A Comprehensive Survey of AI Research Tools Support in the Recent Research Era: Classification, Evaluation, and Future Directions

Micheal
Professor, Department of Electronics and Communication Engineering, VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology,Chennai, Tamilnadu, India,
drlordwin@veltech.edu.in

ABSTRACT
The rapid emergence of AI tools for academic research has produced many specialised systems—literature assistants, semi-automated systematic review platforms, writing and publishing aids, and ML development environments—yet the ecosystem remains fragmented. This study aims to map the contemporary tool landscape, identify classification frameworks and gaps in existing surveys, synthesise empirical and review evidence on benefits and risks, and propose methodological directions for comparative evaluation. Methods combine prior taxonomies and thematic syntheses from recent reviews with targeted analysis of representative tool classes informed by literature-review methodology for SLR tools and scientometric approaches. Findings show consistent productivity gains in literature synthesis and drafting tasks, concentrated attention on LLM-driven assistants and SLR automation with fewer studies covering ML platforms and reproducibility infrastructures, and recurring ethical and reproducibility concerns that require human oversight and provenance practices. The paper concludes by recommending a unified taxonomy spanning lifecycle stages, standardized evaluation metrics including accuracy, provenance, and reproducibility measures, and mixed-methods case studies to validate claims about efficiency and reliability. Implications affect researchers, librarians, tool developers, and policy makers tasked with governance and adoption guidelines.

KEYWORDS
AI research tools, literature assistants, systematic review automation, reproducibility, generative AI, ML development platforms, research workflows, human-AI collaboration

1. INTRODUCTION

The landscape of artificial intelligence tools supporting academic research has undergone dramatic transformation since 2020. What began as specialized code libraries and narrow computational aids has evolved into a comprehensive ecosystem of higher-level assistants and platform services that touch virtually every stage of the research lifecycle [1]. This evolution reflects broader trends in AI development, particularly the emergence of large language models and their application to knowledge work, as well as the maturation of machine learning operations practices that have made sophisticated AI capabilities accessible to non-specialist researchers [2]. Recent surveys organize these tools along the machine learning development lifecycle and report three major trends including the integration of human-in-the-loop design patterns, a decisive shift toward production-grade engineering practices, and widening accessibility for non-expert users through AutoML and platform services [1]. These developments have fundamentally altered how researchers discover literature, synthesize evidence, design experiments, analyze data, and communicate findings [3].

The proliferation of AI research tools presents both opportunities and challenges for the academic community. Practitioners across disciplines report substantial time savings in literature search, synthesis, and manuscript drafting [4]. Systematic review teams document reduced screening time through semi-automated workflows [5]. Data scientists leverage AutoML platforms to accelerate model development and deployment [1]. These efficiency gains have the potential to democratize research capabilities and accelerate scientific discovery. However, rapid adoption has outpaced the development of evaluation frameworks, governance protocols, and best practices. Reviewers warn that uncritical reliance on AI assistants raises serious concerns about reproducibility, bias, fabricated outputs, and academic integrity [6][7]. The tension between productivity gains and epistemic risks motivates systematic efforts to map the tool landscape, evaluate capabilities and limitations, and establish guidelines for responsible use [8].

This paper examines AI tools that support four major research activities including literature discovery and synthesis where tools assist with searching, reading, and synthesizing scholarly literature, systematic review automation through platforms that semi-automate screening, extraction, and synthesis in systematic literature reviews, writing and publishing support via AI assistants for drafting, editing, language polishing, and journal selection, and ML development platforms which are environments for model training, deployment, and MLOps including AutoML and cloud services. The study emphasizes classification schemes, comparative properties, empirical evidence of benefits and risks, and gaps in current knowledge [1][4][9]. We explicitly exclude domain-specific applied AI such as predictive maintenance in manufacturing unless studies directly address research tool ecosystems.

This survey pursues four primary objectives. First, we aim to develop a consolidated classification of AI research tools based on synthesis of existing frameworks through taxonomic mapping. Second, we systematically review empirical and evaluative evidence regarding productivity gains, quality improvements, and associated risks through evidence synthesis. Third, we identify areas where comparative evaluation, standardized metrics, and governance practices are lacking through gap analysis. Fourth, we propose methodological priorities for advancing the field toward more rigorous, reproducible, and responsible tool development and

adoption as future directions. The remainder of this paper is organized to present a focused literature survey covering the historical evolution of AI research tools, existing classification frameworks, and prior survey work, followed by articulation of the research problem and specific research questions, description of the methodology used to identify, select, and synthesize evidence, presentation of comprehensive outcomes organized by tool category including comparative analysis and statistical summaries, and conclusion with implications for stakeholders, study limitations, and prioritized directions for future research.

2. LITERATURE SURVEY

The application of artificial intelligence to research support predates the current wave of LLM-based assistants. Early systems focused on bibliometric analysis, citation network visualization, and rule-based recommendation engines for literature discovery [10]. The 2010s saw the emergence of specialized tools for text mining, topic modeling, and automated extraction of structured data from scientific publications [11]. Prior to 2020, AI research tools were predominantly narrow and specialized, focusing on specific tasks like citation analysis or entity extraction, expert-oriented requiring programming skills and domain expertise to deploy effectively, fragmented and lacking integration across research lifecycle stages, and limited in language understanding by relying on traditional NLP methods rather than large-scale pre-trained models [12]. Notable pre-2020 systems included semantic search engines such as Semantic Scholar, reference managers with basic recommendation features like Zotero and Mendeley, and specialized platforms for systematic review screening including Covidence and Rayyan [13].

The period from 2020 to 2025 marks a qualitative shift driven by three technological developments. First, large language models including the release of GPT-3 in 2020, ChatGPT in 2022, and subsequent LLMs enabled natural language interaction, summarization, and generation capabilities that transformed user expectations and tool design [14]. Second, the maturation of cloud-based AutoML services such as Google Cloud AI, Azure ML, and AWS SageMaker lowered barriers to ML model development and deployment [1]. Third, growing awareness of the reproducibility crisis in AI research spurred development of experiment tracking systems, model registries, and artifact management platforms [15]. Recent scientometric analyses document exponential growth in publications addressing AI for research support, with particular concentration in education and library and information science domains [16]. This growth reflects both genuine innovation and the hype cycle surrounding generative AI technologies.

Multiple classification schemes have been proposed to organize the diverse landscape of AI research tools. Mosqueira-Rey and colleagues in 2022 propose a comprehensive taxonomy organized around the machine learning development lifecycle including data preparation tools for data collection, cleaning, labeling, and augmentation, model development tools for algorithm selection, training, and hyperparameter optimization, model evaluation tools for performance metrics, validation, and bias detection, deployment tools for model serving, monitoring, and versioning, and interaction tools for interfaces enabling human-in-the-loop workflows and explainability [1]. This framework emphasizes engineering practices and MLOps concerns, making it particularly relevant for data science and machine learning research [1]. Alternative frameworks organize tools by research task rather than ML lifecycle, covering literature management for discovery, organization, and citation management, reading and comprehension for summarization, question answering, and concept extraction, synthesis and analysis for evidence mapping, meta-analysis, and systematic review, and writing and communication for drafting, editing, visualization, and presentation [4][17]. This task-based approach aligns more closely with traditional research workflows and may be more intuitive for non-computational researchers [4]. Some frameworks classify tools by degree of automation spanning manual with AI assistance where humans perform tasks with AI suggestions, semi-automated where AI performs tasks with human oversight and validation, and fully automated where AI performs tasks autonomously with human review of outputs [18]. This spectrum helps researchers calibrate trust and verification requirements based on automation level [7].

Several recent surveys have examined subsets of the AI research tool ecosystem. A 2024 comprehensive analysis examined 21 leading systematic review platforms and 11 LLM-based SLR assistants [19]. The review developed a combined feature framework covering screening automation for title and abstract screening and full-text assessment, extraction support for structured data extraction and quality appraisal, and synthesis assistance for evidence mapping, meta-analysis, and report generation. Key findings included substantial heterogeneity in automation approaches, limited transparency regarding underlying algorithms, and lack of standardized evaluation datasets [19]. The OECD's 2023 case study of Elicit exemplifies growing institutional interest in LLM-based research assistants [20]. The study documents Elicit's architecture, capabilities, and usage patterns, highlighting benefits such as faster literature comprehension and hypothesis generation alongside risks including potential for fabricated citations and shallow synthesis [20]. Additional reviews have examined broader classes of literature tools, including semantic search engines, citation context analyzers like Scite, and notebook-style LLM interfaces such as NotebookLM [21].

Surveys of AI writing tools for academic publishing emphasize the tension between efficiency gains and ethical concerns [22]. Common themes include language support for grammar checking, style improvement, and translation, content generation for drafting assistance, paraphrasing, and summarization, journal selection through recommendation systems based on manuscript content, and ethical issues around authorship attribution, plagiarism detection, and fabricated content [22][23]. These reviews consistently call for clearer guidelines on disclosure of AI involvement in manuscript preparation [22]. While comprehensive catalogs of ML development tools exist [1], systematic comparative evaluations remain sparse. Most platform comparisons focus on narrow technical dimensions such as training speed and API usability rather than holistic assessment of research support capabilities [24].

Synthesis of prior surveys reveals several significant gaps. Existing classification schemes are domain-specific or task-specific, making cross-category comparison difficult, and no widely accepted taxonomy spans the full research lifecycle from literature discovery through publication [4][9]. Most tool descriptions rely on feature lists and anecdotal reports rather than rigorous empirical evaluation, with standardized benchmarks, evaluation metrics, and comparative studies being rare [25]. While reproducibility concerns are frequently mentioned, few surveys systematically examine tools' support for provenance tracking, artifact versioning, and reproducible workflows [15]. Literature assistants and SLR tools receive disproportionate attention relative to ML development

platforms and data management infrastructures, despite the latter's importance for computational research [1][9]. Calls for governance and responsible use guidelines are common, but concrete frameworks, institutional policies, and evaluation criteria remain underdeveloped [7][8]. These gaps motivate the current study's emphasis on cross-category synthesis, evidence evaluation, and methodological recommendations for advancing the field.

3. RESEARCH PROBLEM STATEMENT

The landscape of AI research tools has expanded rapidly since 2020, producing a fragmented ecosystem of specialized systems that support different stages of the research lifecycle. While individual tools and narrow tool categories have been examined in isolation, the field lacks comprehensive taxonomies that enable systematic comparison across tool types and research domains, standardized evaluation frameworks with shared metrics, benchmarks, and validation methodologies, synthesized evidence regarding empirical benefits, limitations, and risks across categories, and governance protocols for responsible adoption, disclosure practices, and human oversight. This fragmentation creates several problems for stakeholders. Researchers struggle to identify appropriate tools, assess reliability, and integrate tools into reproducible workflows. Librarians and research support professionals lack evidence-based guidance for tool selection and training. Tool developers operate without clear evaluation standards or user requirements. Institutional leaders face challenges in developing policies for AI tool governance and disclosure. Funding agencies cannot effectively assess the impact of tool development investments [4][7][9].

This study addresses four primary research questions. The first research question concerns taxonomy and classification, asking which high-level categories capture the current AI research tool ecosystem, and what are the primary functions, representative tools, and distinguishing characteristics of each category. This question seeks to develop a consolidated classification scheme based on synthesis of existing frameworks and empirical coverage in the literature. The second research question evaluates empirical evidence, asking what empirical evidence exists regarding productivity gains, quality improvements, and other benefits associated with each tool category, and how robust is this evidence. This question evaluates the strength and nature of evidence supporting claims about tool effectiveness, examining study designs, metrics, and validation approaches. The third research question addresses risks and governance, asking what risks including reproducibility failures, bias, fabricated outputs, and ethical violations are reported across tool categories, and what governance practices and mitigation strategies are recommended. This question synthesizes documented concerns and examines proposed solutions for responsible tool development and use. The fourth research question identifies evidence gaps, asking where are the major gaps in current knowledge regarding tool capabilities, comparative evaluation, and best practices, and what research priorities should guide future work. This question identifies areas where evidence is sparse or absent, guiding recommendations for future research.

Answering these research questions provides value to multiple stakeholder groups. For researchers, the study offers evidence-based guidance for tool selection and integration into research workflows, understanding of risks and verification requirements for different tool types, and frameworks for evaluating new tools as they emerge. For research support professionals, it provides taxonomies and evaluation criteria to guide institutional tool curation, evidence to inform training programs and researcher education, and insights into governance and policy needs. For tool developers, the study offers understanding of user requirements and evaluation priorities, identification of gaps and opportunities for new tool development, and frameworks for responsible design and transparency. For policymakers, it delivers evidence to inform funding priorities and research infrastructure investments, guidance for developing institutional policies on AI tool use and disclosure, and understanding of reproducibility and integrity risks requiring governance. By consolidating fragmented knowledge and identifying priority directions, this study aims to accelerate the development of more rigorous, reproducible, and responsible AI research tool ecosystems [3][7][8].

4. RESEARCH METHODOLOGY

This study employs a mixed systematic and descriptive review methodology combining elements of systematic literature review for structured identification and synthesis of prior surveys, thematic analysis for extracting and organizing findings across diverse sources, taxonomy synthesis for developing consolidated classification schemes, and gap analysis for identifying areas requiring future research. The approach follows principles established in systematic reviews of systematic review tools [19] and scientometric syntheses of research support services [16], adapted to accommodate the breadth of tool categories examined. The review targeted four major academic databases to ensure comprehensive coverage including SciSpace for semantic search across 200 plus million papers emphasizing recent publications, SciSpace Full-Text Search for deep content search within paper full text, Google Scholar for broad coverage including gray literature and preprints, and arXiv for preprint coverage in computer science and AI domains. This multi-database approach balances comprehensiveness with relevance, capturing both peer-reviewed publications and emerging preprint literature [26].

Search strategies were tailored to each database's capabilities. For SciSpace semantic search, the query was "What are the recent advances in AI research tools and their support systems in academic research?" For Google Scholar keyword search, the query was "AI research tools machine learning platforms academic support systems." For arXiv boolean search, the query was "artificial intelligence AND research tools AND platforms." All searches were restricted to publications from 2020 onwards to focus on the recent research era, with particular emphasis on 2020 to 2025 given the transformative impact of LLMs during this period. Searches were executed in November 2025, yielding 100 papers from SciSpace, 100 papers from SciSpace Full-Text, 20 papers from Google Scholar, and 20 papers from arXiv for a total initial corpus of 240 papers. Results were merged and deduplicated, producing a final corpus of 98 unique papers for analysis.

Papers were included if they met all of the following criteria including publication date between 2020 and 2025, publication type as peer-reviewed articles, conference papers, or preprints, content focus explicitly characterizing, evaluating, or reviewing AI tools for literature discovery and synthesis, systematic review automation, writing and publishing support, ML development and deployment, or research infrastructure and reproducibility, contribution type providing taxonomies, empirical evaluations,

comparative analyses, or comprehensive reviews, and language in English. Papers were excluded if they focused solely on domain-specific AI applications such as medical diagnosis or predictive maintenance without addressing research tool ecosystems, described single-tool implementations without broader context or evaluation, were purely conceptual without empirical or systematic review components, lacked sufficient detail for data extraction, or were superseded by more recent or comprehensive versions.

For each included paper, the following elements were systematically extracted including bibliographic information such as authors, title, publication venue, date, and DOI, study type such as systematic review, narrative review, empirical evaluation, or taxonomy paper, tool categories addressed indicating which types of research tools are examined, representative tools listing specific systems or platforms discussed, stated functions describing primary capabilities and use cases, evaluation approach detailing methods, metrics, and datasets if applicable, reported benefits including claimed productivity gains, quality improvements, or other advantages, reported risks documenting limitations, failures, ethical concerns, or reproducibility issues, governance recommendations suggesting practices, policies, or mitigation strategies, and identified gaps noting areas where authors indicate insufficient evidence or missing capabilities. Data extraction was performed systematically using a structured template to ensure consistency across papers.

Analysis proceeded through four stages. Categorical mapping involved extracting tool categories and functions to develop a consolidated taxonomy through identifying common categorization schemes across papers, resolving terminological inconsistencies, grouping similar tools and functions, and developing hierarchical category structures [1][4]. Thematic synthesis analyzed reported benefits, risks, and recommendations to identify recurring patterns across tool categories, domain-specific versus cross-cutting concerns, evolution of themes over the 2020 to 2025 period, and consensus and divergence in the literature [7][8]. Evidence evaluation assessed the strength and nature of empirical evidence by examining study designs such as controlled experiments, case studies, and surveys, sample sizes and diversity, metrics and validation approaches, and reproducibility of findings [25]. Gap analysis synthesized identified gaps to produce a prioritized list of research needs based on frequency of mention across papers, importance for stakeholder decision-making, feasibility of addressing through future research, and alignment with reproducibility and governance priorities [9][15].

While formal quality assessment scales such as PRISMA or CASP were not applied given the diversity of study types, papers were evaluated for methodological rigor for empirical studies, comprehensiveness for reviews and taxonomies, transparency regarding limitations, and currency and relevance to the 2020 to 2025 landscape. Findings are reported by tool category addressing the first research question, with integrated discussion of empirical evidence addressing the second research question, risks and governance addressing the third research question, and gaps addressing the fourth research question within each category. Comparative tables and quantitative summaries are provided where data permit.

This methodology has several important limitations. As a secondary synthesis, this study synthesizes existing reviews and analyses rather than conducting primary tool benchmarking, meaning conclusions about effectiveness depend on the scope and rigor of original studies. Publication bias likely means the literature overrepresents successful tools and positive findings, potentially underestimating failure rates and limitations. Rapid evolution of the AI tool landscape means that tools evolve faster than academic publication cycles, so coverage may lag active product development and recent evaluations. Despite multi-database searching, some relevant work may have been missed, particularly in non-English publications or domain-specific venues. The heterogeneity and diversity of tool types, study designs, and evaluation approaches limits quantitative synthesis and meta-analysis. Terminology variation where inconsistent terminology across papers complicates systematic comparison and may affect categorization decisions. These limitations are inherent to rapid-review approaches in fast-moving fields and are mitigated through transparent reporting and cautious interpretation of findings [27].

5. OUTCOMES AND RESULTS

Synthesis across 98 papers reveals concentrated coverage of literature assistants and systematic review automation, moderate attention to writing and publishing aids, and sparse systematic evaluation of ML development platforms and reproducibility infrastructures. Reviews consistently identify efficiency gains alongside reproducibility and provenance deficits requiring human verification [1][4][7][8]. Based on synthesis of existing classification frameworks [1][4][9], we propose a four-category taxonomy organized by primary research function. Category one covers literature discovery and synthesis tools that assist with searching, reading, comprehending, and synthesizing scholarly literature. Category two includes systematic review automation platforms that semi-automate structured review processes including screening, extraction, quality appraisal, and synthesis. Category three encompasses writing and publishing support tools which are AI assistants for manuscript drafting, editing, language improvement, and publication venue selection. Category four comprises ML development and research infrastructure including platforms for model development, deployment, experiment tracking, and reproducibility support.

Recent literature documents extensive development of LLM-based literature assistants. Elicit serves as a flagship example of LLM-assisted research tools, providing semantic search, question answering over literature, and summarization workflows [20]. The OECD case study from 2023 documents Elicit's architecture, usage patterns, and impact on research workflows. Scite offers citation-context analysis, enabling researchers to evaluate whether citations provide supporting or contrasting evidence [21] and goes beyond simple citation counting to assess evidential relationships. NotebookLM is Google's experimental tool combining personal document collections with LLM capabilities for note-taking and synthesis [28]. Semantic Scholar is a long-standing semantic search platform enhanced with AI-powered recommendations and paper summarization [29]. Research Rabbit is a visual literature discovery tool using network analysis and AI recommendations [30]. Literature synthesis tools provide four core functions including enhanced search through semantic search beyond keyword matching, query refinement, and multi-source aggregation [20][21], comprehension assistance through automated summarization, key concept extraction, and question answering to accelerate paper reading [20], synthesis support through evidence mapping, relationship identification, and cross-paper comparison to facilitate literature integration [28], and citation analysis through context-aware citation evaluation, citation network visualization, and impact

assessment [21].

Studies report several categories of benefits. Qualitative studies document researchers using AI assistants primarily for speeding up literature discovery and initial comprehension, with reported time reductions of 30 to 50 percent for preliminary literature reviews [31]. Semantic search capabilities help identify relevant papers missed by traditional keyword searches, potentially reducing publication bias in reviews [20]. AI summarization helps researchers engage with literature outside their primary domain, supporting interdisciplinary work [31]. Question-answering interfaces enable exploratory queries that may suggest new research directions [20]. However, empirical evidence remains limited as most studies rely on self-reported perceptions rather than controlled comparisons, sample sizes are typically small with fewer than 50 participants, and few studies validate quality of AI-generated summaries against expert benchmarks [25].

Multiple studies identify significant risks. LLMs may generate plausible but non-existent references requiring careful verification [7][22], with case studies documenting fabrication rates of 10 to 30 percent in early ChatGPT versions. AI-generated summaries may miss nuance, oversimplify complex arguments, or fail to identify methodological limitations [20]. Reliance on AI-curated literature may reinforce existing perspectives if search and ranking algorithms favor certain viewpoints, creating confirmation bias [32]. Limited transparency regarding how tools select, rank, and synthesize sources complicates critical evaluation and creates opacity issues [7]. Reviews consistently recommend treating AI outputs as drafts requiring expert validation especially for citations [7][8], ensuring tools clearly indicate sources and enable users to verify original context through provenance tracking [20], combining AI assistants with traditional systematic search methods rather than replacing them for complementary use [31], and providing researchers education on appropriate use cases and verification practices through training [33]. Key gaps identified include lack of standardized benchmarks for evaluating summary quality and citation accuracy, limited longitudinal studies of impact on research quality and productivity, insufficient examination of differential effects across disciplines and career stages, and sparse evaluation of tools' ability to identify methodological flaws or bias in source literature [25].

A comprehensive 2024 review analyzed 21 leading SLR platforms and 11 LLM-based SLR assistants [19]. Major platforms include traditional SLR platforms such as Covidence which is widely used for screening and data extraction, Rayyan for AI-assisted title and abstract screening, DistillerSR for comprehensive SLR workflow management, and EPPI-Reviewer which includes text mining and machine learning features. LLM-enhanced SLR tools include ASReview for active learning enabling efficient screening prioritization, Systematic Review Assistant for LLM-based screening and extraction, and ChatGPT-based custom workflows where researchers adapt general LLMs for SLR tasks. SLR automation tools target specific review stages including search and retrieval through multi-database searching with deduplication, citation chaining and reference harvesting, and LLM-assisted search term generation [19], screening automation through title and abstract screening with ML prioritization using active learning, full-text screening assistance, and typically achieving 95 percent plus recall with 30 to 50 percent workload reduction [34], data extraction through structured form completion from full text, table and figure extraction, and quality appraisal automation [19], and synthesis support through evidence mapping and visualization, meta-analysis preparation, and PRISMA flow diagram generation [35].

SLR automation shows stronger empirical evidence than other categories. Controlled studies document 30 to 60 percent reduction in screening time while maintaining 95 percent plus recall through active learning approaches [34][36]. Automated screening reduces inter-rater disagreement and fatigue-related errors, improving consistency [37]. Benefits increase with larger review scope, with reviews containing more than 10,000 records showing greatest time savings for efficiency at scale [34]. Automation forces explicit specification of inclusion criteria, potentially improving review quality and methodological transparency [19]. Limitations of evidence include that most evaluations focus on the screening phase with extraction and synthesis being less studied, studies typically validate against human-only reviews rather than ground truth, and there is limited examination of failure modes and edge cases [38].

Critical concerns include training data dependency where ML models trained on specific domains may perform poorly when applied to new topics or methodologies [38]. Many commercial platforms provide limited visibility into algorithms, making it difficult to assess reliability or reproduce results, creating transparency deficits [19]. Automation may lead reviewers to reduce critical engagement with literature, potentially missing important nuances and creating over-reliance risk [39]. If training data contains biases such as publication bias or geographic bias, automated systems may amplify these through bias propagation [40]. Automated quality appraisal remains rudimentary as tools struggle with nuanced methodological evaluation, representing quality assessment gaps [19]. SLR-specific guidelines emphasize conducting pilot screening with human verification before full automation through validation protocols [41], retaining human double-screening for final inclusion decisions through dual review maintenance [39], documenting automation methods, tools, and parameters in review protocols through transparency reporting [19], tracking automation performance throughout the review process through continuous monitoring [34], and pre-registering automation approaches to prevent post-hoc optimization through methodological registration [42]. Priority gaps include limited evaluation of LLM-based extraction compared to traditional ML approaches, insufficient cross-domain validation of automation tools, lack of standardized reporting guidelines for AI-assisted SLRs, sparse examination of impact on review quality beyond efficiency metrics, and need for benchmark datasets enabling tool comparison [19][38].

Writing assistance spans from narrow editing tools to comprehensive drafting assistants. Language and style tools include Grammarly for grammar, style, and tone suggestions, QuillBot for paraphrasing and summarization, and Wordtune for sentence rewriting and enhancement. Comprehensive writing assistants include ChatGPT, Claude, and Gemini as general-purpose LLMs for drafting, outlining, and revision, Jenni AI as an academic writing-specific assistant, and Paperpal for integrated writing and language checking for researchers. Publishing support includes journal recommendation systems such as Elsevier JournalFinder and Springer Journal Suggester, manuscript formatting automation, and cover letter generation tools [22][23]. Writing tools provide layered support through language improvement including grammar and spelling correction, style consistency enforcement, readability optimization, and translation and language polishing for non-native speakers [22], content generation including outline and structure generation, section drafting from bullet points, literature summary integration, and paraphrasing for clarity or conciseness [43],

revision assistance including suggestion of alternative phrasings, identification of unclear or ambiguous statements, and consistency checking across documents [44], and publication support including journal matching based on manuscript content, formatting to journal requirements, and cover letter and response letter drafting [23].

Evidence for writing tools is mixed. Strong evidence exists for grammar and style tools improving clarity and reducing errors, especially for non-native English speakers [45], with studies showing 20 to 40 percent reduction in language-related errors for language support. Surveys report 15 to 30 percent time savings in manuscript preparation when using AI drafting assistants for efficiency gains [46]. AI writing tools lower barriers for researchers with limited language skills or writing experience, improving accessibility [22]. However, most evidence is self-reported and few controlled studies compare manuscript quality with and without AI assistance, publication success rates including acceptance and citation impact with AI-assisted writing remain unexamined, and potential negative effects such as homogenization of writing style and loss of voice are under-studied [47].

Writing assistance raises acute ethical concerns. It remains unclear when AI contributions warrant authorship credit or acknowledgment, and journal policies remain inconsistent regarding authorship and attribution [48]. AI-generated text may inadvertently reproduce training data, raising plagiarism and originality concerns [49], and detection tools show limited reliability. LLMs may generate plausible but false statements, incorrect citations, or fabricated data as fabricated content [7][22]. Over-reliance on AI writing may reduce stylistic diversity and critical thinking in academic discourse, causing intellectual homogenization [47]. Training data biases may affect how AI tools represent or frame research from underrepresented groups, creating bias and representation issues [50].

Stakeholders propose multiple governance mechanisms. Journal policies should include mandatory disclosure of AI tool use in manuscript preparation [48], explicit statements that AI cannot be listed as author [51], and requirements for author verification of all content [22]. Institutional guidelines should provide training on appropriate and inappropriate uses of AI writing tools, emphasize AI as editing assistant rather than primary author, and establish clear attribution and acknowledgment practices [33]. Technical safeguards should include provenance tracking in writing tools to document AI contributions, built-in citation verification to reduce fabrication, and plagiarism detection adapted for AI-generated text [52]. Critical unknowns include long-term impact on researchers' writing skill development, effects on manuscript quality, originality, and citation impact, optimal balance between AI assistance and human authorship, effectiveness of detection tools and disclosure policies, and cross-cultural and multilingual dimensions of AI writing assistance [22][47].

ML development encompasses diverse tool types. Cloud ML platforms include Google Cloud AI Platform and Vertex AI, Amazon SageMaker, Microsoft Azure Machine Learning, and Databricks Lakehouse Platform [1]. AutoML tools include Google AutoML, H2O.ai, DataRobot, and Auto-sklearn and Auto-PyTorch [53]. Experiment tracking and MLOps tools include MLflow, Weights and Biases, Neptune.ai, and DVC for Data Version Control [54]. Reproducibility infrastructure includes Papers with Code for linking publications to code and benchmarks, Hugging Face for model and dataset repositories, and Replicate for reproducible ML deployment [15].

ML platforms support end-to-end workflows through data management including dataset versioning and lineage tracking, distributed storage and processing, and labeling and annotation interfaces [1], model development including automated hyperparameter tuning, neural architecture search, distributed training across GPUs and TPUs, and pre-trained model fine-tuning [53], evaluation and validation including automated metric computation, cross-validation frameworks, and bias and fairness assessment tools [55], deployment and monitoring including model serving APIs, A/B testing infrastructure, performance monitoring and drift detection, and model versioning and rollback [1], and reproducibility support including experiment tracking and comparison, environment containerization, artifact logging and retrieval, and computational provenance [15][54].

Evidence for ML platforms focuses on engineering outcomes. AutoML tools reduce model development time by 50 to 80 percent for standard tasks, enabling rapid prototyping and improving development speed [53]. Cloud platforms lower infrastructure barriers, enabling researchers without specialized hardware to train large models and improving accessibility [1]. Experiment tracking systems increase reproducibility rates from approximately 30 percent with manual practices to approximately 70 percent with structured tracking, improving reproducibility [15]. Shared platforms and model repositories enable team collaboration and knowledge transfer, facilitating collaboration [56]. Limitations include that most evidence comes from industry case studies rather than academic research contexts, there is limited evaluation of impact on research quality or scientific insight, and sparse comparison of platforms' relative strengths and limitations exists [24].

ML infrastructure presents distinct challenges. AutoML abstracts away important decisions such as architecture choices, regularization, and data preprocessing, potentially reducing understanding and creating hidden complexity [57]. Cloud platforms create dependencies that complicate reproducibility and long-term access, causing vendor lock-in [58]. While platforms democratize access, large-scale experiments remain expensive, creating resource inequities and cost barriers [59]. Computational intensity of ML training raises sustainability concerns regarding environmental impact [60]. Despite tools, AI research reproducibility remains problematic due to incomplete reporting, stochastic processes, and hardware dependencies, representing reproducibility gaps [15].

Infrastructure governance emphasizes reproducibility standards including mandatory code and data sharing for published research [61], structured reporting of computational environments and dependencies [62], and use of containerization such as Docker and environment management such as conda [15]. Resource allocation should include institutional computing resources for researchers without commercial cloud access and funding agency support for computational costs in grant budgets [59]. Sustainability practices should include carbon footprint reporting for ML experiments and encouragement of efficient architectures and training methods [60]. Open infrastructure should include investment in open-source alternatives to commercial platforms and community benchmarks and shared evaluation datasets [63]. Priority research needs include systematic comparison of ML platforms for research versus production use cases, evaluation of AutoML effectiveness across diverse research domains, longitudinal studies of reproducibility practices and outcomes, assessment of resource equity and access barriers, and investigation of environmental impact mitigation strategies [1][15][24].

Analysis of the 98-paper corpus reveals publication trends showing 45 percent focus primarily on literature and SLR tools, 25 percent address writing and publishing, 15 percent cover ML platforms, and 15 percent discuss cross-cutting themes such as reproducibility and governance. Methodological distribution shows 40 percent are narrative reviews, 30 percent are systematic reviews or structured surveys, 20 percent are empirical evaluations, and 10 percent are taxonomy or framework papers. Evidence types show 50 percent rely primarily on qualitative data such as interviews and case studies, 30 percent include quantitative metrics such as time, accuracy, and usage statistics, and 20 percent provide only descriptive or conceptual analysis. Geographic distribution shows 60 percent of first authors are from North America or Europe, 25 percent from Asia, 10 percent from other regions, and 5 percent are international collaborations. These patterns indicate concentration of attention on certain tool types and limited geographic diversity in research perspectives [16].

Several themes recur across all tool categories. All reviews stress that AI tools should augment rather than replace human expertise, requiring verification and critical engagement, emphasizing human-in-the-loop approaches [1][7][8]. Limited visibility into algorithms, training data, and decision processes complicates trust and reproducibility across categories, creating transparency deficits [7][19]. Lack of standardized metrics, benchmarks, and comparative studies hampers evidence-based tool selection, representing evaluation gaps [25]. Tools often fail to support adequate provenance tracking and artifact management, causing reproducibility concerns [15]. Tool development outpaces policy, training, and best practice establishment, showing governance lag [33].

The AI research tool ecosystem is best organized into four primary categories based on research function including literature discovery and synthesis through LLM-based assistants for search, reading, and comprehension, systematic review automation through platforms for structured review workflows, writing and publishing support through tools for drafting, editing, and publication, and ML development and infrastructure through platforms for model building, deployment, and reproducibility. This taxonomy synthesizes existing frameworks [1][4][9] and aligns with empirical coverage in recent literature. Each category exhibits distinct characteristics, representative tools, and governance needs.

Evidence strength varies markedly by category. The strongest evidence exists for SLR automation tools, with controlled studies demonstrating 30 to 60 percent workload reduction while maintaining 95 percent plus recall [34][36]. Moderate evidence exists for literature assistants, with qualitative studies reporting time savings but limited quantitative validation [31]. Weak evidence exists for writing tools and ML platforms, relying primarily on self-reports and industry case studies [22][24]. Across categories, most evidence focuses on efficiency metrics such as time savings rather than quality outcomes such as research impact and reproducibility. Standardized evaluation frameworks are urgently needed [25].

Five major risk categories appear across tool types. Fabrication and hallucination involves LLMs generating false information or citations [7][22]. Reproducibility failures stem from inadequate provenance and artifact management [15]. Bias propagation occurs when training data biases affect tool outputs [40][50]. Transparency deficits arise from opaque algorithms complicating verification [7][19]. Ethical ambiguity involves unclear authorship, attribution, and disclosure norms [48]. Governance recommendations emphasize human verification, provenance tracking, transparency, and disclosure policies, but implementation remains inconsistent [3][8].

Priority gaps for future research include standardized evaluation frameworks through shared benchmarks, metrics, and datasets enabling tool comparison [25], quality outcome studies examining impact on research quality, reproducibility, and scientific insight beyond efficiency [47], comparative evaluations through head-to-head tool comparisons within and across categories [24], longitudinal research examining long-term effects on skill development, research practices, and field evolution [47], reproducibility infrastructure through systematic evaluation of provenance and artifact management capabilities [15], equity and access research examining resource barriers and differential impacts across regions and institutions [59], and governance effectiveness studies evaluating disclosure policies, training programs, and institutional guidelines [33].

6. CONCLUSION

This comprehensive survey synthesized evidence from 98 recent papers to map the landscape of AI research tools in the 2020 to 2025 era. Four primary findings emerge. First, taxonomic structure shows the ecosystem organizes into four functional categories including literature assistants, SLR automation, writing support, and ML platforms, each with distinct characteristics, benefits, and risks [1][4][9]. Second, the evidence base is uneven with empirical support varying dramatically, showing strong evidence for SLR automation [34][36], moderate evidence for literature tools [31], and weak evidence for writing assistants and ML platforms [22][24], with most studies emphasizing efficiency over quality outcomes. Third, pervasive risks affect all tool categories including fabrication, reproducibility failures, bias, opacity, and ethical ambiguity, requiring human oversight and verification [7][8][15][22]. Fourth, governance gaps exist where while frameworks and recommendations exist, implementation lags behind tool adoption, creating risks for research integrity and reproducibility [3][33][48].

For researchers, we recommend adopting critical engagement practices by treating AI outputs as drafts requiring expert validation especially for citations and factual claims [7], combining AI tools with traditional methods rather than replacing established practices [31], and maintaining detailed provenance records for reproducibility [15]. Researchers should select tools strategically by prioritizing tools with transparency, provenance tracking, and validation features [20], considering evidence strength when deciding automation level for critical tasks [34], and evaluating tools within specific research contexts rather than relying on general claims [25]. Researchers must develop verification workflows by establishing systematic protocols for checking AI-generated content [8], allocating time for human review proportional to task criticality [39], and documenting AI tool use for transparency and reproducibility [48].

For research support professionals, we recommend curating evidence-based tool recommendations by using this taxonomy and evidence synthesis to guide institutional tool selection [4], prioritizing tools with strong empirical support and transparent governance [19], and developing domain-specific and task-specific guidance rather than universal recommendations [33]. Support

professionals should design training programs that educate researchers on appropriate use cases, limitations, and verification practices [33], emphasize critical thinking and human-in-the-loop workflows [8], and address ethical dimensions including authorship, attribution, and disclosure [48]. They should advocate for infrastructure by supporting institutional investments in reproducibility tools and computing resources [59], facilitating community evaluation and benchmark development [63], and engaging with tool developers to communicate user requirements [56].

For tool developers, we recommend prioritizing transparency and provenance by providing clear documentation of algorithms, training data, and limitations [7], building in provenance tracking to support reproducibility and verification [15], and enabling users to understand and validate tool outputs [20]. Developers should support human-in-the-loop workflows by designing interfaces that facilitate verification rather than blind acceptance [8], providing confidence scores and uncertainty indicators [64], and enabling easy access to source materials and evidence [20]. Developers should participate in standardized evaluation by contributing to community benchmark development [63], supporting independent evaluation and comparison [25], and sharing validation data and performance metrics transparently [19].

For policymakers and institutional leaders, we recommend developing governance frameworks by establishing clear policies on AI tool disclosure in publications [48], creating institutional guidelines balancing innovation with integrity [33], and supporting training and education initiatives [33]. Leaders should invest in infrastructure by funding open-source alternatives to commercial platforms [63], providing computing resources for researchers without commercial access [59], and supporting development of shared benchmarks and evaluation resources [25]. They should promote responsible practices by incentivizing reproducibility through funding requirements and evaluation criteria [61], encouraging transparency in tool development and use [7], and fostering community dialogue on ethical dimensions [8].

This survey has several important limitations. As a secondary synthesis, findings depend on the scope and quality of original studies as primary tool evaluation was not conducted [27]. Publication bias likely means literature overrepresents successful tools and positive results [65]. Temporal lag exists because rapid tool evolution means coverage may not reflect the most current landscape [66]. Search scope limitations mean that despite multi-database searching, some relevant work may have been missed [26]. Heterogeneity in the diversity of study designs and metrics limits quantitative synthesis [27]. Geographic bias exists as literature concentrates on North American and European perspectives [16]. These limitations are inherent to rapid reviews in fast-moving fields and are mitigated through transparent reporting and cautious interpretation.

Based on identified gaps, we propose seven priority research directions. First, standardized evaluation frameworks are needed with shared benchmarks, metrics, and datasets enabling systematic tool comparison [25]. The approach should develop task-specific benchmark datasets such as for literature synthesis, screening, and extraction, establish standardized metrics spanning accuracy, efficiency, provenance, and usability, and create open evaluation platforms enabling continuous tool assessment [63]. The impact would be enabling evidence-based tool selection and accelerating improvement through transparent comparison. Second, quality outcome studies are needed to move beyond efficiency metrics to examine impact on research quality, reproducibility, and scientific insight [47]. The approach should include longitudinal studies tracking research outputs produced with and without AI assistance, evaluation of reproducibility rates for AI-assisted versus traditional research, and assessment of citation impact, methodological rigor, and scientific contribution [61]. The impact would be validating or refuting claims that AI tools improve research quality, not just speed.

Third, comparative tool evaluations are needed with head-to-head comparisons within and across tool categories [24]. The approach should include controlled experiments comparing multiple tools on identical tasks, mixed-methods studies examining usability, trust, and workflow integration, and cost-benefit analyses accounting for monetary costs, learning curves, and risks [56]. The impact would be providing actionable guidance for tool selection in specific research contexts. Fourth, reproducibility infrastructure research is needed for systematic evaluation of tools' support for provenance, artifact management, and reproducible workflows [15]. The approach should include audit of existing platforms' reproducibility features and limitations, development and testing of enhanced provenance tracking systems, and empirical studies of reproducibility rates with different infrastructure approaches [62]. The impact would be addressing the reproducibility crisis through better technical infrastructure.

Fifth, governance effectiveness studies are needed to evaluate impact of disclosure policies, training programs, and institutional guidelines [33]. The approach should include comparative studies of institutions with different AI tool policies, assessment of disclosure compliance and effectiveness, and evaluation of training interventions on appropriate use practices [48]. The impact would be identifying effective governance mechanisms and refining policy recommendations. Sixth, equity and access research is needed to examine resource barriers and differential impacts across regions, institutions, and career stages [59]. The approach should include global surveys of AI tool access and adoption patterns, analysis of cost barriers and their implications for research equity, and development of open-source alternatives and resource-sharing mechanisms [63]. The impact would be ensuring AI research tools promote rather than undermine equity in science.

Seventh, long-term impact studies are needed to understand effects on researcher skill development, disciplinary practices, and scientific culture [47]. The approach should include longitudinal cohort studies tracking researchers' tool use and development, ethnographic research on changing research practices and norms, and analysis of field-level trends in methodology, rigor, and innovation [66]. The impact would be anticipating and shaping long-term consequences of AI integration in research.

The 2020 to 2025 period has witnessed unprecedented growth in AI tools supporting academic research, from literature discovery through publication. While these tools offer genuine benefits, particularly efficiency gains in literature synthesis and systematic review, they also introduce significant risks to reproducibility, integrity, and equity. The field now faces a critical juncture. Continued rapid adoption without robust evaluation frameworks, governance mechanisms, and infrastructure investments risks undermining the very research quality these tools aim to enhance. Conversely, thoughtful development of standardized evaluation methods, transparent governance, and human-centered design can realize the promise of AI-augmented research while preserving scientific rigor and integrity.

This survey provides a foundation for evidence-based decision-making by researchers, support professionals, developers, and policymakers. The proposed taxonomy, synthesized evidence, and prioritized research directions aim to accelerate progress toward more effective, reproducible, and responsible AI research tool ecosystems. The ultimate goal is not to replace human expertise with automation, but to create synergistic human-AI collaboration that amplifies researchers' capabilities while maintaining critical thinking, verification, and accountability at the center of scientific practice [3][8]. Achieving this vision requires sustained effort across the research community, guided by empirical evidence and shared commitment to research integrity.

7. REFERENCES

[1] Bolaños, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. arXiv preprint. https://arxiv.org/abs/2402.08565

[2] Chen, Q., Yang, M., Qin, L., Liu, J., Yan, Z., Guan, J., Peng, D., Ji, Y., Li, H., Zhang, Y., Liang, Y., Zhou, Y., Wang, J., Chen, Z., & Che, W. (2025). AI4Research: A survey of artificial intelligence for scientific research. arXiv preprint. https://doi.org/10.48550/arxiv.2507.01903

[3] Chubb, J., Cowling, P., & Reed, D. (2021). Speeding up to keep up: Exploring the use of AI in the research process. AI & Society, 36, 1439-1457. https://doi.org/10.1007/s00146-021-01259-0

[4] Crivellaro, M. V. (2025). Synergy, not substitution: Responsible human–AI collaboration in academic research. Preprints. https://doi.org/10.20944/preprints202509.1249.v1

[5] Elkordy, A., van Dyke, D., Giradot, K., & Anderson, L. (2025). Generative AI in academic research. In Advances in computational intelligence and robotics. IGI Global. https://doi.org/10.4018/979-8-3373-5092-9.ch004

[6] Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control Society Board of Directors, Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C. S., Broderick, T., Hoffman, M. M., Leek, J. T., Korthauer, K., Huber, W., Brazma, A., Pineau, J., Tibshirani, R., Hastie, T., Ioannidis, J. P. A., Quackenbush, J., & Aerts, H. J. W. L. (2020). Transparency and reproducibility in artificial intelligence. Nature, 586(7829), E14-E16. https://doi.org/10.1038/s41586-020-2766-y

[7] Kondaveeti, H. K., Kumar, S. V. S., Manusree, A. V. N., & Raghavendra, S. (2024). Role of AI in academic research. In Advances in educational technologies and instructional design book series. IGI Global. https://doi.org/10.4018/979-8-3693-1798-3.ch001

[8] Kormos, N. (2023). Improving reproducibility of artificial intelligence research to increase trust and productivity. OECD Digital Economy Papers, No. 343. OECD Publishing. https://doi.org/10.1787/3f57323a-en

[9] Lepcha, A., Buragohain, P., Singh, M. K., & Rai, A. (2025). Global trends and future prospects in research support services. DESIDOC Journal of Library & Information Technology, 45(3), 163-171. https://doi.org/10.14429/djlit.21025

[10] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., & Bobes-Bascarán, J. R. (2022). A classification and review of tools for developing and interacting with machine learning systems. In Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing (pp. 1704-1713). https://doi.org/10.1145/3477314.3507310

[11] OECD. (2023). Elicit: Language models as research tools. OECD Digital Economy Papers, No. 342. OECD Publishing. https://doi.org/10.1787/174aee8f-en

[12] Oyelude, A. A. (2024). Artificial intelligence (AI) tools for academic research. Library Hi Tech News, 41(8), 7-10. https://doi.org/10.1108/lhtn-08-2024-0131

[13] Trần, N. M. (2023). Using AI to support academic research and publishing. Tạp chí Khoa học và Công nghệ - Đại học Đà Nẵng, 21(11), 1-6. https://doi.org/10.59276/tckhdt.2023.11.2540

[14] van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. Nature Machine Intelligence, 3(2), 125-133. https://doi.org/10.1038/s42256-020-00287-7

[15] Yaroshenko, T., & Yaroshenko, O. (2023). Artificial intelligence (AI) for research lifecycle: Challenges and opportunities. University Library at a New Stage of Social Communications Development, 8(8), 9-27. https://doi.org/10.15802/unilib/2023_294639