or the Cochrane Library. They can only generate query strings that human researchers should then adapt and run in each database's specific interface. Furthermore, LLM-generated strategies rarely employ advanced search techniques such as truncation, wildcards, or proximity operators that information specialists routinely use to maximize retrieval [27]. Given these constraints, current evidence suggests that LLMs may serve as useful starting points for search strategy development but cannot replace the expertise of trained medical librarians or information specialists. Human validation remains essential before implementing any LLM-generated search strategy in an SR protocol. Importantly, the lack of standardized benchmarks for evaluating LLM-generated search strategies against expert-crafted ones limits the generalizability of current findings, and most existing evaluations assess only a single LLM version, making it difficult to track performance trajectories across rapidly evolving model iterations.

1.3. Literature Screening and Study Selection

The application of LLMs to literature screening represents one of the most mature use cases in SR methodology. Traditional screening, which requires assessing thousands of abstracts against predetermined inclusion criteria, consumes approximately 30–40% of total review time while being prone to human error, fatigue, and inconsistency [28]. The promise of LLM assistance lies not just in time savings but also in the potential to apply criteria more consistently and to process volumes of literature that would overwhelm human reviewers [29].

Studies examining LLM performance in the literature screening showed mixed results. Issaiy et al. tested ChatGPT on 1198 radiology abstracts from three subfields and reported 95% sensitivity [30]. The model correctly excluded over 50% of irrelevant citations without missing any eligible studies. Structured prompts outperformed narrative instructions. This suggests LLMs handle step-by-step decision-making more effectively than overall assessments [30]. The development of more advanced approaches has yielded improved results. The LARS-GPT system developed by Cai et al. uses a more sophisticated dual-phase screening approach [31]. It first identifies studies with high confidence as clearly irrelevant, then flags borderline cases for human review. This strategy achieved recall rates above 0.9 while reducing human screening workload by 40% [31].

Significant limitations have also been identified. Khraisha Q and colleagues conducted perhaps the most thorough evaluation to date, testing GPT-4 across multiple languages and publication types [32]. Their findings revealed dramatic performance degradation for non-English texts, with data extraction sensitivity dropping from 75% for English articles to just 36% for non-English publications [32]. This language bias represents a critical limitation for comprehensive SRs that aim to minimize it, particularly given that important research from non-English-speaking countries may be systematically excluded if reviewers rely too heavily on LLM assistance.

Prompt engineering significantly affects screening performance, though effects remain unpredictable. Kohandel Gargari et al. found inherent trade-offs between sensitivity and specificity, with some presumably beneficial modifications (e.g., expert role assignment) paradoxically degrading accuracy [33]. Prompt selection should, therefore, align with review priorities—sensitivity for comprehensive reviews, specificity when workload reduction is paramount [33].

Model parameters such as temperature settings influence output consistency, though their impact on screening accuracy remains underexplored. Most screening studies use temperature = 0 for reproducibility, as this eliminates response variation between identical queries [34]. However, recent evidence suggests that temperatures between 0.0 and 1.0 produce statistically equivalent classification accuracy, with significant degradation occurring only above 1.5 [35]. The practical recommendation is to set temperature = 0 and document this choice explicitly for methodological transparency.

Table 1 summarizes recent evidence on LLM performance in SR screening. As these studies demonstrate, performance varies considerably across models, with distinct trade-offs between sensitivity and specificity [36]. When combining outputs using an ensemble method that included citations flagged by any model, sensitivity improved while specificity decreased, suggesting that strategically integrating multiple LLMs can enhance screening coverage at the expense of increased false positives requiring human review [36]. Similarly, multi-agent collaborative frameworks—where distinct LLMs provide initial analyses, review each other's outputs, and converge on consensus decisions through majority voting—have shown promising results in diagnostic accuracy tasks, with one study reporting 98% accuracy among the top three diagnoses compared to 71–96% for individual models [37]. The concept of an “LLM Council,” popularized by Karpathy, extends this approach by having multiple models anonymously evaluate and rank each other's responses before a designated “chairman” model synthesizes the final answer, though this method may favor verbosity over conciseness [38]. These complementary strengths of different models suggest significant potential for multi-model approaches in high-stakes SRs where missing relevant studies carry significant consequences [39].

Table 1. Performance of LLMs in systematic review screening.

Study	Model(s)	Number of Studies	Sensitivity	Specificity	Key Findings
Oami et al. 2025 [36]	GPT-4o	16,669	0.85	0.97	Higher specificity, lower sensitivity
	Gemini 1.5 Pro		0.94	0.85	Higher sensitivity, lower specificity
	Claude 3.5 Sonnet		0.94	0.80	Higher sensitivity, lowest specificity
	Llama 3.3 70B		0.88	0.93	Trade-off: sensitivity vs. specificity
	Ensemble (OR rule)		Improved	Decreased	Trade-off: sensitivity vs. specificity
Matsui et al. 2024 [34]	GPT-4 (3-layer)	4527	0.81–0.88	0.86–1.00	Layered screening approach effective
	GPT-3.5 (3-layer)		0.69–0.75	0.95–0.98	Lower sensitivity than GPT-4
Guo et al. 2024 [39]	GPT-3.5/GPT-4	24,307	0.76	0.91	No pretraining required
Kohandel Gargari et al. 2024 [33]	GPT-3.5 Turbo	200	0.38–0.69	0.25–0.85	Prompt structure critical; trade-offs inevitable
Cai et al. 2025 [31]	LARS-GPT (multi-LLM)	N/A	>0.90	N/A	40% workload reduction with dual-phase approach
Khraisha et al. 2024 [32]	GPT-4	2421	0.75 (English)	N/A	Sensitivity drops to 0.36 for non-English texts

LLM, large language model; GPT, Generative Pre-trained Transformer; N/A, not applicable.

1.4. Data Extraction and Evidence Synthesis

The transition from screening to data extraction represents a major increase in complexity, requiring LLMs not only to identify relevant information within lengthy and often poorly structured texts but also to accurately transcribe specific values, understand statistical presentations, and maintain consistency across heterogeneous reporting styles. The challenge is compounded by the variety of data types encountered in medical research, from simple demographic information to complex statistical analyses and subtle clinical outcomes [40].

Recent evaluations of LLM performance in data extraction reveal both impressive capabilities and concerning limitations (Table 2) [41,42]. Single LLM approaches achieve approximately 80% overall accuracy, with variation across domains (82% in clinical, 72% in social science studies) [41]. Simple data elements, such as participant characteristics, interventions, and study locations, are typically extracted with high accuracy (80–90%). Meanwhile, more complex information, such as outcomes, causal inference methods, and study design, shows notably lower performance [41]. A collaborative dual-LLM approach using GPT-4-turbo and Claude-3-Opus significantly improves accuracy to 94% when both models agree, while reducing hallucination rates from approximately 2.5% to 0.25% [42].

However, performance drops markedly for non-English texts, with sensitivity falling to 36% [32].

Table 2. LLM performance in data extraction for systematic reviews.

Extraction Approach	Performance	Key Findings
Overall extraction (single LLM) [41]	Accuracy ~80%	82% clinical, 80% animal, 72% social science studies
PICO elements [41]	Accuracy >80% (P, I, C), lower for O	Participants/Intervention well-extracted; Outcomes challenging
Collaborative dual-LLM (concordant) [42]	Accuracy 94%	GPT-4-turbo + Claude-3-Opus agreement; hallucination rate 0.25%
Single LLM (discordant cases) [42]	Accuracy 41–50%	GPT-4-turbo 41%, Claude-3-Opus 50%; hallucination rate ~2.5%
Non-English texts [32]	Sensitivity 36%	Significant performance drop in non-English literature
PDF-dependent extraction [43,44]	68.8–100%	Automated PDF parsing 68.8% vs. manual text selection 100%

LLM, large language model; GPT, Generative Pre-trained Transformer; P, participants; I, interventions; C, comparisons; O, outcomes.

Document format significantly affects extraction accuracy. Portable document format (PDF) parsing quality was identified as the primary determinant of success, with GPT-4 achieving only 68.8% accuracy using automated PDF parsing versus 100% with manually selected text [43,44]. While multimodal vision-language models can process documents without separate optical character recognition (OCR) steps, current models still exhibit notable limitations requiring careful validation [45–47].

Language barriers present another significant challenge, with GPT-4’s data extraction sensitivity dropping from 75% for English articles to 36% for non-English publications [32]. This has important implications for SRs aiming for coverage of global literature. Teams must either restrict their reviews to English-language publications—potentially introducing language bias—or maintain significant human resources for processing non-English literature.

Several research groups have developed structured extraction protocols to improve performance. Khan et al. implemented a collaborative two-LLM approach using GPT-4-turbo and Claude-3-Opus in parallel, with consensus between models serving as a quality filter that substantially outperformed single-model extraction [42]. Pre-specified extraction schemas with standardized prompts have shown 95–97% test–retest reliability across multiple extraction rounds [44]. Task chunking—separating complex extractions into smaller independent tasks—has been specifically recommended for improving numerical data accuracy [48].

The phenomenon of hallucinated data in extraction tasks represents a critical concern. Motzfeldt Jensen and colleagues identified a 5.2% false-data rate where LLMs fabricated outcome values rather than reporting them as missing [49]. When information was not explicitly reported in papers, reproducibility decreased from 94.1% to 77.2% as models attempted to infer data based on training patterns. Such systematic errors would directly corrupt meta-analytic effect estimates, emphasizing the importance of human verification, particularly for outcome data and statistical results, rather than complete automation of the extraction process. These limitations are particularly concerning, given that most evaluation studies have been conducted in controlled settings and published in well-structured papers; real-world extraction from diverse source types (conference abstracts,

government reports, gray literature) is likely to yield substantially lower accuracy than reported benchmarks suggest.

1.5. Risk-of-Bias Assessment

Risk-of-bias assessment is one of the most challenging applications of LLMs in SRs, requiring not only information extraction but also careful judgment about study conduct and reporting quality. The subjective nature of many bias assessments, combined with the need to infer information that may not be explicitly stated, creates unique challenges that test the limits of current LLM capabilities. Recent studies evaluating LLM performance on standardized risk-of-bias tools have revealed notable limitations, with agreement between AI and human reviewers remaining modest across multiple evaluation frameworks [12].

Recent studies have evaluated LLM performance in risk-of-bias assessment using the Cochrane RoB 2 tool for RCTs. Pitre and colleagues assessed ChatGPT-4 against Cochrane author judgments across 157 RCTs from 34 reviews, finding only slight agreement with Cohen's κ of 0.16 for overall risk-of-bias assessment [50]. Similarly, Kuitunen et al. evaluated ChatGPT-4o on 100 RCTs from high-impact journals and reported slight agreement for the overall assessment ($\kappa = 0.24$) and the randomization domain ($\kappa = 0.31$), with no agreement to poor agreement in the other domains [51]. A subsequent study by the same group analyzing 61 neonatal RCTs reported moderate overall agreement ($\kappa = 0.43$) [52]. The available evidence suggests variable performance across different bias domains, with LLMs showing better performance on objective criteria, such as allocation concealment ($\kappa = 0.73$), compared to subjective judgments, such as incomplete outcome data assessment ($\kappa = -0.03$) [52].

These performance gaps likely reflect fundamental limitations in how current LLMs process implicit methodological information—risk-of-bias judgments frequently require inferring study conduct from what is not reported, a task that challenges models trained primarily on explicit textual patterns [53]. Nevertheless, several promising strategies for improvement have emerged. Structured prompting strategies incorporating domain-specific guidelines have demonstrated substantially higher accuracy, with one study achieving 89.5% correct assessment rates using well-designed prompts compared to lower performance with basic approaches [54]. Human-in-the-loop workflows that combine LLM screening with expert verification appear particularly promising, reducing assessment time by over 90% while maintaining accuracy comparable to conventional methods [55]. Future advances may require domain-specific fine-tuning on large corpora of completed Cochrane assessments, development of retrieval-augmented approaches that can access methodological guidance documents, and better handling of the contextual reasoning required for subjective domains. Until such advances materialize, LLMs are best positioned as assistive tools that accelerate the mechanical aspects of risk-of-bias assessment while preserving human judgment for the complex interpretations that define rigorous evidence synthesis.

1.6. Large Language Models in Narrative Review Writing

Augmenting Scientific Writing

The application of LLMs to scientific writing extends far beyond simple grammar correction, encompassing advanced capabilities in structure development, argumentation refinement, and stylistic consistency. For narrative reviews, which require synthesizing diverse literature into coherent narratives while maintaining author voice and critical perspective, LLMs offer both promising opportunities and unique challenges [56].

Non-native speakers particularly benefit from LLM assistance [57]. With over 80% of indexed scientific journals publishing in English, non-native speakers face significant disadvantages in manuscript preparation, peer review communication, and career advance-

ment [57]. LLMs can help bridge this gap by improving grammatical accuracy, sentence structure, and overall clarity, potentially democratizing access to international publication venues [5,58,59].

The process of scientific writing involves multiple layers of complexity that LLMs address through different mechanisms. At the most basic level, these models excel at correcting grammatical errors and improving sentence structure, tasks that require understanding of both linguistic rules and scientific conventions [58]. Their capabilities extend to higher-order concerns such as logical flow and argumentative coherence, though their effectiveness in these areas depends heavily on the quality of human guidance and the specific requirements of the writing task [60].

Table 3 summarizes the common issues identified in LLM-generated scientific text along with detection methods and mitigation strategies. Hallucination poses serious risks in scientific writing. Analysis of LLM-generated medical texts reveals high rates of fabricated references. Bhattacharyya et al. found that among 115 references generated by ChatGPT-3.5, 47% were completely fabricated, 46% were authentic but contained inaccuracies, and only 7% were both authentic and accurate [48]. Walters and Brainard reported even more variation across model versions: GPT-3.5 produced 55% fabricated citations compared to 18% for GPT-4, demonstrating that newer models show improvement but remain unreliable [61]. These fabricated references often appear entirely plausible, making detection challenging without systematic verification [62]. Additionally, LLMs may misrepresent study findings when paraphrasing, introduce inaccurate statistical claims, or generate content that sounds authoritative but lacks a factual basis [63].

Table 3. Common issues in LLM-generated scientific text.

Issue Type	Estimated Frequency	Detection Method	Mitigation Strategy
Fabricated references	18–55% (model-dependent) [48]	Database verification	Verify citations
Inaccurate citations	24–46% [48,61]	Original source check	Verify bibliographic details
Incorrect PMID	93% of papers [48]	PubMed verification	Cross-check all PMIDs
Oversimplification	Common (not quantified) [64]	Expert review	Maintain technical precision
Lost nuance	Common (not quantified) [65]	Domain expert check	Preserve complexity
Style homogenization	Common (not quantified)	AI detection tools, stylometric analysis	Maintain author voice, iterative refinement

LLM, large language model; PMID, PubMed Identifier.

The development of verification protocols has become essential for the safe implementation of LLM-assisted writing. These protocols typically involve multiple layers of checking, beginning with automated verification of references against bibliographic databases, followed by validation of numerical claims against original sources, and culminating in expert review of technical accuracy [63]. While time-consuming, such verification is necessary to maintain scientific integrity and prevent the propagation of errors through the literature. A critical concern is that the verification burden may paradoxically negate the time savings offered by LLM-assisted writing, particularly for non-expert users who

may lack the domain knowledge to identify subtle inaccuracies or misrepresentations of study findings. The net efficiency gain, therefore, depends heavily on the user's baseline expertise, a factor that remains underexplored in the existing literature.

The challenge of maintaining technical accuracy while improving readability represents a persistent tension in LLM-assisted writing. Models may oversimplify complex concepts in pursuit of clarity, potentially altering meaning or omitting important nuances [64]. This tendency requires careful human oversight to ensure that improvements in readability do not compromise scientific precision. Researchers must remain vigilant for subtle changes in meaning that could affect the interpretation of findings or the validity of conclusions.

Iterative, multi-stage prompting outperforms single-prompt generation for writing tasks. The Self-Refine approach—where LLMs generate output, provide self-feedback, then iteratively refine—improved quality by approximately 20% compared to one-step generation [66]. Similarly, prompt chaining (sequential drafting, critiquing, and refining) outperformed single-stepwise prompts in 77 of 100 text summarization evaluations, as stepwise prompts often produced “simulated refinement”, where models intentionally introduced errors to subsequently correct them [67]. These findings support decomposing writing tasks into sequential stages to allow human verification at each step.

1.7. Literature Synthesis and Thematic Analysis

The synthesis of disparate research findings into coherent narratives represents one of the most intellectually demanding aspects of review writing. LLMs show strong capabilities in identifying patterns across large bodies of literature, though their effectiveness depends critically on how they are deployed and supervised [65]. The ability to process and synthesize information from multiple sources simultaneously enables the identification of connections that might not be apparent when reading studies sequentially.

Studies comparing LLM-generated analyses with those produced by experienced researchers show considerable overlap in identified themes, though with important differences in depth and point [68]. LLMs tend to identify surface-level patterns effectively but may miss subtle theoretical connections or methodological implications that experienced researchers would recognize. They also tend to impose coherence where none exists, potentially obscuring genuine controversies or contradictions in the literature.

The iterative refinement of narratives through human-AI collaboration appears more effective than either approach alone. This collaborative model uses the LLM's ability to process large volumes of information while maintaining the human researcher's critical judgment and domain expertise [60]. For multi-author projects, LLMs may assist with harmonizing writing styles, standardizing terminology, and maintaining consistency across sections, though empirical validation of these applications remains limited.

1.8. Large Language Models in Clinical Research and Data Analysis

Statistical Programming and Analysis

The application of LLMs to statistical programming represents a particularly promising yet complex domain. Recent evaluations demonstrate that LLM-generated statistical code achieves 32–93% accuracy depending on prompt specificity and task complexity [69]. For descriptive statistics, LLMs achieve near-perfect accuracy, but performance declines substantially for complex analyses requiring assumption verification and appropriate method selection [69,70].

Studies comparing LLM-generated analyses with traditional software (SAS, SPSS, R) reveal consistent results for basic calculations but significant discrepancies for advanced methods [70]. ChatGPT-4 cannot autonomously select appropriate analyses without specific

user instructions, and complex procedures such as Cox regression and MANOVA show notable error rates, including miscalculated degrees of freedom and implausible confidence intervals [71,72]. In survival analysis, LLMs consistently underestimate required sample sizes due to systematic errors in applying statistical formulas. Meanwhile, meta-analyses show high variability and inappropriate model selection based solely on heterogeneity thresholds [71].

The development of statistical analysis plans (SAPs) through LLM assistance shows preliminary promise, with case studies demonstrating acceptable SAP generation within 15 min [73]. However, rigorous validation comparing LLM-generated SAPs with those from human biostatisticians remains lacking. A particular risk is that researchers without strong statistical backgrounds may accept LLM-generated analyses uncritically, potentially propagating errors in assumption checking, model selection, and result interpretation that could compromise study conclusions. Table 4 summarizes key considerations for LLM-assisted statistical programming based on empirical evidence [69,71,74–77].

Table 4. Key considerations for LLM-assisted statistical programming.

Consideration	Challenge	Evidence	Recommendation
Assumption checking	Often omitted without explicit prompting; 43.8% accuracy with basic prompts [69]	Fails normality verification, inappropriate test selection [76]	Always verify assumptions manually
Model selection	May choose inappropriate tests; incorrect method selection was the most common error (66%, n = 51 of 77 total errors) [69]	44% of errors involved knowledge recall (wrong test selection, statistical vs. causal method confusion) [77]	Require statistical expertise for selection
Complex designs	Poor performance on hierarchical models, survival analysis, or meta-analysis [71]	R code for survival analysis worked without corrections in 7/10 sessions [71]	Use only for simple analyses initially
Reproducibility	Identical prompts yield different results across sessions [71]	High variability in meta-analysis outputs [71]	Verify across multiple runs
Stochasticity reporting	Stochastic outputs even at temperature = 0; model version changes alter results [74]	Only 15.1% of studies adequately reported stochasticity handling [75]	Document per MI-CLEAR-LLM; use temperature = 0; archive model versions

LLM, large language model.

1.9. Clinical Data Processing

The extraction and processing of clinical data from electronic health records (EHRs) represents an area where LLMs have demonstrated strong capabilities. Recent studies showed that adapted LLMs can outperform medical experts in clinical text summarization, with GPT-4 summaries rated equivalent or superior to expert summaries in 81% of evaluations [78]. LLMs showed promise for various EHR-related tasks, including diagnosis extraction, medication reconciliation, and outcome ascertainment, with GPT-4 producing the highest-quality phenotyping algorithms when generating executable Structured Query Language (SQL) queries for patient identification [79].

The challenge of maintaining privacy while using LLM capabilities has led to the development of specialized approaches for handling sensitive data. These typically involve

complete de-identification before processing, use of locally deployed models that never transmit data externally, and careful audit logging of all data access [80]. Privacy-preserving frameworks using open-source models such as Llama 2 have achieved 100% sensitivity and 96% specificity for clinical information extraction while running entirely on-premises, eliminating the need for cloud data transfer and addressing GDPR/HIPAA compliance concerns [81].

Real-world clinical documentation presents unique challenges that test the limits of LLM capabilities. Clinical text often contains abbreviations, misspellings, negations, and implicit information that models may misinterpret. Recent work demonstrates GPT-4 can achieve 98% accuracy for clinical acronym disambiguation in zero-shot settings, though performance drops for non-English texts, and smaller models frequently produce hallucinations [82].

The temporal complexity of clinical data—with events across multiple encounters—poses additional challenges, though time-aware approaches have shown modest improvements in longitudinal reasoning [83]. For cohort construction and clinical trial matching, LLMs show potential for initial screening, though current systems tend toward overly restrictive or overly broad phenotyping [84].

1.10. Clinical Trial Protocol Development

LLMs are reshaping clinical trial protocol development across multiple dimensions. In a protocol-writing study, Markey et al. evaluated GPT-4's ability to generate protocol sections, including endpoints and eligibility criteria. Off-the-shelf GPT-4 performed well on content relevance and medical terminology (scores > 80%) but showed limited performance in clinical thinking/logic and transparency/references (scores 40% or less). However, retrieval-augmented generation (RAG) incorporating regulatory guidance and ClinicalTrials.gov data markedly improved these weaker dimensions to approximately 80%, demonstrating that hybrid architectures greatly enhance practical usability for protocol writing [85].

For patient-facing materials, ensuring accessibility remains critical. Ali et al. demonstrated that GPT-4 could reduce the reading level of consent forms from the college freshman level to the eighth-grade level while maintaining completeness and legal validity. Their AI-human collaborative framework—validated by physicians and medical malpractice attorneys—achieved sixth-grade readability for procedure-specific forms while achieving perfect scores on consent quality metrics [86]. Though focused on surgical consents, this approach extends to clinical trial informed consent documents facing similar literacy barriers.

SAP generation and adaptive trial design represent emerging applications, though rigorous validation studies remain limited. The complexity of regulatory requirements and the need for methodological precision demand careful human oversight in these domains.

Patient-trial matching presents perhaps the most mature LLM application. Jin et al. developed TrialGPT, an end-to-end framework that retrieves candidate trials, predicts criterion-level eligibility with explanations, and ranks trials for patients. TrialGPT achieved 87.3% accuracy on eligibility predictions and reduced clinician screening time by 42.6% in user studies, while recalling over 90% of relevant trials from just 6% of initial collections [84]. This automated matching addresses a critical bottleneck in trial recruitment.

2. Methodological Considerations

2.1. Prompt Engineering and Optimization

The effectiveness of LLMs in medical research depends critically on prompt engineering, yet optimal strategies remain largely empirical. A scoping review of 114 studies identified prompt design as the most prevalent approach, though terminology remains in-

consistent across the field [87]. Structured prompts with explicit role definition, clear output specifications, and step-by-step instructions generally improve performance, but “prompt brittleness” means even minor wording changes can markedly alter outputs [74,88].

CoT prompting shows particular promise for clinical reasoning, with o1-mini achieving 88.4% accuracy on clinical question-answering tasks [89]. However, CoT benefits primarily large models (≈ 100 B parameters); smaller models often produce illogical reasoning that degrades performance [90]. Self-consistency sampling improved MedQA accuracy by over 7% but decreased performance on other datasets, highlighting the need for task-specific validation [91].

Temperature settings have a smaller impact than commonly assumed. GPT-4o maintained consistent accuracy (98.7–99.0%) across temperatures 0.0–1.5, with degradation only above 1.75 [35]. For reproducibility, temperature = 0 is recommended, though true determinism remains elusive due to hardware-level variations.

2.2. Validation and Quality Assurance

Robust validation is essential for LLM-assisted research. The MI-CLEAR-LLM checklist identifies six critical reporting items affecting reproducibility: LLM identification, stochasticity handling, prompt documentation, prompt structuring, optimization details, and test data independence [92]. However, an analysis of studies published in top medical journals found that only 15.1% adequately reported stochasticity handling [75]. Common error types include reference hallucination, numerical transposition, and context misunderstanding [93,94]. Multi-layered validation, combining automated checks and expert review, can catch most errors before they propagate. Reproducibility remains challenging: model versions change, outputs vary stochastically, and prompt sensitivity means minor variations produce different results [95].

2.3. Ethical and Regulatory Considerations

Publication Ethics and Attribution

Major medical journals and organizations have established policies requiring disclosure of AI use, though requirements vary considerably [15,96]. The consensus that LLMs cannot be listed as authors reflects fundamental principles about responsibility and accountability—AI tools cannot fulfill ICMJE authorship criteria requiring intellectual contribution and responsibility for accuracy [96]. As of 2024, over 80% of publisher policies require disclosure statements when AI is used in manuscript preparation [97].

Questions remain about appropriate attribution when LLMs contribute substantially to analysis or writing. The challenge of maintaining transparency while protecting intellectual property creates tension: some journals require submission of complete prompts and interaction logs, potentially revealing proprietary methods or sensitive information [98]. Balancing transparency with practical considerations remains an ongoing challenge for journals and researchers.

2.4. Data Privacy and Security

The use of LLMs with clinical data raises critical privacy concerns requiring careful management [99]. Regulatory frameworks such as HIPAA/GDPR impose strict requirements that may be incompatible with commercial LLM services, while local deployment requires technical expertise not widely available [100,101]. International variations in privacy regulations add complexity, and developing compliant workflows that satisfy multiple regulatory frameworks remains challenging [102,103].

2.5. Access Limitations and Information Bias

LLMs are predominantly trained on web-crawled data, with academic content comprising only 2–5% of training tokens. The Pile dataset, used by multiple open-source models, explicitly includes only open-access sources such as PubMed Central and arXiv, with no paywalled journal content [104]. Common Crawl, which provides over 80% of GPT-3's training tokens, systematically underrepresents academic literature through authentication barriers that prevent access to subscription content [105]. This creates a “paywall blind spot” where LLMs may have limited exposure to methodological details in premium publications from Elsevier, Springer Nature, and Wiley. Researchers must supplement LLM-assisted synthesis with subscription database queries to ensure comprehensive literature coverage.

2.6. Bias and Fairness

Systematic biases in LLM outputs pose risks for medical research that could perpetuate health disparities. A SR found that over 90% of studies identified demographic biases in medical LLMs [106]. These biases manifest in clinically consequential ways: ChatGPT, GPT-4, and Claude propagate debunked race-based medicine, including false claims about racial differences in kidney function and pain thresholds [107]. GPT-4 was more likely to rate Black patients as abusing opioids when presented with identical clinical information [108]. Mitigation strategies—including bias education prompts and diverse prompt testing—show promise but remain insufficiently validated [109]. Researchers should implement systematic demographic auditing before deploying LLMs in clinical applications.

2.7. Integrating Scientific Integrity into LLM Workflows

Minimizing risks associated with bias and model inaccuracy requires embedding principles of scientific integrity at every stage of LLM-assisted research workflows. We propose a five-component framework. First, transparency and reproducibility should be ensured through mandatory documentation of model version, temperature settings, complete prompts, and API parameters, as recommended by the MI-CLEAR-LLM checklist [92]; notably, only 15.1% of studies in top medical journals adequately reported stochasticity handling [75]. Second, multi-layered verification should combine automated checking (reference verification against bibliographic databases), cross-model consensus (dual-LLM approaches reducing hallucination from approximately 2.5% to 0.25% [42]), and expert review as the final arbiter. Third, systematic bias auditing should be implemented as a standard protocol before any LLM deployment in research, given that over 90% of studies have identified demographic biases in medical LLMs [106]. Fourth, institutions should establish human-in-the-loop governance that specifies which research tasks are appropriate for LLM assistance versus those that require fully human execution, addressing the dual risks of automation bias and automation neglect [110–113]. Fifth, research training programs should ensure foundational skills in manual literature screening, data extraction, and critical appraisal before introducing AI assistance, thereby preventing the “never-skilling” phenomenon in which trainees fail to develop independent analytical capabilities [114].

2.8. Limitations and the Human-AI Partnership

Despite impressive capabilities in pattern recognition and text generation, current LLMs exhibit limitations in logical reasoning and causal inference, which are critical for medical research. While excelling at identifying correlations across vast datasets, they struggle with counterfactual reasoning and may fail to recognize confounding factors—for instance, correctly identifying drug-outcome associations in observational data while missing why an RCT might yield different results [115]. This reasoning gap necessitates

a “human in the loop” approach where researchers provide causal understanding while LLMs handle information synthesis.

Beyond limitations in reasoning, human-AI interaction challenges raise additional concerns. While automation bias—over-reliance on AI outputs—has received attention, the converse phenomenon of automation neglect poses equal risks in research contexts. Automation neglect occurs when experienced researchers dismiss AI recommendations due to overconfidence in their own judgment or distrust of AI systems [110–112]. In clinical AI studies, experts were significantly more likely than non-experts to ignore accurate AI recommendations, with up to 16% of correct outputs being dismissed [113]. In medical research, this may manifest as senior investigators rejecting valid LLM-identified studies during screening or dismissing accurate data extractions, potentially introducing systematic errors that paradoxically favor human fallibility over AI accuracy.

Excessive reliance on LLM-generated syntheses poses a subtle risk through cognitive offloading—delegating mental processes to external systems. Stadler et al. demonstrated this paradox experimentally: students using LLMs for scientific inquiry reported significantly lower cognitive load but produced arguments with significantly lower validity compared to those using traditional methods [116]. When researchers rely on automated “deep research” functions that provide pre-digested results, they may bypass the critical cognitive processes essential for developing expertise and recognizing novel patterns. Recent neuroscience research provides biological evidence for these concerns. A study tracking brain activity during AI-assisted cognitive tasks demonstrated that users showed up to 55% reduced neural connectivity in frequencies associated with deep thinking, with an impaired ability to recall content they had just produced—a phenomenon termed “cognitive debt”, where immediate convenience creates long-term cognitive costs [117]. Medical researchers should therefore maintain proficiency in traditional methods and actively engage with primary sources, recognizing that apparent efficiency gains may represent a trade-off against long-term analytical capability [118].

These concerns extend particularly to research training. The phenomenon of “never-skilling”—where trainees who learn SR methods exclusively with AI assistance fail to develop independent analytical capabilities—poses risks for the next generation of researchers [114]. Additionally, “mis-skilling,” where AI errors or biases are learned and perpetuated by trainees as correct methodology, may systematically compromise research quality. Just as clinical educators now advocate for periodic AI-free practice to preserve diagnostic competence, research training programs should ensure foundational skills in manual literature screening, data extraction, and critical appraisal before introducing AI assistance.

3. Research Question and Hypothesis Generation

LLMs show emerging potential for generating novel research questions and scientific hypotheses. By synthesizing patterns across vast bodies of literature, LLMs can identify knowledge gaps and propose testable hypotheses that might not be apparent to individual researchers. A recent study experimentally validated this capability: GPT-4 was tasked with hypothesizing novel synergistic drug combinations for breast cancer treatment, and laboratory experiments confirmed that 3 of 12 AI-generated hypotheses (25%) demonstrated synergy scores exceeding those of positive controls [119]. In a subsequent iterative round, 3 of 4 additional AI-suggested combinations also showed positive synergy. While these results suggest LLMs can serve as valuable sources of scientific hypotheses, concerns remain about their tendency to reinforce existing paradigms rather than proposing truly innovative directions, and validation comparing AI-generated research questions with expert-derived hypotheses remains limited.

4. Future Directions

Several technological advances promise to address current limitations. Multimodal models processing text, images, and structured data simultaneously will enable more comprehensive analysis of complex medical information [120]. Retrieval-augmented generation, combining LLM reasoning with real-time database access, could address concerns about hallucination and outdated information [121]. Specialized medical models trained on biomedical literature show promise for improved domain-specific performance, though validation frameworks and bias assessment remain essential [122].

The recent success of AI-discovered therapeutics, including the first AI-identified drug showing efficacy in Phase IIa trials, demonstrates that LLMs are transitioning from assistive tools to active partners in hypothesis generation [123]. Future applications may include autonomous experimental design, real-time adaptive trial modifications, and continuous evidence synthesis that automatically incorporates new findings. However, realizing this potential requires the development of explainable AI for medical research, the integration of causal reasoning capabilities, and ethical frameworks for attribution when AI contributes substantively to discovery [39,76,77].

Emerging Open-Source and Cost-Effective Models

While this review has focused predominantly on GPT-series models—reflecting the composition of the published evidence base through May 2025—the rapid emergence of open-source alternatives with competitive performance at substantially lower cost represents a significant development for the democratization of AI-assisted medical research. DeepSeek-R1 (671 B parameters, mixture-of-experts architecture, MIT license) achieved 92% accuracy on USMLE questions, approaching GPT-4o's 95% [124]. In 125 standardized patient cases, DeepSeek-R1 performed on par with GPT-4o in clinical decision-making tasks ($p = 0.31$) [125]. At approximately \$0.28 per million input tokens—roughly 9–24 times cheaper than GPT-4o—and with open-weight deployment eliminating cloud data transfer requirements, DeepSeek addresses both cost and privacy barriers simultaneously.

Similarly, Qwen (Alibaba Cloud, Apache 2.0 license) demonstrated strong performance on Chinese-language medical tasks, achieving 88.9% accuracy on the Chinese National Nursing Licensing Examination compared to GPT-4o's 80.7% [126]. However, on English-language medical benchmarks, Qwen generally trails GPT-4o (e.g., 0.57 vs. 0.73 accuracy in cancer genetic variant classification; [127]). Both models support local deployment, enabling institutions to process sensitive clinical data without cloud transmission—a critical advantage for HIPAA/GDPR compliance. However, important limitations remain: DeepSeek-R1 lacks native multimodal capability, generates verbose responses with increased latency, and its reasoning module does not consistently improve clinical performance over its base model. These emerging models underscore the need for review frameworks that transcend any single model's capabilities and instead evaluate the general principles of human-AI collaboration in medical research.

5. Conclusions

This review synthesizes current evidence on LLM applications across SRs, scientific writing, and clinical research. LLMs demonstrate variable but promising performance: literature screening shows high sensitivity with substantial workload reduction, while tasks requiring subjective judgment, such as risk-of-bias assessment, remain insufficiently validated for standalone use. Hallucination and demographic bias represent critical concerns demanding rigorous verification protocols and systematic auditing before clinical deployment. The cognitive offloading paradox presents an underappreciated risk: while LLMs

reduce cognitive burden and increase efficiency, excessive reliance may systematically weaken researchers’ analytical capabilities.

This review has several limitations. As a narrative review rather than a systematic review, our literature search and study selection, though structured, were not exhaustive. The rapid pace of LLM development means that some findings reviewed here may already be outdated. Publication bias toward positive results may overestimate LLM capabilities, and the heterogeneity of evaluation metrics across studies limits direct comparisons. Furthermore, most evidence derives from studies using proprietary commercial models (e.g., GPT-4), whose underlying architectures and training data are not fully transparent, limiting reproducibility and generalizability of findings.

We recommend a structured approach (Figure 1): start with low-risk applications, implement multi-layered validation, maintain reproducible settings, and preserve human judgment for tasks requiring causal reasoning. LLMs are powerful but inherently unstable instruments requiring constant calibration—success depends on researchers maintaining their roles as critical overseers rather than passive consumers of AI-generated content. In practice, this means adopting iterative, step-by-step refinement rather than expecting polished output from single prompts, and rigorously verifying every AI-generated citation and claim against primary sources.

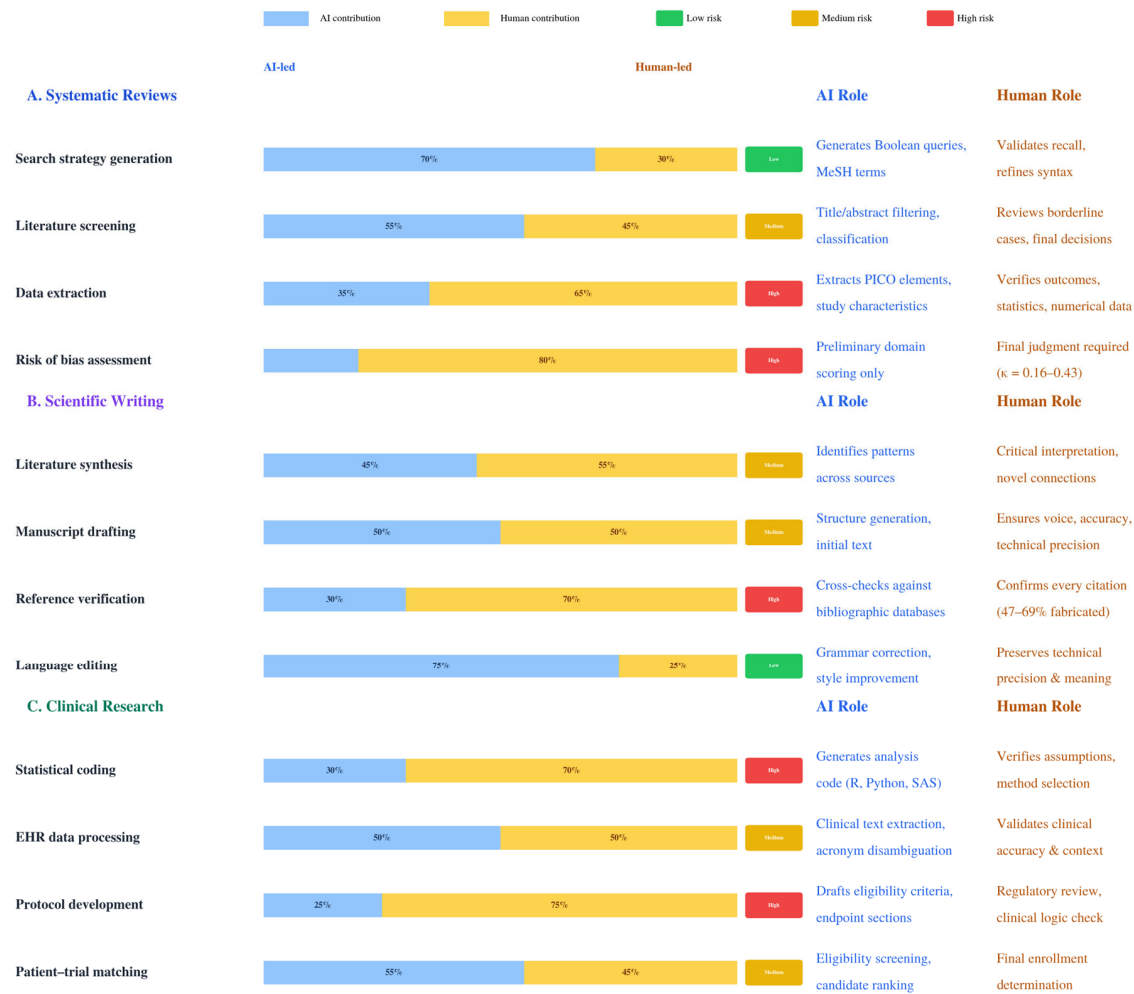


Figure 1. Recommended human–AI collaborative workflow framework for LLM applications across three medical research domains (systematic reviews, scientific writing, and clinical research). Each

task is classified by risk level: low (LLM-led, acceptable with spot checks), medium (LLM-assisted, with systematic human review), or high (human judgment essential, with the LLM as a preliminary tool only). Core principles include universal verification, full documentation per MI-CLEAR-LLM guidelines [92], systematic bias auditing [106], and preservation of traditional methodological skills to prevent cognitive offloading [116,117].

Author Contributions: Conceptualization, C.S.B. and E.J.G.; Data curation, C.S.B., E.J.G., and Y.S.S.; Formal analysis, C.S.B.; Investigation, C.S.B. and E.J.G.; Methodology, C.S.B.; Project administration, E.J.G.; Resources, C.S.B.; Writing—original draft, E.J.G. and C.S.B.; Writing—review and editing, E.J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available upon request to the corresponding author. All investigators have access to the final dataset.

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

References

1. van Dis, E.A.M.; Bollen, J.; Zuidema, W.; van Rooij, R.; Bockting, C.L. ChatGPT: Five priorities for research. *Nature* **2023**, *614*, 224–226. [[CrossRef](#)] [[PubMed](#)]
2. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [[CrossRef](#)] [[PubMed](#)]
3. Gong, E.J.; Bang, C.S. Evaluating the role of large language models in inflammatory bowel disease patient information. *World J. Gastroenterol.* **2024**, *30*, 3538–3540. [[CrossRef](#)]
4. Gong, E.J.; Bang, C.S. Revolutionizing gastrointestinal endoscopy: The emerging role of large language models. *Clin. Endosc.* **2024**, *57*, 759–762. [[CrossRef](#)]
5. Gong, E.J.; Bang, C.S.; Lee, J.J.; Park, J.; Kim, E.; Kim, S.; Kimm, M.; Choi, S.-H. Large Language Models in Gastroenterology: Systematic Review. *J. Med. Internet Res.* **2024**, *26*, e66648. [[CrossRef](#)] [[PubMed](#)]
6. Liu, J.; Wang, C.; Liu, S. Utility of ChatGPT in Clinical Practice. *J. Med. Internet Res.* **2023**, *25*, e48568. [[CrossRef](#)]
7. Kim, H.J.; Gong, E.J.; Bang, C.-S. Application of Machine Learning Based on Structured Medical Data in Gastroenterology. *Biomimetics* **2023**, *8*, 512. [[CrossRef](#)]
8. Borah, R.; Brown, A.W.; Capers, P.L.; Kaiser, K.A. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* **2017**, *7*, e012545. [[CrossRef](#)]
9. Landhuis, E. Scientific literature: Information overload. *Nature* **2016**, *535*, 457–458. [[CrossRef](#)]
10. Marshall, I.J.; Wallace, B.C. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Syst. Rev.* **2019**, *8*, 163. [[CrossRef](#)] [[PubMed](#)]
11. Linardon, J.; Messer, M.; Anderson, C.; Liu, C.; McClure, Z.; Jarman, H.K.; Goldberg, S.B.; Torous, J. Role of large language models in mental health research: An international survey of researchers’ practices and perspectives. *BMJ Ment. Health* **2025**, *28*, e301787. [[CrossRef](#)]
12. Qureshi, R.; Shaughnessy, D.; Gill, K.A.R.; Robinson, K.A.; Li, T.; Agai, E. Are ChatGPT and large language models “the answer” to bringing us closer to systematic review automation? *Syst. Rev.* **2023**, *12*, 72. [[CrossRef](#)] [[PubMed](#)]
13. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [[CrossRef](#)]
14. Anonymous. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* **2023**, *613*, 612. [[CrossRef](#)]
15. Flanagin, A.; Bibbins-Domingo, K.; Berkwits, M.; Christiansen, S.L. Nonhuman “Authors” and Implications for the Integrity of Scientific Publication and Medical Knowledge. *JAMA* **2023**, *329*, 637–639. [[CrossRef](#)] [[PubMed](#)]
16. Bedi, S.; Liu, Y.; Orr-Ewing, L.; Dash, D.; Koyejo, S.; Callahan, A.; Fries, J.A.; Wornow, M.; Swaminathan, A.; Lehmann, L.S.; et al. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA* **2025**, *333*, 319–328. [[CrossRef](#)]
17. Scherbakov, D.; Hubig, N.; Jansari, V.; Bakumenko, A.; Lenert, L.A. The emergence of large language models as tools in literature reviews: A large language model-assisted systematic review. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 1071–1086. [[CrossRef](#)]

18. Lieberum, J.L.; Toews, M.; Metzendorf, M.I.; Heilmeyer, F.; Siemens, W.; Haverkamp, C.; Böhringer, D.; Meerpohl, J.J.; Eisele-Metzger, A. Large language models for conducting systematic reviews: On the rise, but not yet ready for use—A scoping review. *J. Clin. Epidemiol.* **2025**, *181*, 111746. [CrossRef] [PubMed]
19. Ahn, S. The transformative impact of large language models on medical writing and publishing: Current applications, challenges and future directions. *Korean J. Physiol. Pharmacol.* **2024**, *28*, 393–401. [CrossRef]
20. Omar, M.; Nadkarni, G.N.; Klang, E.; Glicksberg, B.S. Large language models in medicine: A review of current clinical trials across healthcare applications. *PLoS Digit. Health* **2024**, *3*, e0000662. [CrossRef]
21. Ferrari, R. Writing narrative style literature reviews. *Med. Writ.* **2015**, *24*, 230–235. [CrossRef]
22. Greenhalgh, T.; Thorne, S.; Malterud, K. Time to challenge the spurious hierarchy of systematic over narrative reviews? *Eur. J. Clin. Investig.* **2018**, *48*, e12931. [CrossRef]
23. Sukhera, J. Narrative reviews: Flexible, rigorous, and practical. *J. Grad. Med. Educ.* **2022**, *14*, 414–417. [CrossRef] [PubMed]
24. Baethge, C.; Goldbeck-Wood, S.; Mertens, S. SANRA—A scale for the quality assessment of narrative review articles. *Res. Integr. Peer Rev.* **2019**, *4*, 5. [CrossRef]
25. Wang, S.; Scells, H.; Koopman, B.; Zuccon, G. Can ChatGPT write a good boolean query for systematic review literature search? *arXiv* **2023**, arXiv:2302.03495. [CrossRef]
26. Yu, F.; Kincaide, H.; Carlson, R.B. An Empirical Study Evaluating ChatGPT’s Performance in Generating Search Strategies for Systematic Reviews. *Proc. Assoc. Inf. Sci. Technol.* **2024**, *61*, 423–434. [CrossRef]
27. Parisi, V.; Sutton, A. The role of ChatGPT in developing systematic literature searches: An evidence summary. *J. EAHIL* **2024**, *20*, 30–34. [CrossRef]
28. O’Connor, A.M.; Tsafnat, G.; Gilbert, S.B.; Thayer, K.A.; Shemilt, I.; Thomas, J.; Glasziou, P.; Wolfe, M.S. Still moving toward automation of the systematic review process: A summary of discussions at the third meeting of the International Collaboration for Automation of Systematic Reviews (ICASR). *Syst. Rev.* **2019**, *8*, 57. [CrossRef]
29. Clark, J.; Glasziou, P.; Del Mar, C.; Bannach-Brown, A.; Stehlik, P.; Scott, A.M. A full systematic review was completed in 2 weeks using automation tools: A case study. *J. Clin. Epidemiol.* **2020**, *121*, 81–90. [CrossRef]
30. Issaiy, M.; Ghanaati, H.; Kolahi, S.; Shakiba, M.; Jalali, A.H.; Zarei, D.; Kazemian, S.; Avanaki, M.A.; Firouznia, K. Methodological insights into ChatGPT’s screening performance in systematic reviews. *BMC Med. Res. Methodol.* **2024**, *24*, 78. [CrossRef] [PubMed]
31. Cai, X.; Geng, Y.; Du, Y.; Westerman, B.; Wang, D.; Ma, C.; Vallejo, J.J.G. Utilizing Large language models to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. *BMC Med. Res. Methodol.* **2025**, *25*, 116. [CrossRef]
32. Khraisha, Q.; Put, S.; Kappenberg, J.; Warritch, A.; Hadfield, K. Can large language models replace humans in systematic reviews? Evaluating GPT-4’s efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res. Synth. Methods* **2024**, *15*, 616–626. [CrossRef]
33. Kohandel Gargari, O.; Mahmoudi, M.H.; Hajisafarali, M.; Samiee, R. Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. *BMJ Evid.-Based Med.* **2024**, *29*, 69–70. [CrossRef] [PubMed]
34. Matsui, K.; Utsumi, T.; Aoki, Y.; Maruki, T.; Takeshima, M.; Takaesu, Y. Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews. *J. Med. Internet Res.* **2024**, *26*, e52758. [CrossRef]
35. Windisch, P.; Dennstädt, F.; Koechli, C.; Schröder, C.; Aebersold, D.M.; Förster, R.; Zwahlen, D.R.; Windisch, P.Y. The Impact of Temperature on Extracting Information From Clinical Trial Publications Using Large Language Models. *Cureus* **2024**, *16*, e75748. [CrossRef]
36. Oami, T.; Okada, Y.; Nakada, T.-A. Optimal large language models to screen citations for systematic reviews. *Res. Synth. Methods* **2025**, *16*, 859–875. [CrossRef]
37. Sorich, M.J.; Mangoni, A.A.; Bacchi, S.; Menz, B.D.; Hopkins, A.M. The Triage and Diagnostic Accuracy of Frontier Large Language Models: Updated Comparison to Physician Performance. *J. Med. Internet Res.* **2024**, *26*, e67409. [CrossRef] [PubMed]
38. Karpathy, A. *LLM Council*; GitHub: San Francisco, CA, USA, 2025. Available online: <https://github.com/karpathy/llm-council> (accessed on 13 March 2026).
39. Guo, E.; Gupta, M.; Deng, J.; Park, Y.-J.; Paget, M.; Naugler, C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *J. Med. Internet Res.* **2024**, *26*, e48996. [CrossRef]
40. Jonnalagadda, S.R.; Goyal, P.; Huffman, M.D. Automating data extraction in systematic reviews: A systematic review. *Syst. Rev.* **2015**, *4*, 78. [CrossRef] [PubMed]
41. Schmidt, L.; Hair, K.; Graziosi, S.; Campbell, F.; Kapp, C.; Khanteymoori, A.; Craig, D.; Engelbert, M.; Thomas, J. Exploring the use of a large language model for data extraction in systematic reviews: A rapid feasibility study. *arXiv* **2024**, arXiv:2405.14445. [CrossRef]

42. Khan, M.A.; Ayub, U.; Naqvi, S.A.A.; Khakwani, K.Z.R.; Sipra, Z.B.R.; Raina, A.; Zhou, S.; He, H.; Saeidi, A.; Hasan, B.; et al. Collaborative large language models for automated data extraction in living systematic reviews. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 638–647. [\[CrossRef\]](#)
43. Konet, A.; Thomas, I.; Gartlehner, G.; Kahwati, L.; Hilscher, R.; Kugley, S.; Crotty, K.; Viswanathan, M.; Chew, R. Performance of two large language models for data extraction in evidence synthesis. *Res. Synth. Methods* **2024**, *15*, 818–824. [\[CrossRef\]](#)
44. Gartlehner, G.; Kahwati, L.; Hilscher, R.; Thomas, I.; Kugley, S.; Crotty, K.; Viswanathan, M.; Nussbaumer-Streit, B.; Booth, G.; Erskine, N.; et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Res. Synth. Methods* **2024**, *15*, 576–589. [\[CrossRef\]](#)
45. Kim, G.; Hong, T.; Yim, M.; Nam, J.; Park, J.; Yim, J.; Hwang, W.; Yun, S.; Han, D.; Park, S. Ocr-free document understanding transformer. *arXiv* **2021**, arXiv:2111.15664. [\[CrossRef\]](#)
46. Wang, D.; Raman, N.; Sibue, M.; Ma, Z.; Babkin, P.; Kaur, S.; Pei, Y.; Nourbakhsh, A.; Liu, X. Doclm: A layout-aware generative language model for multimodal document understanding. *arXiv* **2023**, arXiv:2401.00908.
47. Jin, Q.; Chen, F.; Zhou, Y.; Xu, Z.; Cheung, J.M.; Chen, R.; Summers, R.M.; Rousseau, J.F.; Ni, P.; Landsman, M.J.; et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *npj Digit. Med.* **2024**, *7*, 190. [\[CrossRef\]](#)
48. Bhattacharyya, M.; Miller, V.M.; Bhattacharyya, D.; Miller, L.E. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus* **2023**, *15*, e39238. [\[CrossRef\]](#)
49. Motzfeldt Jensen, M.; Brix Danielsen, M.; Riis, J.; Assifuah Kristjansen, K.; Andersen, S.; Okubo, Y.; Jørgensen, M.G. ChatGPT-4o can serve as the second rater for data extraction in systematic reviews. *PLoS ONE* **2025**, *20*, e0313401. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Pitre, T.; Jassal, T.; Talukdar, J.R.; Shahab, M.; Ling, M.; Zeraatkar, D. ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: A methods study. *medRxiv* **2023**, medRxiv:2023.11.19.23298727.
51. Kuitunen, I.; Ponkilainen, V.T.; Liukkonen, R.; Nyrhi, L.; Pakarinen, O.; Vaajala, M.; Uimonen, M.M. Evaluating the Performance of ChatGPT-4o in Risk of Bias Assessments. *J. Evid.-Based Med.* **2024**, *17*, 700–702. [\[CrossRef\]](#)
52. Kuitunen, I.; Nyrhi, L.; De Luca, D. ChatGPT-4o in Risk-of-Bias Assessments in Neonatology: A Validity Analysis. *Neonatology* **2025**, *122*, 360–365. [\[CrossRef\]](#)
53. Šuster, S.; Baldwin, T.; Verspoor, K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. *Res. Synth. Methods* **2024**, *15*, 988–1000. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Lai, H.; Ge, L.; Sun, M.; Pan, B.; Huang, J.; Hou, L.; Yang, Q.; Liu, J.; Liu, J.; Ye, Z.; et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. *JAMA Netw. Open* **2024**, *7*, e2412687. [\[CrossRef\]](#)
55. Huang, J.; Lai, H.; Zhao, W.; Xia, D.; Bai, C.; Sun, M.; Liu, J.; Liu, J.; Pan, B.; Tian, J.; et al. Large Language Model-Assisted Risk-of-Bias Assessment in Randomized Controlled Trials Using the Revised Risk-of-Bias Tool: Evaluation Study. *J. Med. Internet Res.* **2025**, *27*, e70450. [\[CrossRef\]](#)
56. Huang, J.; Tan, M. The role of ChatGPT in scientific communication: Writing better scientific review articles. *Am. J. Cancer Res.* **2023**, *13*, 1148–1154. [\[PubMed\]](#)
57. Amano, T.; González-Varo, J.P.; Sutherland, W.J. Languages Are Still a Major Barrier to Global Science. *PLoS Biol.* **2016**, *14*, e2000933. [\[CrossRef\]](#)
58. Dergaa, I.; Chamari, K.; Zmijewski, P.; Ben Saad, H. From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biol. Sport* **2023**, *40*, 615–622. [\[CrossRef\]](#)
59. Gong, E.J.; Woo, J.; Lee, J.J.; Bang, C.S. Role of artificial intelligence in gastric diseases. *World J. Gastroenterol.* **2025**, *31*, 111327. [\[CrossRef\]](#)
60. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; et al. “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* **2023**, *71*, 102642. [\[CrossRef\]](#)
61. Walters, W.H.; Wilder, E.I. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci. Rep.* **2023**, *13*, 14045. [\[CrossRef\]](#)
62. Else, H. Abstracts written by ChatGPT fool scientists. *Nature* **2023**, *613*, 423. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Salvagno, M.; Taccone, F.S.; Gerli, A.G. Can artificial intelligence help for scientific writing? *Crit. Care* **2023**, *27*, 75. [\[CrossRef\]](#)
64. Patel, S.B.; Lam, K. ChatGPT: The future of discharge summaries? *Lancet Digit. Health* **2023**, *5*, e107–e108. [\[CrossRef\]](#)
65. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J. Med. Syst.* **2023**, *47*, 33. [\[CrossRef\]](#) [\[PubMed\]](#)
66. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegrefe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-refine: Iterative refinement with self-feedback. *arXiv* **2023**, arXiv:2303.17651.
67. Sun, S.; Yuan, R.; Cao, Z.; Li, W.; Liu, P. Prompt chaining or stepwise prompt? Refinement in text summarization. *arXiv* **2024**, arXiv:2406.00507. [\[CrossRef\]](#)
68. Lee, P.; Bubeck, S.; Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med.* **2023**, *388*, 1233–1239. [\[CrossRef\]](#)

69. Ruta, M.R.; Gaidici, T.; Irwin, C.; Lifshitz, J. ChatGPT for Univariate Statistics: Validation of AI-Assisted Data Analysis in Healthcare Research. *J. Med. Internet Res.* **2025**, *27*, e63550. [\[CrossRef\]](#) [\[PubMed\]](#)
70. Huang, Y.; Wu, R.; He, J.; Xiang, Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: A comparative analysis with SAS, SPSS, and R. *J. Glob. Health* **2024**, *14*, 04070. [\[CrossRef\]](#) [\[PubMed\]](#)
71. Dobler, D.; Binder, H.; Boulesteix, A.L.; Igelmann, J.; Köhler, D.; Mansmann, U.; Pauly, M.; Scherag, A.; Schmid, M.; Al Tawil, A.; et al. ChatGPT as a Tool for Biostatisticians: A Tutorial on Applications, Opportunities, and Limitations. *Stat. Med.* **2025**, *44*, e70263. [\[CrossRef\]](#)
72. Shahrul, A.I.; Syed Mohamed, A.M.F. A Comparative Evaluation of Statistical Product and Service Solutions (SPSS) and ChatGPT-4 in Statistical Analyses. *Cureus* **2024**, *16*, e72581. [\[CrossRef\]](#)
73. Evans, R.; Pozzi, A. Using CHATGPT to develop the statistical analysis plan for a randomized controlled trial: A case report. *Research Square* **2023**. [\[CrossRef\]](#)
74. Lee, J.H.; Shin, J. How to Optimize Prompting for Large Language Models in Clinical Research. *Korean J. Radiol.* **2024**, *25*, 869–873. [\[CrossRef\]](#)
75. Suh, C.H.; Yi, J.; Shim, W.H.; Heo, H. Insufficient Transparency in Stochasticity Reporting in Large Language Model Studies for Medical Applications in Leading Medical Journals. *Korean J. Radiol.* **2024**, *25*, 1029–1031. [\[CrossRef\]](#)
76. Ordak, M. ChatGPT's Skills in Statistical Analysis Using the Example of Allergology: Do We Have Reason for Concern? *Healthcare* **2023**, *11*, 2554. [\[CrossRef\]](#)
77. Liu, X.; Wu, Z.; Wu, X.; Lu, P.; Chang, K.-W.; Feng, Y. Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. *arXiv* **2024**, arXiv:2402.17644. [\[CrossRef\]](#)
78. Van Veen, D.; Van Uden, C.; Blankemeier, L.; Delbrouck, J.B.; Aali, A.; Bluethgen, C.; Pareek, A.; Polacin, M.; Reis, E.P.; Seehofnerová, A.; et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **2024**, *30*, 1134–1142. [\[CrossRef\]](#) [\[PubMed\]](#)
79. Yan, C.; Ong, H.H.; Grabowska, M.E.; Krantz, M.S.; Su, W.C.; Dickson, A.L.; Peterson, J.F.; Feng, Q.; Roden, D.M.; Stein, C.M.; et al. Large language models facilitate the generation of electronic health record phenotyping algorithms. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1994–2001. [\[CrossRef\]](#)
80. Jonnagaddala, J.; Wong, Z.S. Privacy preserving strategies for electronic health records in the era of large language models. *npj Digit. Med.* **2025**, *8*, 34. [\[CrossRef\]](#)
81. Wiest, I.C.; Ferber, D.; Zhu, J.; van Treeck, M.; Meyer, S.K.; Juglan, R.; Carrero, Z.I.; Paech, D.; Kleesiek, J.; Ebert, M.P.; et al. Privacy-preserving large language models for structured medical information retrieval. *npj Digit. Med.* **2024**, *7*, 257. [\[CrossRef\]](#)
82. Kugic, A.; Schulz, S.; Kreuzthaler, M. Disambiguation of acronyms in clinical narratives with large language models. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 2040–2046. [\[CrossRef\]](#)
83. Cui, H.; Unell, A.; Chen, B.; Fries, J.A.; Alsentzer, E.; Koyejo, S.; Shah, N.H. TIMER: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digit. Med.* **2025**, *8*, 577. [\[CrossRef\]](#)
84. Jin, Q.; Wang, Z.; Floudas, C.S.; Chen, F.; Gong, C.; Bracken-Clarke, D.; Xue, E.; Yang, Y.; Sun, J.; Lu, Z. Matching patients to clinical trials with large language models. *Nat. Commun.* **2024**, *15*, 9074. [\[CrossRef\]](#)
85. Markey, N.; El-Mansouri, I.; Rensonnet, G.; van Langen, C.; Meier, C. From RAGs to riches: Utilizing large language models to write documents for clinical trials. *Clin. Trials* **2025**, *22*, 626–631. [\[CrossRef\]](#) [\[PubMed\]](#)
86. Ali, R.; Connolly, I.D.; Tang, O.Y.; Mirza, F.N.; Johnston, B.; Abdulrazeq, H.F.; Lim, R.K.; Galamaga, P.F.; Libby, T.J.; Sodha, N.R.; et al. Bridging the literacy gap for surgical consents: An AI-human expert collaborative approach. *npj Digit. Med.* **2024**, *7*, 63. [\[CrossRef\]](#) [\[PubMed\]](#)
87. Zaghir, J.; Naguib, M.; Bjelogrić, M.; Névél, A.; Tannier, X.; Lovis, C. Prompt Engineering Paradigms for Medical Applications: Scoping Review. *J. Med. Internet Res.* **2024**, *26*, e60501. [\[CrossRef\]](#)
88. Gong, E.J.; Bang, C.S. Interpretation of Medical Images Using Artificial Intelligence: Current Status and Future Perspectives. *Korean J. Gastroenterol.* **2023**, *82*, 43–45. [\[CrossRef\]](#)
89. Jeon, S.; Kim, H.G. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. *Comput. Biol. Med.* **2025**, *196*, 110614. [\[CrossRef\]](#)
90. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
91. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [\[CrossRef\]](#) [\[PubMed\]](#)
92. Park, S.H.; Suh, C.H.; Lee, J.H.; Kahn, C.E.; Moy, L. Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM). *Korean J. Radiol.* **2024**, *25*, 865–868. [\[CrossRef\]](#) [\[PubMed\]](#)
93. Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv* **2023**, arXiv:2302.04023. [\[CrossRef\]](#)

94. Alkaissi, H.; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus* **2023**, *15*, e35179. [[CrossRef](#)] [[PubMed](#)]
95. Busch, F.; Hoffmann, L.; Rueger, C.; van Dijk, E.H.; Kader, R.; Ortiz-Prado, E.; Makowski, M.R.; Saba, L.; Hadamitzky, M.; Kather, J.N.; et al. Current applications and challenges in large language models for patient care: A systematic review. *Commun. Med.* **2025**, *5*, 26. [[CrossRef](#)]
96. Zielinski, C.; Winker, M.A.; Aggarwal, R.; Ferris, L.E.; Heinemann, M.; Lapeña, J.F., Jr.; Pai, S.A.; Ing, E.; Citrome, L.; Alam, M.; et al. Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. *Colomb. Med.* **2023**, *54*, e1015868. [[CrossRef](#)] [[PubMed](#)]
97. Ganjavi, C.; Eppler, M.B.; Pekcan, A.; Biedermann, B.; Abreu, A.; Collins, G.S.; Gill, I.S.; Cacciamani, G.E. Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: Bibliometric analysis. *BMJ* **2024**, *384*, e077192. [[CrossRef](#)]
98. Hill-Yardin, E.L.; Hutchinson, M.R.; Laycock, R.; Spencer, S.J. A Chat(GPT) about the future of scientific publishing. *Brain Behav. Immun.* **2023**, *110*, 152–154. [[CrossRef](#)] [[PubMed](#)]
99. Murdoch, B. Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Med. Ethics* **2021**, *22*, 122. [[CrossRef](#)]
100. Price, W.N., 2nd; Cohen, I.G. Privacy in the age of medical big data. *Nat. Med.* **2019**, *25*, 37–43. [[CrossRef](#)]
101. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311. [[CrossRef](#)]
102. Vayena, E.; Blasimme, A.; Cohen, I.G. Machine learning in medicine: Addressing ethical challenges. *PLoS Med.* **2018**, *15*, e1002689. [[CrossRef](#)]
103. Morley, J.; Machado, C.C.V.; Burr, C.; Cows, J.; Joshi, I.; Taddeo, M.; Floridi, L. The ethics of AI in health care: A mapping review. *Soc. Sci. Med.* **2020**, *260*, 113172. [[CrossRef](#)]
104. Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; et al. The pile: An 800 GB dataset of diverse text for language modeling. *arXiv* **2020**, arXiv:2101.00027.
105. Baack, S. A critical analysis of the largest source for generative ai training data: Common crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2024. [[CrossRef](#)]
106. Omar, M.; Sorin, V.; Agbareia, R.; Apakama, D.U.; Soroush, A.; Sakhuja, A.; Freeman, R.; Horowitz, C.R.; Richardson, L.D.; Nadkarni, G.N.; et al. Evaluating and addressing demographic disparities in medical large language models: A systematic review. *International journal for equity in health. Int. J. Equity Health* **2025**, *24*, 57. [[CrossRef](#)] [[PubMed](#)]
107. Omiye, J.A.; Lester, J.C.; Spichak, S.; Rotemberg, V.; Daneshjou, R. Large language models propagate race-based medicine. *npj Digit. Med.* **2023**, *6*, 195. [[CrossRef](#)]
108. Zack, T.; Lehman, E.; Suzgun, M.; Rodriguez, J.A.; Celi, L.A.; Gichoya, J.; Jurafrsky, D.; Szolovits, P.; Bates, D.W.; Abdulnour, R.-E.E.; et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *Lancet Digit. Health* **2024**, *6*, e12–e22. [[CrossRef](#)]
109. Haltaufderheide, J.; Ranisch, R. The ethics of ChatGPT in medicine and healthcare: A systematic review on Large Language Models (LLMs). *npj Digit. Med.* **2024**, *7*, 183. [[CrossRef](#)]
110. Parasuraman, R.; Manzey, D.H. Complacency and bias in human use of automation: An attentional integration. *Hum. Factors* **2010**, *52*, 381–410. [[CrossRef](#)] [[PubMed](#)]
111. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. [[CrossRef](#)]
112. Quek, S.X.Z.; Ho, K.Y. Artificial Intelligence in Upper Gastrointestinal Diagnosis. *Korean J. Helicobacter Up. Gastrointest. Res.* **2025**, *25*, 251. [[CrossRef](#)] [[PubMed](#)]
113. Roser, D.; Meinikheim, M.; Muzalyova, A.; Mendel, R.; Palm, C.; Probst, A.; Nagl, S.; Scheppach, M.W.; Römmele, C.; Schnoy, E.; et al. Artificial Intelligence-assisted Endoscopy and Examiner Confidence: A Study on Human-Artificial Intelligence Interaction in Barrett's Esophagus (with Video). *DEN Open* **2026**, *6*, e70150. [[CrossRef](#)]
114. Abdulnour, R.E.; Gin, B.; Boscardin, C.K. Educational Strategies for Clinical Supervision of Artificial Intelligence Use. *N. Engl. J. Med.* **2025**, *393*, 786–797. [[CrossRef](#)]
115. Marcus, G. Deep learning: A critical appraisal. *arXiv* **2018**, arXiv:1801.00631. [[CrossRef](#)]
116. Stadler, M.; Bannert, M.; Sailer, M. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Comput. Hum. Behav.* **2024**, *160*, 108386. [[CrossRef](#)]
117. Kosmyna, N.; Hauptmann, E.; Yuan, Y.T.; Situ, J.; Liao, X.-H.; Beresnitzky, A.V.; Braunstein, I.; Maes, P. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv* **2025**, arXiv:2506.08872. [[CrossRef](#)]
118. Choudhury, A.; Chaudhry, Z. Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals. *J. Med. Internet Res.* **2024**, *26*, e56764. [[CrossRef](#)]

119. Abdel-Rehim, A.; Zenil, H.; Orhobor, O.; Fisher, M.; Collins, R.J.; Bourne, E.; Fearnley, G.W.; Tate, E.; Smith, H.X.; Soldatova, L.N.; et al. Scientific hypothesis generation by large language models: Laboratory validation in breast cancer treatment. *J. R. Soc. Interface* **2025**, *22*, 20240674. [[CrossRef](#)] [[PubMed](#)]
120. Acosta, J.N.; Falcone, G.J.; Rajpurkar, P.; Topol, E.J. Multimodal biomedical AI. *Nat. Med.* **2022**, *28*, 1773–1784. [[CrossRef](#)]
121. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv* **2020**, arXiv:2005.11401.
122. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950. [[CrossRef](#)] [[PubMed](#)]
123. Ren, F.; Aliper, A.; Chen, J.; Zhao, H.; Rao, S.; Kuppe, C.; Ozerov, I.V.; Zhang, M.; Witte, K.; Kruse, C.; et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat. Biotechnol.* **2025**, *43*, 63–75. [[CrossRef](#)]
124. Tordjman, M.; Liu, Z.; Yuce, M.; Fauveau, V.; Mei, Y.; Hadjadj, J.; Bolger, I.; Almansour, H.; Horst, C.; Parihar, A.S.; et al. Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat. Med.* **2025**, *31*, 2550–2555. [[CrossRef](#)] [[PubMed](#)]
125. Sandmann, S.; Hegselmann, S.; Fujarski, M.; Bickmann, L.; Wild, B.; Eils, R.; Varghese, J. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat. Med.* **2025**, *31*, 2546–2549. [[CrossRef](#)] [[PubMed](#)]
126. Zhu, S.; Hu, W.; Yang, Z.; Yan, J.; Zhang, F. Qwen-2.5 outperforms other large language models in the Chinese National Nursing Licensing Examination: Retrospective cross-sectional comparative study. *JMIR Med. Inform.* **2025**, *13*, e63731. [[CrossRef](#)] [[PubMed](#)]
127. Lin, K.H.; Kao, T.H.; Wang, L.C.; Kuo, C.T.; Chen, P.C.; Chu, Y.C.; Yeh, Y.C. Benchmarking large language models GPT-4o, llama 3.1, and qwen 2.5 for cancer genetic variant classification. *npj Precis. Oncol.* **2025**, *9*, 141. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.