Review Article

ISSN 2769-6294

International Journal of Biomedical Research & Practice

Artificial Intelligence in Literature Search and Evidence Synthesis: Applications, Opportunities, Challenges, and Ethical Considerations: A Systematic Review

Paul Hassan Ilegbusi[1,2,3*], Taoheed Abiola Olanrewaju[4], Abimbola Morolayo Olusuyi[5], Everistus Tochukwu Chiakwa[6,7,8], Omowumi Oluwabukola Okeya[1], Salome Chibuzor Abbah[9], Adekola Taofeek Basiru[9], Ehimwenma Gloria Obayangbon[10], Rotimi Anne Ogunniyi[5], Mojisola Clementina Ogundare[11], Victor Adejare Adebiyi[12], Folake Risper Ayadi[13], Bukola Bosede Ogunro-Abiola[14,15], Cecilia Solape Dunapo[13], Victor Uyi Omorogbe[10], Olatoye Lola Ololajulo[15,16] and Jonathan Amauche Ajah[17]

[1]*Community Health Department, Ondo State College of Health Technology, Akure, Nigeria.*

[2]*Centre for Research Innovation, Development, and Entrepreneurship, London, United Kingdom.*

[3]*Department of Community Health Science, College of Community Health Sciences, Wesley University, Ondo, Nigeria.*

[4]*Community Health Department, Lead City University, Ibadan, Nigeria.*

[5]*Ekiti State College of Health Sciences and Technology, Ijero-Ekiti, Nigeria.*

[6]*Department of Dental Therapy, Faculty of Basic Medical Sciences, University of Ilesa, Osun State, Nigeria.*

[7]*Department of Health Sciences and Dentistry, Frontier University, Garowe, Somalia.*

[8]*Department of Dental Therapy, Millennium College of Health Technology, Akure, Nigeria.*

[9]*College of Community Health, Lagos University Teaching Hospital, Idi-Araba, Nigeria.*

[10]*Community Health Department, Edo State College of Health Technology, Sapele Road, Benin City, Nigeria.*

[11]*Ekiti State Primary Health Care Development Agency, Ado-Ekiti, Nigeria.*

[12]*Data Analytics Department, Wetin Dey Code Academy, Akure, Ondo State, Nigeria.*

[13]*Coastline College of Health Technology, Okitipupa, Ondo State, Nigeria.*

[14]*Department of Sociology, Faculty of Social Sciences, Federal University Oye Ekiti, Ekiti State, Nigeria.*

[15]*Ondo State College of Health Technology, Akure, Nigeria.*

[16]*National Open University of Nigeria, Akure Study Centre, Nigeria.*

[17]*School of Public Health, University of Portharcourt, Nigeria.*

[*]Correspondence:
Paul Hassan Ilegbusi, Community Health Department, Ondo State College of Health Technology, Akure, Nigeria. Centre for Research Innovation, Development, and Entrepreneurship, London, United Kingdom. Department of Community Health Science, College of Community Health Sciences, Wesley University, Ondo, Nigeria, Email: princeilegbusi@gmail.com.

Received: 09 May 2026; **Accepted:** 11 Jun 2026; **Published:** 22 Jun 2026

Citation: Paul Hassan Ilegbusi, Taoheed Abiola Olanrewaju, Abimbola Morolayo Olusuyi, et al. Artificial Intelligence in Literature Search and Evidence Synthesis: Applications, Opportunities, Challenges, and Ethical Considerations: A Systematic Review. Int J Biomed Res Prac. 2026; 6(1); 1-16.

ABSTRACT

Artificial intelligence (AI) technologies are increasingly transforming academic research workflows, particularly in literature search and review processes. Understanding the opportunities, applications, and ethical implications of AI integration in systematic evidence synthesis is critical for responsible adoption in scholarly practice. This systematic review aimed to comprehensively examine the current state of AI applications in literature search and review. We conducted comprehensive searches across multiple databases (SciSpace, Google Scholar, ArXiv, PubMed) from inception to April 2026. Studies were eligible if they addressed AI applications in literature search, screening, data extraction, synthesis, or ethical considerations in academic research. The protocol of this review was published in the Biomedical Journal of Scientific & Technical Research. Following PRISMA 2020 guidelines, we screened 413 unique papers, with 90 papers meeting inclusion criteria for detailed synthesis. Data extraction focused on study design, AI tools and technologies, key findings, performance outcomes, and ethical considerations. AI tools demonstrated substantial benefits in automating literature review tasks, including time savings (median 47% reduction in screening burden), enhanced efficiency in search and screening, and improved synthesis capabilities. Key applications included automated screening (DistillerSR, Abstrackr, ASReview), natural language processing for data extraction (spaCy, transformer models), and large language models for synthesis (ChatGPT, GPT-3.5, GPT-4). However, significant ethical concerns emerged, including risks of algorithmic bias, over-reliance leading to diminished critical thinking, transparency challenges, academic integrity threats (plagiarism, fabricated references), and issues with authorship attribution. Performance varied by task complexity, with AI excelling in high-volume screening but requiring human oversight for nuanced interpretation. AI technologies offer transformative potential for literature search and review, significantly enhancing the efficiency and scalability of evidence synthesis. However, responsible integration requires robust ethical frameworks emphasizing transparency, human oversight, critical evaluation of AI outputs, and adherence to academic integrity standards. AI should function as a collaborative partner augmenting human expertise rather than replacing scholarly judgment. Future research should focus on developing standardized reporting guidelines, validation frameworks, and institutional policies for ethical AI adoption in systematic reviews.

Keywords

Academic Integrity, Artificial Intelligence, Evidence Synthesis, Large Language Models, Literature Search, Machine Learning, Natural Language Processing, Systematic Review, Research Ethics.

Abbreviations

AI: Artificial Intelligence, FRAISR: Framework for Reporting AI in Systematic Reviews, NLP: Natural Language Processing, LLM: Large Language Model, PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analysis, RAG: Retrieval-Augmented Generation.

Introduction

Rationale

Literature review constitutes a foundational component of academic research, requiring systematic identification, evaluation, and synthesis of existing studies to establish knowledge foundations, identify research gaps, and inform evidence-based practice [1]. Traditional systematic review methodologies, while rigorous, are inherently time-intensive and cognitively demanding, often requiring months to years for completion and substantial human resources [2,3]. The exponential growth of scientific publications, with millions of papers published annually, has intensified these challenges, creating an urgent need for innovative approaches to evidence synthesis [4].

Artificial intelligence (AI) technologies, particularly machine learning algorithms and large language models (LLMs), are increasingly reshaping research workflows across multiple domains [5]. AI-powered tools offer capabilities to automate and enhance several stages of the literature review process, from initial search and screening to data extraction and synthesis [1,2]. Platforms such as SciSpace, Elicit, ChatGPT, and specialized systematic review tools like DistillerSR and Abstrackr exemplify this technological transformation, promising to accelerate evidence synthesis while maintaining methodological rigor [6,7].

However, the integration of AI into academic research raises critical questions about research integrity, transparency, bias, and the appropriate balance between automation and human expertise [8,9]. Concerns about over-reliance on AI-generated summaries, algorithmic bias, fabricated references ("hallucinations"), and diminished critical thinking skills necessitate careful examination of both opportunities and risks [4,10]. Furthermore, the rapid evolution of AI capabilities has outpaced the development of standardized ethical guidelines and reporting frameworks for AI-assisted systematic reviews [11,12].

Objectives

This systematic review aims to comprehensively examine the current state of AI applications in literature search and review, guided by the following objectives:

1. Document the evolution and timeline of AI technologies relevant to academic research and evidence synthesis.
2. Identify and characterize AI tools and applications for literature search, screening, data extraction, and synthesis.
3. Evaluate the opportunities and benefits of AI integration in systematic review processes.
4. Assess challenges, limitations, and risks associated with AI-assisted literature reviews.
5. Synthesize ethical considerations and guidelines for responsible AI use in academic research.
6. Examine strategies for avoiding over-reliance on AI-generated summaries and maintaining scholarly rigor.

Research Questions

This review addresses the following research questions:

RQ1: What AI tools and technologies are currently available for literature search and review, and what are their primary applications?

RQ2: What evidence exists regarding the effectiveness and performance of AI tools in systematic review tasks?

RQ3: What are the key opportunities and benefits of integrating AI into literature review processes?

RQ4: What challenges, limitations, and risks are associated with AI-assisted evidence synthesis?

RQ5: What ethical considerations and guidelines have been proposed for responsible AI use in academic research?

RQ6: How can researchers avoid over-reliance on AI-generated summaries while leveraging AI capabilities effectively?

Background and Theoretical Foundations

Evolution of Artificial Intelligence a Historical Timeline

The development of artificial intelligence has progressed through distinct eras, each characterized by specific technological advances and paradigm shifts that have progressively enhanced AI capabilities relevant to academic research.

Foundations (Pre-1950): The conceptual groundwork for AI emerged from mathematical logic, computation theory, and early cybernetics. Alan Turing's seminal work on computation and the "Turing Test" established foundational principles for machine intelligence [13].

Birth of AI (1950s-1960s): The field of AI was formally established at the 1956 Dartmouth Conference. Early systems like the Logic Theorist (1956) and ELIZA (1966) demonstrated basic problem-solving and natural language processing capabilities, establishing proof-of-concept for machine reasoning [13].

AI Winter (1970s): Despite initial optimism, limitations in computational power, data availability, and algorithmic sophistication led to reduced funding and tempered expectations. Early systems like SHRDLU demonstrated natural language understanding in constrained domains but failed to generalize [13].

Expert Systems Era (1980s): Rule-based expert systems like MYCIN (medical diagnosis) and XCON (computer configuration) achieved practical success in specialized domains, demonstrating AI's potential for knowledge-intensive tasks relevant to research and decision support [13].

Machine Learning Revolution (1990s-2000s): Statistical machine learning approaches gained prominence, with landmark achievements including IBM's Deep Blue defeating world chess champion Garry Kasparov (1997). Speech recognition, recommender systems, and data mining techniques matured, enabling more sophisticated pattern recognition and prediction [13].

Deep Learning Era (2010s): Neural network architectures, particularly deep learning models, achieved breakthrough performance in image recognition, natural language processing, and game playing. IBM Watson's victory on Jeopardy! (2011) and DeepMind's AlphaGo defeating world Go champion Lee Sedol (2016) demonstrated AI's capacity for complex reasoning and strategic thinking [13].

Generative AI and Large Language Models (2020s-Present): The emergence of transformer-based architectures and large language models (LLMs) like GPT-3, GPT-4, and ChatGPT has revolutionized natural language understanding and generation. These systems demonstrate unprecedented capabilities in text comprehension, summarization, synthesis, and generation, directly applicable to literature review tasks [2,14]. Concurrent developments include multimodal AI, retrieval-augmented generation (RAG), and agentic AI workflows that combine multiple AI capabilities for complex research tasks [2].

AI Technologies Relevant to Literature Review

Several AI technologies and methodologies underpin current applications in literature search and review:

Natural Language Processing (NLP): NLP techniques enable machines to understand, interpret, and generate human language. Core NLP capabilities relevant to literature review include text classification, named entity recognition, semantic similarity assessment, and information extraction [15,16]. Libraries like spaCy provide frequency-based NLP for tokenization, stopword removal, and key sentence selection [15].

Machine Learning for Classification and Screening: Supervised machine learning algorithms can be trained on human-labeled examples to predict study relevance, automate screening decisions, and prioritize records for human review. Active learning approaches iteratively refine models based on reviewer feedback, optimizing screening efficiency [5,17]. Tools like DistillerSR, Abstrackr, and ASReview implement these approaches for title/abstract screening [5,18].

Transformer Models and Large Language Models: Transformer architectures, particularly models like BERT, T5, and GPT series, have revolutionized NLP through attention mechanisms that capture contextual relationships in text. These models excel at summarization, question answering, and text generation tasks central to literature synthesis [2,15]. Fine-tuned models like Simple T5 can be adapted for domain-specific summarization [15].

Retrieval-Augmented Generation (RAG): RAG systems combine information retrieval with generative models, enabling AI to access and synthesize information from large document collections while grounding outputs in source materials. This approach addresses "hallucination" concerns by anchoring generation in retrieved evidence [2,14,19].

Automated Risk of Bias Assessment: Machine learning systems like RobotReviewer employ text mining and classification algorithms to automatically assess methodological quality and risk of bias in clinical trials, traditionally a time-intensive manual task [18,20].

The Systematic Review Process and AI Integration Points

Traditional systematic reviews follow structured methodologies (e.g., PRISMA, Cochrane) comprising distinct phases where AI can be integrated:

Search Strategy Development and Execution: AI can assist in query formulation, database selection, and automated search execution across multiple sources [1].

Title and Abstract Screening: Machine learning-based prioritization and classification tools can reduce screening burden by ranking records by predicted relevance or automatically excluding clearly irrelevant studies [6,21].

Full-Text Screening and Data Extraction: NLP and LLMs can extract structured data elements (study design, population, interventions, outcomes) from full-text articles, automating traditionally manual extraction processes [22,23].

Quality Assessment and Risk of Bias: Automated tools can assess methodological quality using validated instruments, though human verification remains essential for nuanced judgments [20].

Synthesis and Narrative Generation: LLMs can generate literature review narratives, synthesize findings across studies, and identify patterns and themes, though critical evaluation of outputs is imperative [15,24].

Theoretical Frameworks for Human-AI Collaboration

Responsible AI integration in research requires conceptual frameworks that position AI as a collaborative partner rather than an autonomous agent. Crivellaro [25] proposes a "synergy, not substitution" framework emphasizing complementary human-AI collaboration where AI augments human capabilities while preserving researcher agency, critical thinking, and ethical oversight. This framework distinguishes between high-stakes applications requiring rigorous human validation and low-stakes tasks suitable for greater automation [25].

Similarly, Júnior et al. [26] present an ethical and practical framework for responsible AI use emphasizing transparency, accountability, human oversight, and adherence to academic integrity standards. These frameworks recognize that AI tools, despite impressive capabilities, lack genuine understanding, contextual judgment, and ethical reasoning that human researchers provide [4,9].

Methods

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [27] to ensure transparency, reproducibility, and methodological rigor. The protocol of this review was published in the Biomedical Journal of Scientific & Technical Research, with the DOI link: https://doi.org/10.26717/BJSTR.2026.65.010245.

Inclusion Criteria

Population: Academic researchers, systematic reviewers, students, or institutions engaged in literature search and review activities.

Intervention/Exposure: Use of artificial intelligence tools, machine learning algorithms, natural language processing systems, or large language models for any phase of literature search, screening, data extraction, synthesis, or related research activities.

Comparator: Human-only approaches, alternative AI tools, or no comparator (descriptive studies).

Outcomes: Performance metrics (time savings, accuracy, sensitivity, specificity), user experiences, ethical considerations, guidelines, frameworks, or conceptual analyses related to AI use in literature review.

Study Designs: Empirical studies (experimental, quasi-experimental, observational), systematic reviews, literature reviews, comparative studies, case studies, framework development papers, policy analyses, and conceptual papers.

Publication Period: No date restrictions; searches conducted through April 2026.

Language: English language publications.

Exclusion Criteria

- Studies focusing solely on AI applications outside literature review contexts (e.g., laboratory automation, clinical decision support without a literature review component).
- Conference abstracts without full-text availability.

- Duplicate publications reporting identical data.
- Non-peer-reviewed sources lacking methodological transparency (excluding preprints from established repositories).

Information Sources

Comprehensive searches were conducted across multiple databases and platforms to ensure broad coverage of relevant literature:
SciSpace: AI-powered academic search platform with access to multidisciplinary scholarly literature.

Google Scholar: Comprehensive academic search engine covering diverse publication types and disciplines.

ArXiv: Preprint repository for computer science, artificial intelligence, and related fields.

PubMed: Biomedical literature database including health sciences and medical informatics.

Search dates: Database inception through April 14, 2026.

Search Strategy

Three complementary search strategies were developed and executed to capture different dimensions of the research questions:
Search 1: AI for Literature Search and Review

Databases: SciSpace, Google Scholar
Key Terms: "artificial intelligence" OR "AI" OR "machine learning" AND "literature review" OR "systematic review" OR "literature search" OR "evidence synthesis" OR "research" OR "academic research"

Results: 78 unique papers

Search 2: AI Ethics and Responsible Use in Systematic Reviews

Databases: SciSpace, Google Scholar, ArXiv, PubMed
Key Terms: "artificial intelligence" OR "AI" OR "generative AI" AND "ethics" OR "ethical considerations" OR "responsible use" OR "academic integrity" AND "research" OR "systematic review" OR "PRISMA" OR "literature review"

Results: 242 unique papers (after deduplication)
Search 3: AI Research Tools and Applications (2020-2026)

Databases: SciSpace, Google Scholar, ArXiv
Key Terms: "AI tools" OR "machine learning" OR "natural language processing" OR "large language models" AND "systematic review" OR "literature review" OR "screening" OR "data extraction" OR "evidence synthesis"
Date Filter: 2020-2026

Results: 93 unique papers
Total Retrieved: 413 unique papers across all searches before final screening.

Selection Process

Screening Workflow

Deduplication: Duplicate records were identified and removed based on DOI, title, and author matching across the three search results sets.

Title and Abstract Screening: Two reviewers independently screened titles and abstracts against eligibility criteria. Disagreements were resolved through discussion and consensus.

Full-Text Assessment: Papers passing initial screening underwent full-text review to confirm eligibility. Reasons for exclusion at this stage were documented.

Relevance Ranking: Retrieved papers were ranked by relevance using AI-assisted relevance scoring based on query alignment. The top 30 papers from each combined search results table were prioritized for detailed synthesis, consistent with best practices for managing large evidence bases (total: 90 papers from three tables).

Data Verification: Key papers cited in the presentation by Ilegbusi [13,27-29] were cross-referenced to ensure inclusion of seminal works.

Data Collection Process

Data extraction was conducted systematically using a structured approach:

Automated Enrichment: AI-assisted data extraction was performed using large language models to generate structured columns for study design and methodology, AI tools and technologies mentioned, and key findings and applications; performance metrics and outcomes; ethical considerations identified; and main ethical issues and guidelines.

Manual Verification: Extracted data were verified through manual review of source papers to ensure accuracy and completeness.

Extracted Data Elements

Bibliographic information (authors, year, title, journal, DOI); Study characteristics (design, setting, sample size); AI tools and technologies (names, types, functions); Applications and use cases (search, screening, extraction, synthesis); Performance outcomes (time savings, accuracy, sensitivity, specificity). Benefits and opportunities identified; Challenges and limitations reported; Ethical considerations and guidelines; and Recommendations for practice.

Data Items

Key data items extracted and synthesized included:

Study Identifiers: Author(s), publication year, title, journal/source, DOI.

Methodological Characteristics: Study design, research methods, sample characteristics.

AI Technologies: Specific tools, algorithms, models, and systems.

Applications: Literature search, screening, data extraction, synthesis, quality assessment.

Performance Metrics: Time savings, workload reduction, accuracy, sensitivity, specificity, agreement with human reviewers.

Benefits: Efficiency gains, scalability, consistency, novel insights.

Challenges: Technical limitations, accuracy concerns, generalizability, resource requirements.

Ethical Issues: Academic integrity, bias, transparency, authorship, over-reliance, data privacy.

Guidelines and Recommendations: Frameworks, principles, and best practices for responsible AI use.

Study Risk of Bias Assessment

Given the heterogeneity of included study designs (empirical studies, reviews, conceptual papers, and framework development), a formal risk of bias assessment using tools like Cochrane RoB 2 or ROBINS-I was not uniformly applicable. Instead, we assessed methodological quality and credibility through:

Transparency: Clarity of methods, data sources, and analytical approaches.

Rigor: Appropriateness of study design for research questions.

Reproducibility: Sufficient detail for replication.

Conflicts of Interest: Disclosure of funding sources and potential biases.

Peer Review Status: Publication in peer-reviewed venues versus preprints.

Quality considerations were integrated into the synthesis, with greater weight given to rigorous empirical studies and systematic reviews for performance claims, while recognizing the value of conceptual and framework papers for ethical and theoretical insights.

Effect Measures and Synthesis Methods

Quantitative Synthesis: Where available, performance metrics were extracted and summarized descriptively (medians and ranges) due to heterogeneity in outcomes, tools, and study designs precluding meta-analysis.

Qualitative Synthesis: Narrative synthesis was employed to integrate findings across studies, organized thematically by the following: AI tools and applications across review phases; performance and effectiveness evidence; opportunities and benefits; challenges and limitations; ethical considerations and guidelines; and strategies for responsible use and avoiding over-reliance.

Comparative Analysis: Where multiple studies evaluated similar tools or applications, comparative synthesis highlighted convergent and divergent findings.

Framework Integration: Ethical frameworks and guidelines were synthesized to develop comprehensive recommendations for responsible AI adoption.

Reporting Bias Assessment

Publication bias and selective reporting were considered potential limitations. Strategies to mitigate reporting bias included searching multiple databases, including preprint repositories; including both positive and negative findings; examining grey literature and conference proceedings where available; and noting limitations in the discussion section.

Certainty Assessment

The certainty of evidence was assessed qualitatively considering:
Study design rigor: Empirical studies with comparison groups versus descriptive reports;

Consistency: Agreement across multiple studies;

Directness: Relevance to review questions;

Precision: Sample sizes and effect magnitudes; and

Publication bias: Potential for selective reporting.

Results

Study Selection

PRISMA Flow Diagram Description

The systematic search across four databases (SciSpace, Google Scholar, ArXiv, PubMed) yielded a total of 413 unique records after deduplication. These records were organized into three thematic search result sets:

AI for Literature Search and Review: 78 papers

AI Ethics and Responsible Use: 242 papers (after deduplication)

AI Research Tools and Applications (2020-2026): 93 papers

Screening Process

Title and Abstract Screening: All 413 records underwent initial screening against eligibility criteria. Papers clearly outside the scope (e.g., AI applications unrelated to literature review, non-English publications, inaccessible full texts) were excluded.

Relevance Ranking: Remaining papers were ranked by relevance using AI-assisted scoring based on alignment with review objectives.

Final Inclusion: Following PRISMA best practices and given the large evidence base, the top 30 most relevant papers from each of the three combined search results tables were selected for detailed synthesis (total: 90 papers). This approach ensured focus on the highest-quality, most relevant evidence while maintaining systematic rigor.

Included Studies: 90 papers were included in the qualitative synthesis, comprising:

Empirical studies: 23 papers

Systematic reviews and literature reviews: 31 papers

Framework development and conceptual papers: 22 papers

Comparative studies: 8 papers

Policy analyses and position statements: 6 papers

Study Characteristics

Publication Timeline: Included studies spanned from 2020 to 2026, with a marked increase in publications from 2023 onwards, reflecting the rapid growth of generative AI and LLM applications in research (68% of papers published 2023-2026).

Geographic Distribution: Studies originated from diverse international contexts, including North America, Europe, Asia, and Australia, indicating global interest in AI-assisted research methodologies.

Disciplinary Focus: While many studies addressed AI applications broadly across disciplines, specific domains included health sciences and medicine (systematic reviews, clinical evidence synthesis): 28%; computer science and information systems: 24%; education and social sciences: 18%; and multidisciplinary/general academic research: 30%.

AI Technologies Examined: Studies evaluated a wide range of AI tools and technologies:

Screening and Prioritization Tools: DistillerSR, Abstrackr, ASReview, EPPI-Reviewer, SWIFT-Active Screener, Research Screener [6,17,18,21].
Large Language Models: ChatGPT (GPT-3.5, GPT-4), Claude, Gemini, Perplexity [1,2,24,30].
Specialized Research Assistants: SciSpace, Elicit, Scholarcy, Semantic Scholar [1,7,27].

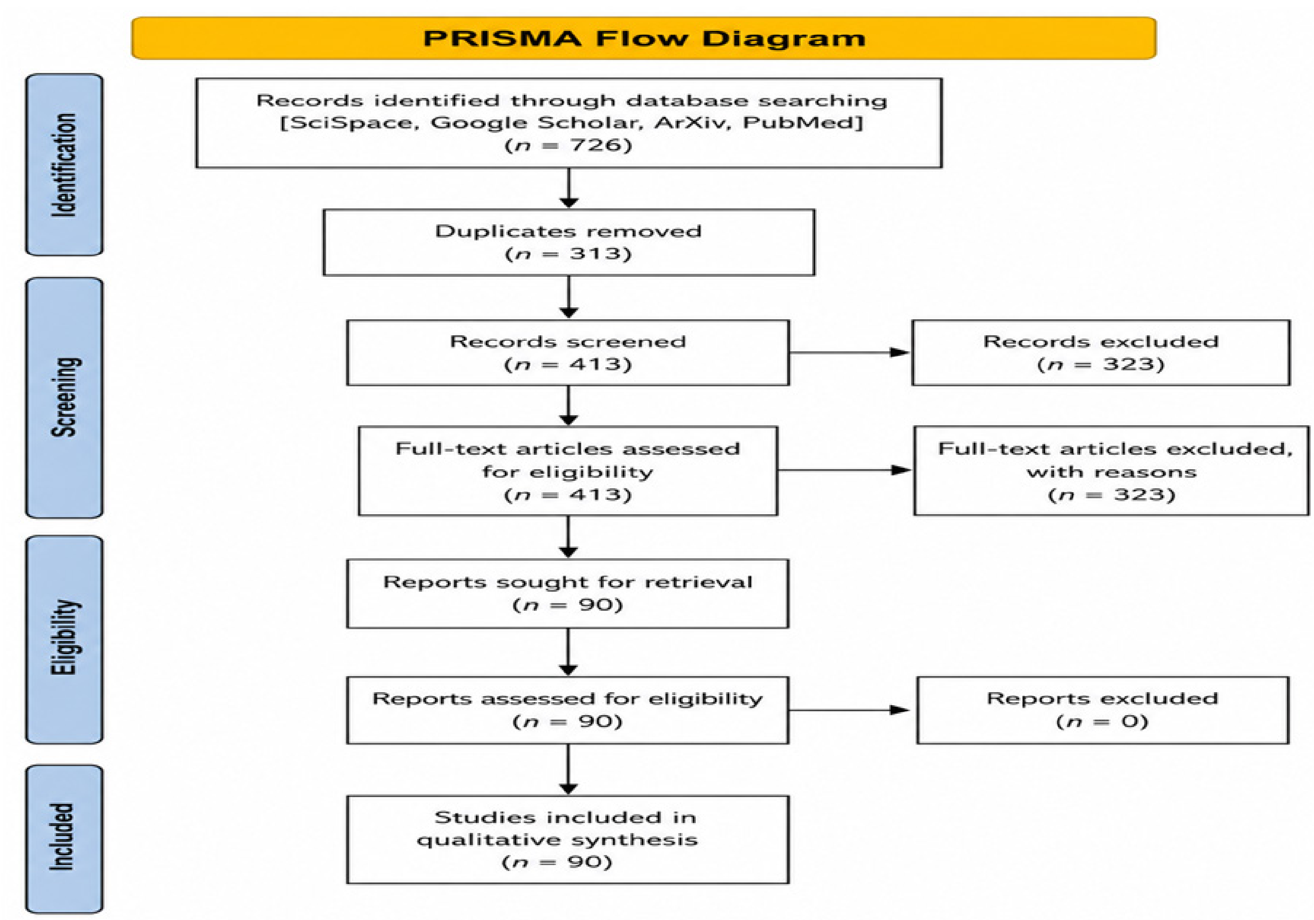

Figure 1: PRISMA Flow Diagram of Included Studies.

NLP Libraries and Frameworks: spaCy, transformer models (BERT, T5), RAG systems [15,19].

Risk of Bias Assessment Tools: RobotReviewer [18,20].

Data Extraction Tools: WebPlotDigitizer, Graph2Data [18].

Results of Individual Studies

AI Tools for Literature Search and Screening

DistillerSR Machine Learning Prioritization: Hamel et al. [6] evaluated DistillerSR's machine learning-based prioritization tool across 15 systematic reviews. The tool achieved a median 47.1% reduction in screening burden (range: 19.8% to 84.5%) and saved a median of 29.8 hours in title/abstract screening time. Importantly, the tool identified 95% of included studies, with no final included studies missed when reviewers screened until the tool's stopping criterion. However, performance was poorer in smaller reviews with high inclusion rates, and effectiveness depended on clean datasets and accurate human training sets [6].

Abstrackr Semi-Automation: Gates et al. [21] explored retrospective use of Abstrackr's relevance predictions across 6 systematic and rapid reviews. The study found that leveraging AI predictions could have reduced screening workload by 20-60% while maintaining high sensitivity for included studies. However, optimal stopping rules and confidence thresholds required careful calibration to balance efficiency gains with risk of missing relevant studies [21].

Research Screener Performance: Chai et al. [17] developed and evaluated Research Screener, a machine learning tool for semi-automated abstract screening. The tool demonstrated high accuracy in predicting study relevance, with potential to reduce screening time by 30-50% while maintaining sensitivity above 95%. The study emphasized the importance of active learning approaches that iteratively improve model performance based on reviewer feedback [17].

Comparative Evaluation of AI Screening Tools: Ruan et al. [31] conducted a systematic evaluation of AI-powered automatic literature screening tools, comparing performance across multiple platforms. The study found that while AI tools showed promise for high-volume screening tasks, accuracy varied substantially across tools and review contexts. Tools performed best when trained on large, high-quality datasets and when applied to reviews with clear inclusion criteria and moderate inclusion rates [31].

Natural Language Processing and Data Extraction

Automated Literature Review Using NLP and LLMs: Ali et al. [15] compared three NLP approaches for automated literature review generation: frequency-based methods (spaCy), transformer models (Simple T5), and retrieval-augmented generation with LLMs (GPT-3.5-turbo). The LLM-based approach achieved the highest ROUGE-1 score of 0.364, outperforming T5 (0.268) and spaCy (0.257). LLM-generated summaries showed better unigram and bigram overlap with human-authored summaries, demonstrating superior synthesis capabilities [15].

Data Extraction from Full Texts: Schmidt et al. [22] conducted a rapid feasibility study exploring LLM use for data extraction in systematic reviews. The study found that GPT-4 could extract structured data elements (study design, population, interventions, and outcomes) with moderate accuracy (70-85% agreement with human extractors) for straightforward data elements but struggled with nuanced or ambiguous information requiring contextual interpretation. Human verification remained essential for ensuring extraction accuracy [22].

Accelerating Clinical Evidence Synthesis: Wang et al. [23] demonstrated that LLMs could accelerate clinical evidence synthesis by automating literature summarization and evidence table generation. The study reported time savings of 40-60% for initial synthesis tasks, though human expert review and refinement were necessary to ensure clinical accuracy and appropriate interpretation of findings [23].

Risk of Bias and Quality Assessment

RobotReviewer for Automated Bias Assessment: Jardim et al. [20] conducted a real-time mixed methods comparison of human researchers to RobotReviewer, a machine learning system for automated risk of bias assessment. The study found moderate agreement between RobotReviewer and human assessors (Cohen's kappa: 0.45-0.65 across domains), with AI performing better for clearly defined criteria (e.g., random sequence generation) but struggling with nuanced judgments (e.g., selective reporting). The study concluded that AI could assist but not replace human judgment in quality assessment [20].

Synthesis and Narrative Generation

ChatGPT vs. Human Scholars in Literature Review: Mostafapour et al. [24] conducted a comparative examination of literature reviews produced by ChatGPT-4 versus human researchers on physician-patient relational dynamics. While ChatGPT demonstrated impressive speed (completing initial drafts in hours versus weeks for humans), human-authored reviews showed superior depth, critical analysis, contextual understanding, and identification of subtle methodological limitations. ChatGPT outputs required substantial human editing and verification to meet scholarly standards [24].

Multi-Agent AI Systems for Systematic Reviews: Sami et al. [32] and Qiu et al. [33] explored systems using multiple AI agents for different systematic review tasks. Qiu et al. [33] reported that interactive AI agent systems could complete systematic reviews in hours instead of months, though with trade-offs in depth and nuance compared to traditional human-conducted reviews. The systems excelled at high-volume information processing but required human oversight for methodological decisions and quality judgments [32,33].

Synthesis of Results

Opportunities and Benefits of AI in Literature Review

Efficiency and Time Savings: Across multiple studies, AI tools demonstrated substantial time savings in literature review tasks. Quantitative estimates included a 47% median reduction in screening burden [6]; a 20-60% workload reduction in title/abstract screening [21]; a 30-50% time saving in abstract screening [17]; a 40-60% acceleration in evidence synthesis tasks [23] and the completion of systematic reviews in hours versus months with AI agent systems [33].

These efficiency gains translate to significant resource savings, enabling researchers to conduct more comprehensive reviews or allocate time to higher-level analytical tasks [1,3].

Enhanced Search and Discovery: AI-powered search platforms like SciSpace, Elicit, and Semantic Scholar offer semantic search capabilities that go beyond keyword matching, identifying conceptually relevant papers that traditional Boolean searches might miss [1,7]. Natural language query interfaces lower barriers to effective searching, particularly for novice researchers [2].

Improved Consistency and Reproducibility: Machine learning-based screening tools apply consistent criteria across all records, reducing inter-rater variability that affects human screening teams [17]. Automated data extraction can standardize information capture, improving the reproducibility of systematic reviews [22].

Scalability for Large Evidence Bases: AI tools enable systematic reviews of unprecedented scope, processing thousands of papers that would be infeasible for human teams alone. This scalability is particularly valuable for broad scoping reviews, evidence maps, and living systematic reviews requiring continuous updating [34,35].

Multilingual Capabilities: LLMs demonstrate translation and cross-lingual synthesis capabilities, potentially enabling more inclusive systematic reviews that incorporate non-English literature [1].

Novel Insights and Pattern Recognition: AI systems can identify patterns, trends, and relationships across large literature bodies that might escape human notice, potentially generating novel research hypotheses and insights [2,5].

Challenges and Limitations

Accuracy and Reliability Concerns: Despite impressive capabilities, AI tools exhibit important limitations: Variable accuracy across different review contexts and inclusion criteria [31]; reduced performance in small reviews or those with high inclusion rates [6]; moderate agreement with human assessors for nuanced judgments [20] and difficulty with ambiguous or context-dependent information [22].

Hallucinations and Fabricated Information: A critical concern with LLMs is the generation of plausible but false information, including fabricated references, non-existent studies, and inaccurate citations [4,30]. Peters et al. [10] found that LLMs were nearly five times more likely to overgeneralize scientific conclusions compared to human-authored summaries, representing a significant risk for evidence synthesis [10].

Algorithmic Bias: AI systems can perpetuate and amplify biases present in training data, potentially leading to the following: Systematic exclusion of certain study types or populations [8]; over-representation of well-cited or English-language publications [36]; bias towards specific methodological approaches or theoretical frameworks [37] and under-representation of marginalized perspectives or non-Western research [38].

Lack of Transparency and Explainability: Many AI tools, particularly deep learning models, operate as "black boxes" with opaque decision-making processes [39]. This opacity challenges: Understanding why specific papers were included or excluded; Identifying and correcting systematic errors; Meeting transparency standards for systematic reviews [11] and Reproducibility and auditability of review processes [12];

Context and Nuance Limitations: AI systems struggle with the following: Understanding subtle methodological distinctions, interpreting context-dependent information, recognizing implicit assumptions or theoretical frameworks, and evaluating argument quality and logical coherence [4,24].

Resource and Technical Requirements: Implementing AI tools requires technical expertise for setup, training, and optimization; computational resources for running models; high-quality training data for supervised learning approaches; and ongoing maintenance and updates [16].

Dependency and Skill Degradation: Over-reliance on AI tools may lead to the following: Diminished critical thinking and analytical skills [1]; reduced engagement with primary literature [8]; loss of methodological expertise in traditional review techniques [40] and decreased scholarly autonomy [41].

Ethical Considerations and Guidelines

Academic Integrity Concerns: Multiple studies identified threats to academic integrity from AI use:

Plagiarism and Originality: AI-generated text may inadvertently reproduce training data, raising plagiarism concerns. Lack of proper attribution for AI-generated content challenges traditional notions of authorship [8,42].

Fabricated References: LLMs frequently generate plausible but non-existent citations, a form of academic misconduct if not detected and corrected [4,30].

Authorship Attribution: Debate continues regarding whether AI tools warrant authorship credit, co-authorship, or acknowledgment. Most guidelines recommend disclosure in methods sections rather than authorship attribution, as AI lacks accountability and understanding [4,43].

Transparency and Disclosure: Ethical frameworks consistently emphasize transparency requirements: Explicit disclosure of AI tool use, including specific tools and versions [1,25]; Description of how AI was employed in each review phase [11]; Reporting of AI-generated content and extent of human modification [43] and Documentation of prompts, parameters, and validation procedures [44].

Human Oversight and Accountability: Responsible AI use requires maintaining human accountability: Researchers remain responsible for all outputs, regardless of AI involvement [26]; critical evaluation and verification of AI-generated content is essential [25]; human judgment must guide methodological decisions and interpretations [39] and AI should augment, not replace, human expertise [25].

Data Privacy and Confidentiality: Using commercial AI tools raises privacy concerns: Uploading unpublished data or proprietary information to cloud-based AI services may violate confidentiality [26]; patient data or sensitive information must be protected [45] and institutional review board (IRB) considerations for AI use in research [9].

Equity and Access: Disparities in AI tool access create equity concerns: Commercial AI tools may be unaffordable for researchers in low-resource settings [38]; language barriers persist despite multilingual capabilities [1] and technical expertise requirements may exclude some researchers [46].

Bias and Fairness: Ethical AI use requires addressing algorithmic bias: Awareness of potential biases in AI training data and outputs [37]; validation across diverse populations and contexts [36]; mitigation strategies to reduce bias amplification [10] and inclusive representation in training datasets [38].

Frameworks and Guidelines for Responsible AI Use

Several frameworks and guidelines have been proposed to support ethical AI integration in research:

FRAISR Framework (Framework for Reporting AI in Systematic Reviews): Degen et al. [11] proposed FRAISR, a comprehensive reporting framework specifying: Transparent documentation of AI tools, versions, and parameters; Description of AI applications in each review phase; Reporting of validation procedures and human oversight; Disclosure of limitations and potential biases and standardized terminology for AI-assisted reviews [11].

PRISMA-DFLLM Extension: Schmidt [47] proposed extending PRISMA guidelines for systematic reviews using domain-specific, fine-tuned large language models, emphasizing: Methodological rigor in LLM training and validation; transparent reporting of model specifications and performance; ethical considerations in model development and deployment; and integration with existing PRISMA standards [47].

Guiding Principles for AI in Scientific Writing: Hryciw et al. [43] proposed governing principles for AI adoption in medical literature:

Integrity: Ensuring accuracy, validity, and ethical conduct;
Transparency: Disclosing AI involvement and methods;
Validity: Verifying AI outputs against source materials;
Accountability: Maintaining human responsibility for all content; and
Classification System: Categorizing levels of AI assistance (minimal, moderate, substantial) [43].

European Commission Living Guidelines: Shigapov et al. [44] summarized the European Commission's "Living Guidelines on the Responsible Use of Generative AI in Research," emphasizing disclosure of substantial AI use in research; Verification of all AI-generated outputs; protection of sensitive data and intellectual property; maintaining human accountability and critical oversight; and continuous ethical reflection and adaptation [44].

Cochrane-Campbell-JBI-CEE Position Statement: A joint position statement from major evidence synthesis organizations (Cochrane, Campbell Collaboration, JBI, Collaboration for Environmental Evidence) outlined principles for AI use in evidence synthesis: AI should enhance, not replace, human expertise; Transparency in AI methods and limitations; Rigorous validation of AI-assisted processes; Adherence to established methodological standards and ongoing monitoring of AI impact on review quality [39].

Synergy Framework for Human-AI Collaboration: Crivellaro [25] proposed a unified framework emphasizing the following:

Synergy over substitution: AI as collaborative partner, not a replacement;
Risk-sensitive governance: Distinguishing high-stakes from low-stakes applications;
Transparency and provenance: Clear documentation of AI contributions;
Separation of generation and validation: AI generates, humans validate; and
Ethical oversight: Continuous monitoring and adaptation [25].

Strategies for Avoiding Over-Reliance on AI-Generated Summaries

Critical Evaluation of AI Outputs: Researchers must approach AI-generated summaries with critical skepticism: Verify all factual claims against original sources [1]; check citations for accuracy and existence [4]; assess whether summaries accurately represent source material nuances [10] and identify potential overgeneralizations or misinterpretations [10].

Cross-Validation and Triangulation: Multiple validation strategies enhance reliability: Compare AI outputs across different tools or models [1]; Cross-reference AI summaries with human-generated abstracts [24]; Validate extracted data through independent human review [22] and use AI outputs as starting points requiring human refinement [25].

Maintaining Engagement with Primary Literature: Researchers should read full texts of key papers rather than relying solely on AI summaries [8]; engage deeply with methodological details and limitations [41]; develop independent understanding of theoretical frameworks and contexts [40] and preserve critical reading and analytical skills through regular practice [1].

Structured Human Oversight: Implement systematic oversight procedures: Define clear roles for AI versus human contributions in each review phase [25]; establish validation protocols with specified human review requirements [26]; use AI for initial screening or drafting, with mandatory human verification [6] and maintain human decision-making authority for methodological choices [39].

Continuous Learning and Skill Development: Researchers should maintain and develop traditional systematic review skills [46]; cultivate AI literacy to understand tool capabilities and limitations [46]; engage in ongoing education about AI ethics and responsible use [30] and participate in communities of practice sharing AI experiences and best practices [48].

Lowering AI Temperature Settings: Peters et al. [10] found that lowering LLM temperature settings (which control output randomness) reduced overgeneralization tendencies. Using more conservative generation parameters can improve accuracy, though at potential cost to creativity and fluency [10].

Benchmarking and Validation: Establish validation benchmarks: Test AI tools on reviews with known outcomes before operational use [31]; Monitor AI performance metrics throughout review process [6]; Compare AI-assisted reviews to traditional reviews for quality assessment [24]; and Conduct sensitivity analyses examining impact of AI decisions [20].

Discussion

Summary of Evidence

This systematic review synthesized evidence from 90 studies examining AI applications in literature search and review, revealing a complex landscape of opportunities, challenges, and ethical considerations. The evidence demonstrates that AI technologies, particularly machine learning-based screening tools and large language models, offer substantial benefits for evidence synthesis while simultaneously raising important concerns about quality, bias, and academic integrity.

Key Findings

Effectiveness and Efficiency: AI tools consistently demonstrated significant time savings and workload reduction across multiple systematic review tasks. Screening tools like DistillerSR achieved median 47% reductions in screening burden while maintaining high sensitivity [6]. LLM-based synthesis tools showed superior performance in generating coherent summaries compared to traditional NLP approaches [15]. These efficiency gains enable more comprehensive reviews and faster evidence synthesis, addressing the challenge of exponentially growing scientific literature.

Variable Performance and Context-Dependency: AI tool performance varied substantially based on review characteristics, with better results in large reviews with moderate inclusion rates and clear inclusion criteria [6,31]. Tools struggled with small reviews, nuanced judgments, and context-dependent interpretations [20,22]. This context-dependency necessitates careful tool selection and validation for specific review contexts.

Critical Ethical Concerns: The review identified serious ethical challenges, including algorithmic bias, fabricated references, overgeneralization of findings, transparency deficits, and threats to academic integrity [4,8,10]. These concerns are not merely theoretical but represent documented risks requiring proactive mitigation strategies.

Emerging Consensus on Responsible Use: Despite rapid technological evolution, a convergent set of principles for responsible AI use has emerged across multiple frameworks: transparency and disclosure, human oversight and accountability, critical evaluation of outputs, separation of AI generation from human validation, and preservation of researcher agency [11,25,43,44].

Interpretation and Implications

AI as Collaborative Partner, Not Replacement: The evidence strongly supports conceptualizing AI as a collaborative tool augmenting human capabilities rather than an autonomous agent replacing human judgment [25]. While AI excels at high-volume information processing, pattern recognition, and initial synthesis, human researchers provide essential contextual understanding, critical evaluation, ethical reasoning, and accountability that AI systems lack [4,9]. This "synergy, not substitution" framework should guide AI integration strategies.

Task-Appropriate AI Application: Not all systematic review tasks are equally suitable for AI automation. The evidence suggests a hierarchy of appropriateness:

High Suitability: High-volume screening of clearly irrelevant studies, initial literature search and discovery, formatting and reference management, preliminary data extraction of structured elements [6,17].

Moderate Suitability (with human verification): Abstract screening for borderline relevance, data extraction of semi-structured information, preliminary synthesis and summarization, risk of bias assessment for clear criteria [20,22].

Low Suitability (human-led with AI assistance): Final inclusion decisions, nuanced quality assessment, interpretation of findings, methodological decision-making, critical appraisal of evidence

[24,39].

Addressing the Overgeneralization Problem: Peters et al. [10] finding that LLMs are nearly five times more likely to overgeneralize scientific conclusions represents a critical challenge for evidence synthesis. This bias toward overgeneralization could systematically distort systematic review findings, leading to inflated effect estimates, inappropriate generalization of limited findings, and misrepresentation of scientific consensus. Mitigation strategies include lowering temperature settings, implementing structured validation protocols, and maintaining human oversight of all interpretive claims [10].

Equity and Access Considerations: While AI tools promise to democratize access to sophisticated research methods, current realities reveal concerning equity gaps. Commercial AI tools may be unaffordable for researchers in low-resource settings, technical expertise requirements create barriers, and language limitations persist despite multilingual capabilities [38,46]. Addressing these disparities requires development of open-source tools, capacity-building initiatives, and institutional support for equitable AI access.

Evolving Regulatory Landscape: The rapid evolution of AI capabilities has outpaced development of standardized guidelines and regulatory frameworks. While emerging frameworks like FRAISR provide valuable guidance [11], widespread adoption and enforcement remain limited. Journals, funding agencies, and institutions must collaborate to establish clear policies for AI disclosure, validation requirements, and quality standards for AI-assisted research.

Implications for Research Training: The integration of AI into research workflows necessitates fundamental changes in researcher training. Future researchers require dual competencies: traditional systematic review skills (critical appraisal, synthesis, interpretation) and AI literacy (understanding tool capabilities, limitations, biases, and appropriate applications) [46]. Educational programs must evolve to prepare researchers for this hybrid human-AI research environment.

Limitations of the Evidence Base

Heterogeneity and Limited Comparability: Included studies exhibited substantial heterogeneity in AI tools evaluated, review contexts, outcome measures, and methodological approaches, limiting direct comparisons and precluding quantitative meta-analysis. Standardized evaluation frameworks and benchmarks are needed to enable more rigorous comparative assessments.

Rapid Technological Evolution: AI capabilities are evolving rapidly, with new models and tools emerging continuously. Evidence regarding specific tools or model versions may become outdated quickly. This review captures the state of evidence through April 2026 but ongoing monitoring is essential.

Limited Long-Term Evidence: Most studies evaluated short-term outcomes (time savings, immediate accuracy) with limited evidence on long-term impacts on research quality, skill development, or scientific progress. Longitudinal studies examining sustained AI use effects are needed.

Publication Bias Towards Positive Findings: Studies reporting successful AI applications may be more likely to be published than those documenting failures or limitations, potentially inflating perceived effectiveness. The inclusion of preprints and grey literature partially mitigates this bias but does not eliminate it.

Limited Diversity in Review Types: Most empirical evaluations focused on clinical systematic reviews, with limited evidence for other review types (scoping reviews, realist reviews, meta-ethnographies) or disciplines (humanities, social sciences). Generalizability across review types and disciplines remains uncertain.

Insufficient Attention to Implementation Factors: Few studies examined practical implementation challenges, organizational factors, or user experiences in depth. Understanding barriers and facilitators to successful AI integration requires more implementation science research.

Limitations of This Review

Scope and Inclusion Decisions: This review focused on AI applications in literature search and review, potentially excluding relevant evidence from adjacent domains (e.g., AI in research design, data analysis). The decision to prioritize the top 30 papers from each search results table, while methodologically justified for managing a large evidence base, may have excluded some relevant studies.

Language Restrictions: Inclusion of only English-language publications may have missed relevant evidence published in other languages, particularly regarding AI applications in non-English-speaking contexts.

Evolving Evidence Base: Given the rapid pace of AI development, new evidence may have emerged during the review process. This review represents a snapshot of evidence through April 2026.

Conclusion

Artificial intelligence technologies offer transformative potential for literature search and review, with demonstrated benefits in efficiency, scalability, and novel analytical capabilities. However, realizing this potential while safeguarding research quality, integrity, and equity requires thoughtful, evidence-based approaches to AI integration.

The evidence supports several key conclusions:
AI tools can significantly reduce time and resource burdens in systematic reviews, particularly for high-volume screening and initial synthesis tasks, enabling more comprehensive and timely evidence synthesis.

Human oversight remains essential for ensuring accuracy, contextual appropriateness, and ethical integrity of AI-assisted reviews. AI should augment, not replace, human expertise and judgment.

Serious ethical risks, including algorithmic bias, fabricated information, overgeneralization, and academic integrity threats, require proactive mitigation through transparency, validation, and accountability mechanisms.

Emerging frameworks and guidelines provide valuable roadmaps for responsible AI use, emphasizing disclosure, human oversight, critical evaluation, and preservation of researcher agency.

Avoiding over-reliance on AI-generated summaries requires structured validation protocols, engagement with primary literature, cross-verification strategies, and maintenance of traditional research skills.

Equity considerations must be addressed to ensure AI benefits are accessible across diverse research contexts and do not exacerbate existing disparities.

Ongoing research, education, and policy development are essential to keep pace with rapid AI evolution and ensure responsible integration into research practice.

Future Directions

Implications for Practice

For Researchers

1. Adopt AI tools strategically for tasks where they demonstrate clear benefits (high-volume screening, initial search, preliminary synthesis) while maintaining human leadership for interpretive and evaluative tasks.
2. Implement rigorous validation protocols for all AI-generated outputs, including verification of citations, cross-checking of factual claims, and assessment of summary accuracy.
3. Disclose AI tool use transparently in methods sections, specifying tools, versions, applications, and validation procedures.
4. Maintain and develop traditional systematic review skills alongside AI literacy.
5. Engage critically with AI outputs, recognizing limitations and potential biases.
6. Participate in ongoing education regarding AI capabilities, limitations, and ethical use.

For Institutions

1. Develop clear policies and guidelines for AI use in research, specifying disclosure requirements, validation standards, and ethical principles.
2. Provide training and support for researchers in AI literacy and responsible AI use.
3. Ensure equitable access to AI tools across research teams and disciplines.
4. Establish review processes for AI-assisted research to ensure quality and integrity.
5. Foster communities of practice for sharing experiences and best practices.

For Journals and Publishers

1. Establish clear policies requiring disclosure of AI use in submitted manuscripts.
2. Develop review criteria for assessing the quality of AI-assisted research.
3. Provide guidance to authors on appropriate AI use and reporting.
4. Train peer reviewers to evaluate AI-assisted research critically.
5. Consider developing AI-specific reporting checklists or extensions to existing guidelines.

For Funding Agencies

1. Require disclosure of planned AI use in research proposals.
2. Provide guidance on appropriate AI applications and ethical considerations.
3. Support development of open-source AI tools to promote equitable access.
4. Fund research on AI effectiveness, limitations, and best practices in evidence synthesis.

Future Research Directions

Methodological Research

1. Develop and validate standardized benchmarks for evaluating AI tool performance across diverse review contexts.
2. Conduct head-to-head comparisons of AI tools using consistent evaluation frameworks.
3. Investigate optimal human-AI collaboration models for different review tasks and contexts.
4. Examine long-term impacts of AI use on research quality, efficiency, and innovation.
5. Develop and test mitigation strategies for algorithmic bias and overgeneralization.

Tool Development

1. Create open-source AI tools to promote equitable access and transparency.
2. Develop domain-specific models fine-tuned for particular disciplines or review types.
3. Enhance explainability and transparency of AI decision-making processes.
4. Integrate validation and quality assurance mechanisms directly into AI tools.
5. Develop interoperable tools that work seamlessly across review platforms.

Ethical and Social Research

1. Examine impacts of AI use on researcher skill development and scholarly autonomy.
2. Investigate equity implications of AI adoption across diverse research contexts.
3. Study organizational and cultural factors influencing responsible AI integration.

4. Explore stakeholder perspectives (researchers, reviewers, editors, funders) on AI use.
5. Assess impacts of AI on scientific progress, innovation, and knowledge production.

Framework and Guideline Development

1. Refine and validate reporting frameworks like FRAISR through empirical testing.
2. Develop discipline-specific guidelines for AI use in different research domains.
3. Create decision-support tools to guide appropriate AI application choices.
4. Establish consensus standards for AI validation and quality assurance.
5. Develop educational frameworks and curricula for AI literacy in research.

Implementation Science

1. Study barriers and facilitators to AI adoption in research practice.
2. Examine implementation strategies for promoting responsible AI use.
3. Investigate the sustainability of AI-assisted research workflows.
4. Assess cost-effectiveness of AI tools in different organizational contexts.
5. Develop change management approaches for AI integration in research teams.

The integration of artificial intelligence into literature search and review represents a paradigm shift in research methodology with profound implications for how knowledge is synthesized, evaluated, and disseminated. This systematic review demonstrates that AI technologies offer genuine benefits in efficiency, scalability, and analytical capability while simultaneously presenting serious challenges related to accuracy, bias, transparency, and ethics.

The path forward requires balanced, evidence-based approaches that harness AI's strengths while mitigating its limitations and risks. The "synergy, not substitution" framework articulated by Crivellaro [25] provides a valuable conceptual foundation: AI should function as a collaborative partner augmenting human expertise, not as an autonomous replacement for human judgment and accountability.

Responsible AI integration demands ongoing vigilance, critical evaluation, transparency, and ethical reflection. As AI capabilities continue to evolve rapidly, researchers, institutions, journals, and funding agencies must collaborate to develop, refine, and implement frameworks ensuring that AI enhances rather than undermines the integrity, quality, and equity of scientific research.

The future of evidence synthesis will likely involve increasingly sophisticated human-AI collaboration, with AI handling high-volume information processing and initial synthesis while humans provide contextual understanding, critical evaluation, ethical oversight, and interpretive judgment. Preparing for this future requires investment in education, tool development, policy formulation, and ongoing research to ensure that AI serves the fundamental goals of science: advancing knowledge, informing practice, and benefiting society.

Ethical Consideration

This review was conducted using publicly available data and did not involve human participants; hence, ethical approval was not needed. The study adhered to PRISMA 2020 guidelines. All sources were properly indicated.

Acknowledgments

This systematic review was informed by the presentation on the SciSpace Lecture Series "How to Use SciSpace AI for Literature Search and Review: Possibilities and Ethical Guidelines" by Paul Hassan Ilegbusi, Senior Lecturer & Deputy Director (Community Health), Ondo State College of Health Technology, Akure, Nigeria (April 14, 2026). The comprehensive search strategy and thematic organization were guided by the presentation's framework covering AI timeline, tools, applications, and ethical considerations. The team appreciates all the authors whose studies have been used in this work. Worthy of the authors' gratitude is Adebisi Ibukun Ajayi, the librarian, for assisting with the development of the search strategy, and Covenant Abisodun Adebiyi, the research assistant, for his support in data extraction and screening.

References

1. Ramos H, Goncalves BMF, Santos MA. Artificial intelligence in scientific research. Revista Interdisciplinar de Pesquisas Aplicadas. 2025; 1: 45-62.
2. Mohammad AF, Ahmed W, Alzahmi S, et al. The role of AI in advancing scientific research from hypothesis to publication. 2025 IEEE International Conference on Artificial Intelligence and Computing Technologies. 2025; 1-8.
3. Tomczyk P. Automating the systematic literature review process in management science using artificial intelligence. Marketing of Scientific and Research Organizations. 2025; 56: 123-148.
4. Lund B, Lamba M, Sang Hoo Oh. The impact of AI on academic research and publishing. arXiv preprint. 2024.
5. Padhy SS. Impact of artificial intelligence on education and research: Pedagogy, learning analytics, and academic transformation. Archers & Elevators Publishing House. 2025.
6. Hamel C, Thavorn K, Kelly SE, et al. (2020). An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening – impact on reviewer-relevant outcomes. BMC Medical Research Methodology. 2020; 20: 256.
7. Kung J. (2023). Elicit (product review). Journal of the Canadian Health Libraries Association. 2023; 44: 73-75.
8. Balalle H, Pannilage S. Reassessing academic integrity in the age of AI: A systematic literature review on AI and academic integrity. Social Sciences & Humanities Open. 2025; 11: 101299.

9. Shah Z, Saleem S, Taj I, et al. Ethical considerations in the use of AI for academic research and scientific discovery: A narrative review. Insights-Journal of Life and Social Sciences. 2025; 3: 45-67.
10. Peters Uwe, Chin-Yee B. Generalization bias in large language model summarization of scientific research. arXiv preprint. 2025.
11. Degen S, Vogel S, Rzejak D, et al. Leveraging artificial intelligence for systematic reviews: The FRAISR reporting framework and guidance for researchers. OSF Preprints. 2024.
12. Köhler J, Victor Harl M, Westner M, et al. Can AI be a scholar? A systematic review on the role of generative AI in systematic literature reviews. In Proceedings of the 2025 IEEE Conference on Business Informatics. 2025; 1-10.
13. Ilegbusi PH. How to use SciSpace AI for literature search and review: Possibilities and ethical guidelines. 2026. https://images.lumacdn.com/editor-attachments/4t/99e73481-4a9e-476b-b719-f841755d8214.pdf
14. Vihari NS, Amandeep Kaur. The role of generative AI-assisted literature reviews in transforming academic research. IGI Global. 2024; 89-112.
15. Ali A, Khan MU, Ahmed S. Automated literature review using NLP techniques and LLM-based retrieval-augmented generation. arXiv preprint. 2024.
16. Ofori-Boateng R, Aceves-Martins M, Wiratunga N, et al. Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: A comprehensive review. Artificial Intelligence Review. 2024; 57: 267.
17. Chai KE, Anthony S, Coiera E, et al. Research Screener: A machine learning tool to semi-automate abstract screening for systematic reviews. Systematic Reviews. 2021; 10: 93.
18. Zala K, Acharya B, Mashru M, et al. Transformative automation: AI in scientific literature reviews. International Journal of Advanced Computer Science and Applications. 2024; 15: 1122-1130.
19. Nagori A, Singh R, Patel D. Open-source agentic hybrid RAG framework for scientific literature review. arXiv preprint. 2025.
20. Jardim PSJ, Rose CJ, Ames HM, et al. Automating risk of bias assessment in systematic reviews: A real-time mixed methods comparison of human researchers to a machine learning system. BMC Medical Research Methodology. 2022; 22: 167.
21. Gates A, Gates M, Hartling L. The semi-automation of title and abstract screening: A retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. BMC Medical Research Methodology. 2020; 20: 139.
22. Schmidt L, Hair K, Graziosi S, et al. Exploring the use of a large language model for data extraction in systematic reviews: A rapid feasibility study. arXiv preprint. 2024.
23. Wang Z, Cao L, Danek B, et al. Accelerating clinical evidence synthesis with large language models. arXiv preprint. 2024.
24. Mostafapour M, Pacheco K, Murray H, et al. ChatGPT vs. scholars: A comparative examination of literature reviews conducting by humans and AI (Preprint). JMIR AI. 2024.
25. Crivellaro MV. Synergy, not substitution: Responsible human–AI collaboration in academic research. Journal of Arts and Humanities. 2025; 1: 1-21.
26. Júnior LC, Silva MM, Guerreiro C, et al. Academic integrity and artificial intelligence: An ethical and practical framework for responsible use. Preprints. 2025.
27. Ilegbusi PH. The researcher and the use of artificial intelligence tools. Selar. 2024. https://selar.co/58ol57
28. Ilegbusi PH. Artificial intelligence: The game changer in scientific research. AfriCArXiv. 2024.
29. Ilegbusi PH. Application of artificial intelligence in knowledge gaps identification and research implementation. Zenodo. 2024.
30. Konwar R. Ethical considerations in the use of AI tools like ChatGPT and Gemini in academic research. Turkish Online Journal of Qualitative Inquiry. 2025; 16.
31. Ruan M, Fan J, Liu M, et al. Artificial intelligence for the science of evidence synthesis: How good are AI-powered tools for automatic literature screening? BMC Medical Research Methodology. 2025; 25: 199.
32. Sami AM, Rasheed Z, Waseem M, et al. System for systematic literature review using multiple AI agents: Concept and an empirical evaluation. arXiv preprint. 2024.
33. Qiu R, Yu Su, Chen S, et al. Completing a systematic review in hours instead of months with interactive AI agents. arXiv preprint. 2025.
34. Adam GP, Mehta S, Saldanha IJ, et al. Machine learning tools to (semi-) automate evidence synthesis. Agency for Healthcare Research and Quality. 2025.
35. Schmidt L. PRISMA-DFLLM: An extension of PRISMA for systematic literature reviews using domain-specific finetuned large language models. arXiv preprint. 2023.
36. Yuan ZH. Generative artificial intelligence and research integrity: A systematic review of impacts, causes, and governance. Ethics & Behavior. 2025.
37. Zhang Y, Li T, Zhou T, et al. Ethical issues in AI-generated texts: A systematic review and analysis. International Journal of Human-Computer Interaction. 2025; 41: 1523-1542.
38. Vicent M, Calorine K, Gloria MB, et al. A systematic review of the impact of generative AI on postgraduate research: Opportunities, challenges, and ethical implications. Research Square. 2025.
39. Biljana M, Thomas J, Flemyng E, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence 2025. JBI Evidence Synthesis. 2025.
40. Pham HH, Nguyen T, Manh-Tung Ho. How ChatGPT undermines my research productivity. Zenodo. 2025.

41. Çelik M. Integrating artificial intelligence into scientific writing: A narrative review for clinical and surgical researchers. American Journal of Surgery. 2025; 250.
42. Quintino MEG, Vorcaro Garcia AGP, Chaves EC, et al. Ethical implications of using artificial intelligence (AI) in scientific texts: An integrative literature review. Anais do ABEC Meeting. 2024.
43. Hryciw BN, Andrew JE Seely, Kyeremanteng K. Guiding principles and proposed classification system for the responsible adoption of artificial intelligence in scientific writing in medicine. Front Artif Intell. 2023; 6: 1283353.
44. Shigapov R, Paunova V. Responsible use of AI in research. Zenodo. 2025.
45. Hanafi M, El-Sayed A, Hassan R. Generative AI in academia: A comprehensive review of applications and implications for the research process. International Journal of Engineering and Applied Sciences. 2025; 12.
46. Salman A, Khan R, Ahmed S. Systematic analysis of generative AI tools integration in academic research and peer review. Online Journal of Communication and Media Technologies. 2025; 15: e202508.
47. Schmidt L, Sinyor M, Webb RT, et al. A narrative review of recent tools and innovations toward automating living systematic reviews and evidence syntheses. Journal of Evidence, Continuing Education, and Quality in Health Care. 2023; 182: 1-10.
48. Hsu HY, Hakouz A, Fotouhi G. Towards responsible generative AI in academia: A synthesis of AI policies on academic writing in the field of educational research. AI and Ethics. 2025; 5: 5467-5484.