Automated Survey Generation, Literature Review Automation, and Intelligent Agentic Systems in Academia: Foundations, Architectures, and Responsible Integration

Abstract

The exponential growth and interdisciplinary complexity of scientific output have surpassed the capacity of traditional scholarly methods, motivating the widespread adoption of artificial intelligence (AI) and agent-based architectures for academic automation. This comprehensive survey examines the technical, methodological, and ethical foundations underpinning automated survey generation, literature review, and knowledge recognition, focusing on the transformative potential and challenges posed by intelligent agentic systems. The review synthesizes advances in large language models (LLMs), hybrid and multi-agent workflows, and automated question generation, while highlighting innovations in quality assurance, explainability, and cross-lingual inclusivity. Key contributions include a taxonomy of generative artificial experts (GAEs), analyses of composable survey architectures, and frameworks for transparent, human-in-the-loop oversight. Empirical case studies—spanning WhatsApp-based survey automation, agent-driven video and behavioral recognition, and peer-augmented writing assistance—demonstrate gains in scalability, equity, and operational efficiency, yet underscore persistent gaps in standardization, interoperability, and evaluation robustness. The survey addresses pressing issues of academic integrity, bias mitigation, and privacy in AI-assisted research, advocating for harmonized protocols, open benchmarking, and participatory governance. It concludes by outlining best practices for responsible system integration and charting future directions that prioritize contestability, reproducibility, and inclusive design to ensure that automation augments, rather than supplants, core academic values.

ACM Reference Format:

. 2025. Automated Survey Generation, Literature Review Automation, and Intelligent Agentic Systems in Academia: Foundations, Architectures, and Responsible Integration. In . ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction and Motivation

1.1 Scope and Motivation

The increasing complexity and volume of contemporary scientific output have rendered traditional scholarly methods insufficient, creating an urgent demand for automation in scientific discovery, scholarly writing, survey generation, and literature review. This imperative arises from the mounting pressures within academia, such as escalating research output, intricate interdisciplinary collaboration requirements, and evolving publication standards. Additionally, broader societal and policy mandates concerning transparency, reproducibility, and equity in scientific communication amplify this necessity [20, 76].

The advent of agentic systems—propelled by advances in artificial intelligence (AI) and intelligent agent-based architectures—represents a transformative shift in the execution and management of knowledge-intensive academic tasks [20, 76, 95]. These systems offer more than just operational efficiency and a reduction in human cognitive burden; they also foster the generation of reproducible, well-structured academic outputs, directly addressing policy-driven imperatives for robustness and accountability.

Among the various facets of academic automation, automated survey generation and agentic delivery systems have exceptionally transformative potential. By automating the identification, synthesis, and assessment of emerging scientific trends, these systems enable rapid, unbiased, and scalable literature reviews and meta-analyses—surpassing human limitations in both speed and coverage [20, 76]. Nevertheless, the integration of automation into academic protocols necessitates rigorous scrutiny, with particular emphasis on transparency, verifiability, and the preservation of scholarly integrity. Consequently, recent research has concentrated on embedding explainability and trackable provenance within AI-assisted tools, thereby setting the stage for more responsible deployment and assessment of automated academic practices [20, 23, 27, 70, 94, 110].

A pivotal conceptual advance in this context is the emergence of Generative Artificial Experts (GAEs)—a distinctive class of generative AI agents designed for complex, knowledge-intensive environments [95]. Unlike general-purpose generative AI systems, GAEs are defined by their generativity, domain-specific expertise, autonomy within well-bounded tasks, adoption of synthetic expert personas, and capacity for multimodal output generation. The foundational work on GAEs articulates a taxonomy structured around seven core traits, which delineate their capabilities and distinguish them clearly from legacy automation systems and conventional large language models. Notably, GAEs embody a hybrid paradigm: they combine the rule-based structure of expert systems with advances in human-AI collaboration and sophisticated generative modeling. This synthesis transcends basic automation, steering towards a synergistic augmentation of human scholarly activity [95].

1.2 Major Themes

The landscape of academia, as reshaped by AI, is characterized by several interrelated and evolving themes. Initially confined to repetitive, rote functions, automation now encompasses high-level

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

scientific reasoning, nuanced knowledge recognition, and even creative undertakings such as scholarly survey composition and scholarly style transfer. Recent studies underscore that leading Large Language Models (LLMs) are capable of automating scholarly content production, simulating expert-level reasoning, and emulating the stylistic idiosyncrasies of specific authors, including those from traditionally underrepresented backgrounds [94, 110]. The shift towards agentic and multi-agent systems further allows for the orchestration of diverse, specialized AI components, facilitating large-scale and collaborative scientific workflows [20, 95].

As agentic systems assume greater responsibility for scientific knowledge recognition—including the identification, contextualization, and explanation of domain-specific concepts—the demand for workflow transparency and robust explainability intensifies [20]. Benchmarking and standardized evaluation frameworks are thus emerging as essential tools—not only to quantify and compare system capabilities but also to build trust among academic stakeholders. Innovations in metadata tracking, notably the inclusion of standardized AI usage reports in scientific manuscripts, are enabling rigorous analyses of AI's impact on scholarly discourse, research transparency, and citation dynamics [23, 27]. In tandem, agentic architectures are broadening their applicability—from servicing under-resourced language communities to empowering high-stakes domains such as law and finance—through adaptable instruction tuning and scalable deployment across diverse contexts [20, 27, 70, 76, 110].

1.3 Contributions and Challenges

AI-assisted text generation and automated survey production yield significant benefits in academic productivity and inclusivity. These systems enable researchers to rapidly synthesize expansive literature bodies, generate structured academic reports, and implement style transfer in challenging out-of-distribution or low-resource settings [20, 27, 110]. Furthermore, the routine documentation of transparency and AI-assisted metadata within academic publishing workflows establishes new standards for accountability, enhances reproducibility, and supports large-scale investigations of generative AI's societal influence [23, 27].

However, several critical challenges impede the responsible and widespread adoption of agentic systems in academia. Chief among these is the ongoing absence of standardized guidelines specifying how, when, and to what extent AI tools should be integrated into research and publication pipelines [20, 23]. The increasing autonomy of agentic systems introduces further complexities surrounding explainability, provenance, and user trust. Moreover, resource disparities—indexed by computational costs and the availability of high-quality data—limit the reach of advanced AI applications, particularly in underrepresented domains and languages [20, 70, 110]. Finally, the evaluation of agentic systems, specifically with respect to task fidelity, robustness, and domain generalizability, remains an open research challenge.

These challenges may be systematically compared based on three core criteria: standardization, resource equity, and evaluation robustness. Table 1 summarizes these challenges, along with their primary implications for the field.

Overcoming these challenges will necessitate sustained interdisciplinary collaboration, refinement of metadata and reporting standards, and principled development of agentic AI architectures. Ultimately, such advances must strive not only for technical excellence and operational flexibility but also for transparency, auditability, and alignment with the fundamental values of academic and societal stakeholders [20, 23, 27, 70, 76, 94, 110].

2 Theoretical, Methodological, and Workflow Foundations

2.1 Scientific Standards and Methodologies

Ensuring scientific rigor and reproducibility is fundamental to both traditional and automated research practices. Core protocols—such as PRISMA for systematic reporting, AMSTAR-2 for methodological assessment, and GRADE for evidence quality—constitute the backbone of evidence synthesis and meta-analysis workflows. As research becomes increasingly digitized and automated, these standards persist as the benchmarks for quality, transparency, and trustworthiness, even as contemporary methodologies adapt their implementation for digital contexts [6, 12, 14, 16, 21, 25, 26, 31, 44, 67–69, 84, 88, 100, 102, 105, 112].

With the adoption of automation, adherence to scientific standards must be complemented by their continual adaptation. This entails explicit, context-specific operationalizations of reporting guidelines, risk of bias assessment, and transparent evidence grading—especially as machine learning (ML) and natural language processing (NLP) deliver unprecedented gains in time and labor efficiency. Automation introduces new complexities in maintaining transparency: for instance, ensuring standard-compliant reporting becomes more challenging when portions of the workflow are abstracted or obfuscated by algorithms.

Methodological innovation across scientific domains now encompasses a broad spectrum of manual, automated, and hybrid workflows. A critical axis of differentiation is the degree of explainability: workflows can be fully explainable, partially explainable (hybrid), or predominantly "black-box" in AI-driven components. The integration of Explainable AI (XAI) and Human-Centered AI (HCAI) strategies aims to maximize interpretability and accountability, thereby preserving scientific integrity and fostering confidence among stakeholders in automated research pipelines [12, 14, 85, 86]. Nonetheless, major limitations persist—particularly for complex deep learning architectures and large language models (LLMs). In these cases, decision pathways are often opaque, presenting formidable challenges to transparency and post-hoc scrutiny [6, 12, 45, 50].

Hybrid systems—which strategically combine automation with domain expert oversight—are gaining traction as they balance efficiency and reliability. These workflows typically allocate high-confidence, routine tasks to AI, while flagging ambiguous or complex cases for human adjudication. Empirical studies demonstrate that, in contexts such as text classification or survey coding, threshold-based partitioning methods can automate over 70% of the workload, while human reviewers handle cases with lower model certainty. This approach enables optimized resource utilization while effectively reducing systematic errors and minimizing bias propagation [4, 10, 15, 47, 79, 91, 93].

Table 1: Principal challenges in integrating agentic systems into academic workflows, mapped to their primary implications.

Challenge	Description	Primary Implication
Lack of Standardization	Absence of unified protocols for AI integration in research and publishing workflows	Inconsistent adoption, difficulty assessing AI contributions, potential ethical/legal ambiguities
Resource Inequity	Disparities in computational resources and access to high-quality data, especially in low-resource settings	Restricts reach and fairness of AI-powered systems, risks bias and underrepresentation
Evaluation and Benchmarking Gaps	Limited standardized methods for benchmarking task fidelity, robustness, and cross-domain generalizability	Unclear performance baselines, barriers to comparative research and validation
Explainability and Provenance	Challenges in making agentic outputs transparent and attributable	Reduced trust, difficulty in auditing and verifying scholarly processes

Table 2: Comparison of Explainability and Oversight Across Workflow Types

Workflow Type	Explainability	Human Oversight	Automation Role
Manual	Full	Complete	N/A
Hybrid (XAI/HCAI)	Partial/High	Targeted	High for routine tasks
Black-box AI	Low	Limited; post-hoc	Predominant; tasks of any complexity

As shown in Table 2, hybrid and explainable approaches afford greater transparency and targeted oversight than black-box systems, directly influencing both the quality of output and the trust of end-users.

Concurrently, educational applications highlight the dual challenges and potentials of automation: automated item generation (AIG) for multiple-choice questions can match the quality of traditional methods, provided that cognitive models and author training are robust. However, the field continues to debate the sufficiency of such tools for evaluating complex reasoning and writing skills, which underscores the ongoing necessity of human input, continuous methodological refinement, and the maintenance of rigorous evaluation criteria [6, 12, 45, 85].

Despite substantial progress, salient challenges remain concerning the standardization of reporting, replicability, methodological rigor, and equitable distribution of the benefits of research automation. The interplay between established methodological protocols, the evolution of best practices, and the responsible adoption of automation is thus situated at the dynamic intersection of reproducibility, efficiency, and scientific accountability [6, 12, 14, 16, 21, 26, 31, 68, 69, 100, 102, 112].

2.2 Workflow and Technical Architecture

The landscape of automation architectures has shifted rapidly from unimodal, siloed ML approaches to highly integrated, multimodal, agentic, and distributed hybrid systems [1, 9, 12, 14, 17, 18, 23, 31, 36, 42, 43, 45, 56, 60, 62, 64, 71, 76, 80, 86, 90, 94, 111, 114]. This transition is propelled by technological advances including the deployment of LLM-based agents for coordinated multi-agent reasoning, the coupling of multi-modal deep learning systems that synchronize across text, images, and structured data, and the integration of workflow platforms that seamlessly connect diverse tools and repositories.

Case studies in applied survey automation illustrate the strategic advantages of composable, cloud-native architectures. For instance, deployments using platforms such as WhatsApp Business API and Twilio in concert with Google Sheets allow for flexible, scalable survey delivery and longitudinal data collection among mobile and hard-to-reach populations. These architectures not only extend reach but also ensure adaptability, though they require careful orchestration of technical details, thoughtful engagement strategies to maintain response rates, and robust safeguards to uphold privacy and data integrity [28].

Automated scholarly writing and review pipelines have matched this technical progression. LLM-driven, automated literature survey generation now achieves high throughput and favorable performance on metrics such as topical coverage and citation alignment. Nevertheless, persistent barriers include limited context understanding, citation inaccuracies, and model misalignments with research aims [12, 18, 31]. Systematic review automation—often configured as a modular pipeline of document retrieval, screening, citation network analysis, and topic clustering—employs iterative human-expert involvement to compensate for the variable quality and focus of automated extraction and synthesis modules [14, 16, 69, 80, 88, 90]. Deliberately modular system design allows architects to isolate potential points of failure or black-box reasoning, providing opportunities for targeted human intervention or supplemental validation via parallel workflows.

The ongoing transition toward multimodal and distributed agentic architectures is driven by advances in multi-agent systems (MAS). Distributed intelligence and decentralized decision-making not only improve scalability and adaptability but also enhance system resilience [9, 17, 23, 42, 43, 56, 62, 64, 71, 76, 94, 111, 114]. Applications range from multi-agent optimization in logistics and smart infrastructure to distributed scientific workflows, all predicated on robust coordination protocols, communication languages—both human-inspired and synthetic—and the development of emergent collective intelligence. However, increased system complexity raises new challenges in inter-agent explainability, system-level transparency, and resistance to adversarial or ambiguous stimuli.

Research workflow optimization strategies have embraced agent-based and hybrid computational frameworks. Multiphase survey pipelines and automated coding systems delegate routine and repetitive tasks to AI modules, reserving complex inferences and qualitative judgments for human experts. This division has proven highly effective in contexts characterized by high data throughput, task ambiguity, or challenging participant recruitment, supporting scalable delivery without sacrificing methodological rigor [4, 10, 12, 14, 15, 18, 45].

Despite the pace of innovation, several open problems remain: the portability of automation architectures across disparate domains, the standardization of interoperability protocols, and harmonization with evolving scientific reporting requirements persist

as key obstacles to fully integrated, cross-disciplinary research workflows [6, 12, 16, 21, 102]. Achieving seamless synergy between automated, agentic, and human-in-the-loop systems will demand both continual technical refinement and robust governance structures, particularly as considerations of data privacy, intellectual property, and professional ethics evolve in tandem with technological change.

2.3 Roadmapping and Evolution

The convergence of traditional academic methodologies with emergent AI-driven, agentic, and distributed paradigms represents both a preservation of foundational scientific standards and a transformative reconfiguration of workflows, labor distribution, and the overall scope of research activity. The progressive shift from exclusively manual procedures toward AI-augmented and, in select fields, fully automated workflows encapsulates both the immense opportunities and significant tensions introduced by academic automation [1, 9, 12, 31, 36, 43, 86, 90, 114].

Adoption strategies have been guided by the imperative to leverage efficiency gains while steadfastly upholding reliability, transparency, and privacy. Incremental adoption—characterized by the initial automation of high-volume, routine, or reproducible tasks—has repeatedly proven to be an effective pathway. This approach is most successful when coupled with continuous reassessment of both methodological and quality frameworks [6, 14, 21, 102]. Retaining the indispensable critical judgment of human experts, especially for complex synthesis, ethical evaluations, and atypical cases, gives rise to advanced hybrid models. These models optimize the division of labor between automated systems and human agents, dynamically allocating tasks based on model confidence, assessed risk, and ethical implications [4, 10, 14, 47, 86].

Privacy concerns and integration challenges are accentuated by the scaling of automated and distributed systems, which routinely process sensitive, proprietary, or high-dimensional personal data. Addressing these issues necessitates rigorous adherence to privacy-by-design tenets, robust access controls, and comprehensive monitoring for bias and misuse. As agentic, decentralized, and cloud-based platforms proliferate, these requirements only become more demanding [12, 17, 31, 42, 69, 94].

As intelligent systems increasingly mediate research workflows, the scholarly community faces mounting demands to formalize software and workflow documentation, establish open and reproducible evaluation frameworks, adopt living reviews and continuous benchmarking, and promote collaborative open science ecosystems [6, 12, 14, 18, 21, 102]. The overarching trajectory is evident: realizing the full benefits of academic automation will depend on sustained, multidisciplinary collaboration that conscientiously reconciles technological innovation with the enduring values of transparency, trust, and stewardship at the core of scientific progress.

2.4 Automated and Hybrid Survey Systems

2.4.1 Innovations in Survey Administration. The evolution of automated and hybrid survey systems has substantially transformed the landscape of large-scale data collection, management, and analysis. These advancements have yielded notable improvements in participant reach, data quality, and overall cost-effectiveness. Central to this transformation are highly scalable platforms that incorporate chat-based user interfaces, advanced branching logic, and automated reminder functionalities. Such features reduce barriers to participation and help mitigate attrition rates, particularly among mobile and hard-to-reach populations [28]. The integration of tools such as Twilio and Google Sheets exemplifies the modularity and extensibility of contemporary survey systems, streamlining workflows from real-time data ingestion through to downstream analytics [28].

Beyond gains in technical efficiency, automation interfaces deeply with the sociotechnical fabric of participation and norm enforcement within distributed environments. As survey systems become increasingly decentralized and collaborative, dynamic processes for the (re)design and revision of shared behavioral norms grow in importance. Ensuring that automated mechanisms remain adaptive and robust in multi-agent contexts necessitates the continuous synthesis, enforcement, and iterative revision of normative guidelines [30]. The complexity of this challenge is heightened by evolving research goals and the diverse expectations of users. Recent advancements in data-driven norm revision offer promising solutions, drawing upon behavioral trace data to incrementally calibrate system-level rule sets and thereby improving the accuracy of distinguishing compliant from non-compliant behaviors [30].

The expansion of automation has not only increased the scale and diversity of survey participation but has also elevated the importance of inferring agent or respondent motivations from observed behaviors. Approaches grounded in inverse reinforcement learning provide frameworks for extracting latent utility functions, equipping survey systems to adapt their interaction strategies and the interpretation of collected data to more accurately reflect the goals and incentives of heterogeneous participant populations [41]. The ability to personalize, mitigate bias, and augment the interpretability and generalizability of responses thus emerges as a cornerstone of modern, distributed, agent-driven survey architectures.

2.4.2 Case Study: WhatsApp-Based Survey Automation. Recent deployments leveraging widely-used messaging platforms, particularly WhatsApp, highlight the progressive shift toward ubiquitous and user-centric research modalities. Utilizing the WhatsApp Business API, coupling with intermediaries like Twilio, and integrating cloud platforms such as Google Sheets, researchers have established survey workflows that engage participants within their preferred channels of communication. This approach has democratized both access and usability, embodying a pivotal advancement in inclusive research methodology [28]. Architectures built upon these platforms enable chat-based interfaces with dynamic branching, scheduled reminders, automatic error handling, and robust support for longitudinal data collection [28].

Empirically, such systems have been shown to significantly increase completion rates and reduce per-respondent costs when compared to traditional modalities such as SMS or IVR. Engagement has been sustained even among highly mobile and marginalized populations—contexts that are conventionally associated with high attrition [28]. Nevertheless, the implementation of these systems introduces certain vulnerabilities: technical setup can be demanding, and dependencies on proprietary APIs may result in fragility, particularly when faced with unpredictable message delivery. Attrition,

though ameliorated, persists as a challenge; and escalating workflow complexity, translation requirements, and incentive mechanisms can undermine uniform data quality [28].

Optimal implementation, therefore, requires a careful balance among accessibility, workflow optimization, translation and localization, randomization techniques, and proactive engagement strategies. As these systems scale, the necessity for next-generation quality assurance methods becomes acute. Key innovations include automated detection of response patterns, disengagement monitoring, and adaptive, agent-driven interventions underpinned by behavioral inference methodologies rooted in inverse reinforcement learning [41]. Looking forward, the field anticipates expansion to additional communication platforms, enhanced multilingual capability, and deeper integration with established research ecosystems to further streamline data analysis and deliver timely feedback.

The structural and performance differences among WhatsApp-based, SMS-based, and IVR-based survey platforms are synthesized in Table 3.

2.4.3 Automated Question Generation and Assessment. Progress in AI-driven question generation (QG) and automated assessment methodologies has further accelerated the modernization of survey systems. Large, open-access datasets such as SQuAD, MS MARCO, RACE, and SciReviewGen serve as the bedrock for the development and benchmarking of QG systems. These systems are capable of generating both objective (e.g., multiple-choice, cloze) and subjective (e.g., short or long answer) questions drawn from diverse source materials [2, 4, 25, 39, 47, 52, 55, 59, 63, 75, 83, 84, 89, 91, 93, 110, 114]. Current QG paradigms encompass rule-based, natural language processing, and deep learning approaches, each exhibiting particular strengths and limitations: rule-based methods facilitate interpretability and precision, whereas data-driven approaches afford scalability and domain adaptability [25, 63, 75, 114].

Automated response assessment has seen parallel advancements. Comparative studies indicate that automatic item generation can produce multiple-choice questions with quality approaching that of human experts, provided that robust cognitive modeling and authorial expertise guide the system [84]. Automated graders have reached high accuracy in evaluating both short and long answers across languages and response formats, employing a synthesis of linguistic, semantic, and statistical features to achieve human-level—sometimes even superhuman—agreement [83, 89]. Introducing hybrid assessment approaches, which combine algorithmic categorization of unambiguous responses with targeted human review for more complex cases, has resulted in considerable efficiency gains without sacrificing accuracy [91, 93].

Nonetheless, significant obstacles remain. Automated assessment frameworks at scale frequently falter when tasked with providing reliable feedback on open-ended or higher-order reasoning responses, and standardization across content domains remains a challenge [25, 63, 83, 84]. Ongoing research is focused on extending these systems to handle multimedia inputs, detect automatically generated or low-quality content, and enhance explainability, fairness, and adaptivity [39, 47, 59, 110, 114]. Furthermore, hybrid systems integrating comprehensive automation with strategically deployed human oversight are emerging as best practices in high-stakes or methodologically complex survey environments [55, 83, 93].

Altogether, these innovations firmly position automated and hybrid survey systems at the forefront of methodological advancement. They offer compelling paradigms for scalable, inclusive, and adaptive data collection and analysis, while also illuminating challenges inherent in automation—from technical workflow optimization and behavioral inference to ongoing normative and response assessment refinement in dynamic, multi-agent research settings [2, 4, 25, 28, 30, 39, 41, 47, 52, 55, 59, 63, 75, 83, 84, 89, 91, 93, 110, 114].

3 AI and Agentic Systems in Academic Knowledge Recognition and Survey Automation

3.1 AI for Scientific Knowledge Recognition and Automation

Recent advancements in artificial intelligence (AI) and natural language processing (NLP) have fundamentally reshaped approaches to scientific knowledge recognition, introducing new paradigms for semantic parsing, terminology tracking, and the systematic synthesis of the burgeoning academic literature. Central to these innovations is the pursuit of automating literature review processes to achieve scalability and reproducibility in response to the exponential acceleration of scholarly output across disciplines [76]. Seminal systems—such as AutoSurvey, SurveyX, and SurveyForge—epitomize the current state of the art. These platforms leverage large language models (LLMs), retrieval-augmented generation (RAG), and advanced knowledge base structuring (e.g., AttributeTree) to facilitate survey generation automation, citation validation, and optimized content coverage [3, 4, 11, 12, 14–16, 21, 26, 39, 47–50, 52, 65, 69, 74, 89, 91, 93, 94, 100, 102, 105, 109, 111–113].

These systems deliver significant improvements in efficiency and accuracy. For example, AutoSurvey not only multiplies survey throughput by several orders of magnitude relative to manual efforts but also preserves high citation recall and precision, thus decreasing both cost and human labor demands. Nonetheless, persistent limitations temper their efficacy. Among the most salient challenges are citation misalignment—where misinterpretation and overgeneralization constitute the predominant sources of error—contextual window constraints inherent to LLMs, and difficulties with domain adaptation, as most models rely on pre-prints (e.g., arXiv) rather than peer-reviewed sources [65, 105, 109]. SurveyX, characterized by a multi-phase, hybrid retrieval pipeline and structured knowledge bases, partially ameliorates issues of reference relevance and organizational coherence, bringing automated surveys notably closer to human quality. However, this progress continues to be bounded by an intrinsic trade-off: greater breadth of coverage often occurs at the expense of granularity or interpretability in the synthesized outputs [65].

Automated scientific knowledge recognition extends beyond literature review to encompass the semantic analysis of terminology, tracing the origins, evolution, and contextual usage of terms across multiple disciplines. This functionality is crucial for tracking conceptual diffusion and clarifying lexical ambiguities, yet it remains

Table 3: Comparative Features of Automated Survey Platforms Utilizing WhatsApp, SMS, and IVR

Feature	WhatsApp-Based	SMS-Based	IVR-Based
Chat-based interface	Yes	Limited	No
Dynamic branching	Yes	Moderate	Limited
Multilingual support	Moderate–High	Low–Moderate	Moderate
Automated reminders	Yes	Yes	Yes
Integration with cloud tools	High	Moderate	Low
Cost per respondent	Low	Moderate–High	High
Technical setup complexity	High	Moderate	Low
Resilience to attrition	High	Moderate	Low
Accessibility for marginalized	High	Moderate	Moderate

an insufficiently addressed component within contemporary automated frameworks [76]. While leading LLM-powered workflows have introduced preliminary features for glossary extraction and scientific entity annotation, adaptive systems explicitly modeling the temporal and cross-disciplinary semantic drift of key concepts are lacking, thus marking a crucial avenue for future research [52, 76].

The dependability of AI-powered literature review and knowledge extraction is significantly augmented by human-in-the-loop (HITL) methodologies and adaptive machine learning paradigms. Empirical studies indicate that integrating human judgment—through iterative relevance feedback or retrospective validation—yields higher accuracy and a stronger foundation of trust than fully autonomous solutions [11, 12, 15, 50, 52, 91, 93, 113]. Additionally, the establishment of open benchmarks and standardized reporting frameworks, such as PRISMA and GRADE, is essential. Such standards are not merely procedural; they are foundational for reproducibility, comparability, and fair evaluation, especially as AI-driven tools diversify in architecture and target domain [6, 12, 61, 94, 110]. Despite these initiatives, actual adoption and harmonization remain stymied by inconsistent terminology, fragmented evaluation metrics, and the lack of universally accessible annotated datasets [11, 12, 91, 110].

While NLP and AI techniques—including neural text classification, active or weak supervision, and embedding-based retrieval—have achieved impressive technical results, limitations in usability, transparency, and accessibility for end users remain significant obstacles [14, 26, 49, 91, 111, 113]. Many tools are optimized for technically proficient researchers and lack crucial interpretability features—such as rationale extraction or explainable AI components—that are vital for adoption in medicine and multidisciplinary fields [4, 15, 48, 111]. In summary, although contemporary systems provide notable advancements in automating evidence synthesis and academic content organization, broader adoption will necessitate heightened emphasis on interpretability, adherence to open standards, and sustained human oversight. This recalibration is essential to counteract risks of bias, misclassification, or degradation of methodological rigor [11, 14, 15, 26, 49, 74, 76, 94].

As illustrated in Table 4, these systems embody diverse methodological choices and design trade-offs, underscoring the need for comprehensive benchmarks and harmonized evaluation criteria to ensure cross-system comparability and continued progress.

3.2 Agent-Based Recognition Systems in Video and Academic Applications

Agentic systems—distinguished by autonomy, adaptability, and collaborative capability—are exerting a transformative influence on both behavioral video analysis and the deployment of distributed intelligence in academic and medical contexts. In the realm of video-based behavioral recognition, systems such as facial micro-expression recognition (FMER) prominently demonstrate the strengths of agentic frameworks: through contour extraction and distance-based metrics, these platforms efficiently decode subtle social signals, achieving high accuracy and workflow throughput [34]. The significance of micro-expression recognition extends beyond technological achievement in pattern recognition; it also exemplifies the adaptability of modular, agent-driven architectures for parsing complex, temporally resolved data streams—a paradigm readily transferable to the analysis of multimodal academic and clinical datasets.

More broadly, agent-based architectures provide the foundational framework for advanced systems in the medical Internet of Things (IoT) landscape, such as the Smart Agent-based Privacy Preservation and Threat Mitigation Framework (SAPPTMF) [87]. SAPPTMF demonstrates the collaborative efficacy of distributed, privacy-focused agents in Internet of Medical Things (IoMT) models for monitoring and neutralizing threats to sensitive health data. By simulating a range of adversarial scenarios and applying analytic hierarchy processes to prioritize security interventions, these agent-based systems validate how modularity, adaptivity, and formal model-driven reasoning translate into heightened practical robustness. Specifically, SAPPTMF achieves notable gains in accuracy, precision, and recall for threat detection tasks, testifying to the applicability of agentic theory in high-stakes environments [87].

The agent-based paradigm further streamlines workflow integration by coordinating distributed sensing, computational reasoning, and decision-making processes across both physical (e.g., wearables, autonomous vehicles) and informational (e.g., video analytics, automated survey workflows) domains. This integration offers system-level advantages, including scalable multi-agent orchestration, extensibility for emerging data modalities, and the incorporation of dynamic feedback from human supervisors or other AI components—a process aligning closely with the adaptive learning cycles central to modern machine learning [6, 12, 110]. Nevertheless, to

Table 4: Comparison of Leading AI-Driven Survey Automation Systems

System	Core Technologies	Key Strengths	Main Limitations
AutoSurvey	LLM; RAG; AttributeTree knowledge base	High citation recall and throughput, scalable automation	Citation misalignment; domain adaptation issues (pre-prints oriented)
SurveyX	Hybrid retrieval pipeline; structured KB	Improved reference organization, multi-phase processing	Trade-off between breadth and granularity of content
SurveyForge	LLM with citation validation, open benchmarks	Content coverage optimization, ties to reporting frameworks	Development dependent on inconsistent evaluation standards

realize these potential benefits, it is essential to address persistent challenges such as ensuring data privacy, reinforcing robustness against distributed attacks, and developing interpretable outputs that are readily consumable by human decision-makers to sustain trust and accountability [11, 15, 50, 110].

3.3 Intelligent Agent-Based Survey Delivery

Automation of survey data acquisition and analysis constitutes a pivotal application for intelligent agent-driven systems. Conventional survey strategies, grounded in static questionnaires, rigid operational structures, and centralized data management, are fundamentally constrained in their adaptability, scalability, and capacity for respondent engagement. In contrast, contemporary surveying solutions increasingly adopt multi-agent, modular, and distributed architectures that enable real-time monitoring, adaptive sensing, and granular targeting of participant subgroups [17][42][28].

Technological progress in this arena can be observed in multi-agent platforms that orchestrate mobile surveys—ranging from WhatsApp-based and IoT-enabled delivery systems to comprehensive, real-time data quality surveillance and instant feedback integration. These frameworks support flexible deployment, continuous longitudinal engagement, and context-sensitive adjustments, thereby enhancing both data quality and response rates while easing the administrative demands traditionally associated with large-scale data collection [17][42]. Moreover, agent-driven survey platforms demonstrate unique value in engaging hard-to-reach or highly mobile populations, as evidenced in implementations serving refugee communities or geographically dispersed cohorts [28]. These systems routinely outperform traditional modalities such as SMS or interactive voice response (IVR) channels in cost efficiency and completion rates.

The principal advantage of agentic survey infrastructures lies in their ability to integrate hybrid human-AI workflows. For instance, the automation of straightforward open-ended response classification can be paired with human coder intervention for complex or ambiguous cases, delivering substantial efficiency improvements—with manual workload reductions of up to 80% documented in various studies—while preserving the reliability of coding outcomes [17][42]. Additionally, distributed architectural designs facilitate automatic cross-checks for data integrity, scheduled participant engagement reminders, and secure, privacy-aware data handling, all of which are essential for robust management of large, diverse, and sensitive respondent pools.

Despite these advances, challenges persist regarding system integration, interoperability, and the mitigation of sample attrition or engagement decline across extended study durations [28]. Continuous benchmarking against traditional survey methodologies remains imperative for accurately quantifying the value added by agent-based approaches across varied research domains. As intelligent agents move toward deeper integration within survey infrastructures, high-priority objectives include the development of robust threshold protocols for amalgamated human/AI coding, automated detection and rectification of technical biases or model drift, and the enhancement of decision-making transparency through interpretable modules [110][42].

4 Advanced Agent-Based Modeling and Multi-Agent Systems

4.1 Agent-Based Modeling Paradigms

Agent-Based Modeling (ABM) has established itself as a key paradigm for representing complex systems characterized by heterogeneous, interacting entities. Its versatility extends across domains such as transportation, logistics, and collaborative systems. Unlike traditional aggregate or equation-based models, ABMs inherently capture discrete, autonomous agents whose micro-level interactions generate emergent system-level behaviors that elude simple analytical prediction [71][56]. This modeling flexibility is crucial for studying non-linear dynamics and intricate dependencies that typify real-world systems, as seen in simulations of transport networks, collaborative logistical operations, and decentralized decision-making environments [71][56].

A defining strength of ABM is its capacity to represent individual-level heterogeneity and trace how agent interactions propagate to macro-scale phenomena. In transportation systems, for instance, ABMs faithfully model the interplay among user behavior, traffic flow, and service reliability—phenomena that aggregated models often misrepresent [71]. Similarly, logistics and supply chain networks benefit from agent-based simulations that dynamically allocate resources, model congestion, and evaluate resilience under varied perturbations [56]. The evolution of on-demand, decentralized services has further highlighted ABM's relevance; distributed models effectively represent reactive responses in on-demand transport scenarios, integrating operational constraints via decentralized agent architectures and heuristics such as A*-based routing [71][56].

Nevertheless, the expressiveness of ABMs presents significant methodological and computational hurdles. Scalability poses a persistent challenge, especially in high-agent-count systems, those with intricate behavioral rules, or scenarios demanding real-time execution. Consequently, recent research advocates for decentralized and on-demand ABM designs that partition computational loads and enhance integration with real-world automation tasks [71][56]. By endowing agents with local autonomy, these approaches mitigate central bottlenecks, yet they also introduce complex issues concerning model validation and synchronization. To address optimization challenges within these agentic systems, researchers have increasingly leveraged metaheuristic and nature-inspired approaches, such as multi-agent Particle Swarm Optimization (PSO),

wherein dynamic neighborhoods and cognitive agent autonomy improve global search efficacy and solution quality relative to conventional methods [23].

As ABM methodologies mature, the focus has shifted towards automated and adaptive design processes. The Automated Design of Agentic Systems (ADAS) marks a substantial advancement, utilizing meta-agent frameworks to autonomously generate, code, and evolve agents using Turing-complete languages. In this paradigm, a meta-agent iteratively constructs and refines an archive of agentic solutions, enabling the autonomous discovery of novel agent architectures, prompt compositions, and tool integrations that can outperform manually designed counterparts across coding, scientific, and mathematical domains [70]. These self-improving ABM frameworks not only expedite innovation but also promote systematic exploration of agentic design spaces, signifying a promising trajectory for the automation and optimization of agent-based systems.

4.2 Hybrid and Decentralized Agentic Architectures

The escalating heterogeneity and scale inherent in modern multi-agent environments have precipitated the development of hybrid and decentralized architectures designed to facilitate robust, privacy-preserving collaboration. Contemporary frameworks routinely integrate software, physical, and human agents across diverse applications—spanning autonomous vehicles, cyber-physical-social systems (CPSS), traffic management, and large-scale recommendation systems [1, 9, 10, 18, 23, 36, 38, 43, 53, 60, 63, 64, 68, 73, 80, 92, 106, 115]. The core challenge lies in orchestrating seamless agent interactions, fostering adaptability to evolving contexts, and balancing efficiency, privacy, and robustness.

The rise of Large Language Models (LLMs) has catalyzed a new era in Multi-Agent Systems (MAS), enabling agents instantiated with distinct profiles to collaborate via perception, reasoning, planning, and structured communication protocols [20, 23, 27]. Unified MAS frameworks commonly outline fundamental modules including agent perception, self-action, mutual communication, and evolutionary learning. These frameworks find application in collaborative software engineering, robotics, scientific knowledge synthesis, and the simulation of sophisticated virtual societies [23, 27]. Empirical studies reveal the emergence of credible, human-like autonomous collectives, exhibiting capacities for coordinated intelligence and cross-disciplinary problem solving [23].

Hybrid architectures frequently incorporate both symbolic and sub-symbolic reasoning. In CPSS contexts, for example, agents model hybrid attacks involving cyber, physical, and social vectors. Through reinforcement learning, these agents adapt to adversarial scenarios—such as denial-of-service or misinformation campaigns—enabling the study of evolving opinion dynamics, resilience, and collective utility [110]. This data-driven, adaptive simulation paradigm signifies a substantive shift from traditional static or strictly rational game-theoretic methods, permitting richer, more actionable insight for defense-in-depth strategies in critical infrastructures [110]. Additionally, the integration of knowledge graphs and first-order logic into multi-agent simulations, as in autonomous driving systems, empowers knowledge-fusion agents to synthesize deep learning with rule-based reasoning, which enhances safety, rule compliance, and environmental perception compared with purely data-driven approaches [17].

Decentralized agentic architectures have grown increasingly prominent, particularly in distributed data environments. For example, the Multiple Coordinative Data Fusion Modules (MCDFM) framework orchestrates distributed preprocessing, filtering, and decision-making both locally and across networks using software agents, Kalman filters, and fuzzy logic [80]. By supporting adaptive, real-time traffic light control, such multi-agent frameworks enable resilient information exchange and optimize resource utilization amidst heterogeneous, asynchronous data streams.

From an algorithmic and representational standpoint, graph-based models—including multi-layer synchronous dynamical systems and graph neural networks (GNNs)—extend ABM reasoning to richly structured, multi-relational interaction networks [1, 92, 106]. These models create a bridge between discrete agent decision-making and the expressive modeling capabilities of graph-based learning and inference, facilitating significant advancements in areas like collaborative filtering, personalized recommendation, and epidemic modeling.

A concise comparison of central approaches in contemporary multi-agent system design is provided in Table 5, highlighting paradigmatic differences in architecture, reasoning, and typical applications.

Despite significant progress, persistent challenges remain. Achieving scalable agent coordination, developing robust privacy mechanisms, ensuring transparency and explainability, and integrating learning, reasoning, and planning at scale are ongoing obstacles [20, 23, 27, 70, 110]. LLM-based MAS, while demonstrating great potential, are hindered by black-box dynamics, susceptibility to biases, and high computational demands [23, 27]. Advancement in the field now hinges on formulating rigorous validation frameworks, designing adaptable reward functions for reinforcement learning in multi-agent contexts, developing scalable privacy-preserving protocols, and establishing standardized evaluation benchmarks [20, 23, 27, 42, 56, 70, 71, 110]. Key research directions encompass the adoption of inverse reinforcement learning for preference elicitation, the refinement of interpretable meta-agent frameworks, and the expansion of hybrid architectures into emerging domains, such as edge computing and privacy-sensitive collaborative environments [27, 70, 110].

In summary, the synthesis of agent-based modeling advances, decentralized architectures, and hybrid agentic frameworks now defines the leading edge of agent-system research. It is the interplay among scalable model architectures, adaptive agent reasoning, and resilient coordination strategies that underpins the transformative potential of agent-based systems across science, engineering, and society [1, 9, 10, 17, 18, 20, 23, 27, 36, 38, 42, 43, 53, 56, 60, 63, 64, 68, 70, 71, 73, 80, 92, 106, 110, 115].

5 Workflow, Automation, and AI Writing Assistance

5.1 Automated and Hybrid Workflows

Recent advancements in artificial intelligence (AI) and natural language processing (NLP) are fundamentally transforming the

Table 5: Comparison of Multi-Agent System Paradigms in Contemporary Applications

Paradigm	Reasoning Approach	Architecture	Representative Applications
Traditional ABM	Rule-based, stochastic	Centralized/decentralized	Transport, logistics, social simulation
Hybrid MAS (Symbolic+Sub-symbolic)	RL, logic, DL fusion	Centralized/distributed	CPSS, robotics, autonomous driving
LLM-based MAS	Language-model reasoning, planning	Distributed, profile-based	Collaboration, software engineering, virtual societies
Graph-based MAS	Structured graph inference	Synchronous/asynchronous, multi-relational	Recommendation, disease modeling, traffic control

landscape of scholarly document processing, particularly in the realm of systematic reviews, evidence synthesis, and academic writing. The availability of integrated, end-to-end pipelines—which encompass stages from retrieval and screening to synthesis and authoring—now enables researchers to address longstanding hurdles stemming from the accelerating growth of scientific literature and the need for rigor, reproducibility, and efficiency [3, 4, 11, 12, 14–16, 21, 26, 39, 45, 47–50, 52, 65, 69, 74, 89, 91, 93, 94, 100, 102, 105, 109, 111–113].

One emergent trend lies in shifting from automating isolated tasks, such as citation screening, to implementing hybrid workflows that seamlessly blend AI, peer, and instructor contributions. These orchestrated processes ultimately enhance both productivity and output quality. For instance, systems including AutoSurvey, SurveyX, and SurveyForge operationalize modular pipelines that commence with advanced reference retrieval. These pipelines utilize large language models (LLMs), embedding-based search, and multi-phase filtering to curate topic-aligned outlines and construct knowledge graphs [45, 74, 105, 109]. Next, LLM-driven content generation proceeds under the guidance of hierarchical outlines, iterative human feedback, and explicit citation-checking stages. This approach consistently delivers outputs with citation accuracy and topical coverage that are comparable to, or at times exceed, those of human-authored surveys, all while achieving unprecedented throughput and cost-efficiency [14, 16, 49, 74, 105, 109]. Despite these benefits, persistent limitations—such as citation hallucinations, misalignments, and challenges in integrating knowledge from diverse sources—continue to be widely reported in the literature. These issues underscore the necessity for robust human oversight and hybridized solutions [14, 39, 48, 49, 65, 69, 74, 105].

Hybrid human-AI workflows are particularly evident in tools for evidence synthesis supporting systematic reviews. Here, deep learning classifiers and active learning methodologies have demonstrated their ability to reduce manual screening burdens by an estimated 60–94% in practice. However, the highest quality outcomes are realized when these machine-generated recommendations are complemented by structured peer or expert interventions, such as triaging low-confidence cases to human evaluators or implementing scenario-specific model tuning [4, 11, 12, 15, 16, 39, 50, 52, 66, 91, 93, 100, 102, 107, 111–113]. This hybrid approach is regularly appraised using multidimensional evaluation frameworks—encompassing precision, recall, and interrater reliability—to ensure not only efficiency gains but also reproducibility, domain adaptation, interpretability, and trustworthiness [4, 11, 12, 15, 16, 21, 26, 39, 45, 47, 50, 52, 69, 89, 93, 102, 111, 113].

The strategic integration of peer, instructor, and algorithmic feedback constitutes the foundation of next-generation academic workflows. Accumulating evidence indicates that researchers are more likely to iteratively refine their writing when supported by synergistic peer and AI-assisted feedback, producing higher quality manuscripts and more robust scholarly discourse [4, 11, 50, 52, 66, 93, 107, 113]. At the same time, human-in-the-loop annotation, hybrid screening, and active learning inform triage processes help to mitigate the dangers of over-reliance on either humans or algorithms—a precaution that is particularly vital in domains such as clinical guidelines or policy-related reviews, where the stakes are high.

5.2 Multilingual and Inclusive Tooling

Advances powering equitable scholarly communication are driving the development of AI tools designed to serve multilingual and under-resourced research communities. The historical predominance of English-centric models and datasets has long perpetuated disparities in research dissemination and global knowledge access [11, 27, 40, 47, 58, 89, 108, 110]. Recent innovations now address these inequities by fine-tuning large language models and domain-adapted systems, such as MindLLM, for underrepresented languages; expanding tokenization strategies; and supporting variable-length, domain-specific texts [27, 40, 47, 58, 89]. Experimental deployments in languages including Amharic and Spanish illustrate that, with targeted data augmentation and adaptive post-processing, tools for automated summarization and evaluation can attain or surpass the performance of their English-language counterparts [11, 47, 58, 108, 110]. In addition, AI-based multilingual modules—encompassing cross-lingual screening, citation recommendation, and structured data extraction—reduce participation barriers and enrich the diversity of global research dialogue [11, 27, 89, 108].

Nonetheless, significant challenges persist in relation to bias, robustness in cross-lingual transfer, and equitable evaluation outcomes. Notably, AI outputs generated for non-English or code-switched content often display lower detection accuracy and elevated error rates, which can inadvertently perpetuate exclusion or misinterpretation [11, 27, 58, 89, 108]. These findings accentuate the continuing need for high-quality, domain-relevant datasets and the formulation of policy interventions directed toward technical and ethical inclusivity.

5.3 Citation and Evaluation Tools

The attainment of trustworthy AI-assisted academic writing is inextricably tied to precision in citation and evaluation. Advanced recommender systems—built upon embedding techniques, such as those used in SPECTER2, and retrieval-augmented generation (RAG) architectures—now facilitate scalable, context-aware citation suggestion and semantic bibliography contextualization [66]. Reliable and multi-dimensional assessment of these systems employs a suite of metrics: Precision@k, Mean Reciprocal Rank (MRR), ROUGE, and entailment-based measures.

Table 6: Key Evaluation Metrics for Citation Recommendation and Scholarly Text Generation

Metric	Description
Precision@k	Proportion of relevant recommendations within the top-k results.
Mean Reciprocal Rank (MRR)	Average of the reciprocal ranks of relevant items across queries.
ROUGE	Compares overlap between machine-generated and reference summaries (precision, recall, F-score).
Entailment-based Evaluation	Assesses whether generated statements are entailed by ground-truth citations or references.

These metrics, detailed in Table 6, support robust, multi-perspective analysis of both algorithmic and human performance. The most effective citation systems are trained on extensive, open bibliographic corpora and undergo rigorous validation—both quantitative and qualitative. Empirical evidence suggests that contemporary citation recommendation models typically rank correct references above distractors, while LLM-driven introduction generation, in combination with entailment verification, can yield scholarly text with factual and contextual accuracy rivaling expert-authored content [66]. Persistent challenges include citation bias, opacity in reference attribution, and risks of domain or linguistic overfitting. Consequently, there are calls for expanding reference pools to encompass multilingual, low-resource, and interdisciplinary domains, as well as implementing transparent, standardized evaluation protocols [66].

5.4 Generative Tools and Policies

The rapid proliferation of generative AI writing tools—exemplified by models such as ChatGPT—has sparked a major transformation in both individual scholarly practices and institutional policy landscapes [5, 58, 110]. These models capably handle language generation, summarization, feedback, and co-authoring, resulting in substantive improvements in efficiency, accessibility, and the customization of writing support. Nevertheless, their adoption has heightened scrutiny of issues including academic integrity, disclosure practices, and the broader ethics of generative AI [5, 110].

Policy analyses and empirical findings highlight a critical paradox: while AI proves highly effective for tasks such as grammar correction, summarization, and preliminary drafting, it remains irreplaceable for developing higher-order skills such as critical thinking, originality, and structured argumentation—competencies integral to advanced scholarly writing [5, 58, 110]. Institutional guidelines are thus evolving to balance utility with responsibility. Globally, universities are establishing nuanced frameworks that permit beneficial use (e.g., facilitating non-native speakers, reducing language barriers) while demarcating boundaries (e.g., requiring disclosure of AI use, restricting generative tool application in summative assessments, enforcing plagiarism prevention) [58, 110].

At the tool-design level, priorities now include transparency, explainability, and user education. Innovations in authorship attribution, watermarking, and adversarial evaluation are being actively explored to reinforce accountability. Broadly, the evidence to date supports responsible integration—rather than blanket restriction or unchecked adoption—as the strategy best aligned with the diverse and evolving needs of the academic community.

5.5 Synthesis

In summary, the convergence of automated and hybrid workflows, multilingual inclusivity, sophisticated citation and evaluation tools, and generative writing assistants is comprehensively restructuring the research and academic writing milieu. These technological advances promise substantial improvements in efficiency and greater inclusivity, but they also surface ongoing imperatives for oversight, equitable access, and ethical stewardship. The continued evolution of both technological frameworks and policy mechanisms remains essential to meet the diverse requirements of scholars, educators, and institutions [3–5, 11, 12, 14–16, 21, 26, 27, 39, 40, 45, 47–50, 52, 58, 65, 66, 69, 74, 89, 91, 93, 94, 100, 102, 105, 107–113].

5.6 Prompt Engineering, Model Optimization, and Specialized Agents

5.6.1 Prompt Design and Instability. Prompt engineering has rapidly become a foundational methodology for utilizing pretrained language models (LLMs) within various automated workflows. Despite its transformative potential, prompt-based automation faces critical challenges regarding prompt instability and reproducibility. Empirical studies reveal that manual prompt construction is inherently precarious: subtle modifications—such as the alteration of a single lexical item—can precipitate disproportionate declines in model performance across established Natural Language Understanding (NLU) benchmarks. This pronounced sensitivity complicates reliable deployment in real-world contexts, as practitioners frequently resort to iterative trial-and-error or exhaustive prompt search to achieve consistent outcomes. These phenomena are primarily attributable to the ad hoc nature of discrete, human-authored prompts, which may fail to robustly anchor the model's internal representations or adequately engage with its latent knowledge structures [70].

Recent methodological innovations seek to address these deficiencies by moving beyond fragile, hand-crafted prompts toward systematically optimized alternatives. A prominent development in this space is P-Tuning, which introduces trainable, continuous prompt embeddings. These embeddings are either concatenated to the discrete input tokens or instantiated as standalone vectors within the model's learned representation space. Such an approach effectively smooths the prompt landscape, mitigating sensitivity to specific wordings and facilitating more stable convergence during training. Empirical results on benchmarks, including LAMA and SuperGLUE, consistently demonstrate substantial gains in performance and robustness. Moreover, the advantages of P-Tuning generalize across both frozen and fine-tuned models as well as across varying supervision regimes, thereby providing a robust mitigation to prompt instability while preserving adaptability [70].

The comparative properties of discrete and continuous prompt engineering methodologies are concisely summarized in Table 7.

5.6.2 Model Adaptation and Specialization. The escalating scale and resource demands of contemporary LLMs highlight a crucial tension between achieving broad generality and domain-specific efficacy. While expansive, general-purpose models attain impressive aggregate performance, their practical utility within specialized sectors—such as law, finance, and scientific research—often depends on targeted adaptation and efficient deployment strategies. A proliferating research agenda investigates how lightweight, customized language models—trained from scratch using domain-relevant corpora and enhanced with optimized prompt-handling architectures—can effectively bridge this divide [110]. For instance, the development of bilingual, parameter-efficient models, such as MindLLMs (ranging from 1.3B to 3B parameters), demonstrates that judicious dataset curation and targeted instruction-tuning can significantly reduce computational overhead. Simultaneously, such approaches preserve, or even surpass, the domain-specific performance of much larger open-source comparators. These findings underscore that integrating continual, parameter-efficient prompt learning strategies—such as P-Tuning—with domain-specific optimization yields models that are both responsive to nuanced task requirements and practical for deployment in real-world scenarios [70][110].

A keystone of effective adaptation and rigorous evaluation for specialized LLMs lies in the adoption of standardized, transparent benchmarks and open evaluation protocols. As LLMs permeate high-stakes or sensitive domains, the importance of standardized metrics for accuracy, robustness, and societal considerations is magnified. Publicly maintained benchmarks facilitate comparative assessments of new architectures and foster measurable innovation by making performance differentials explicit and tractable. Concurrently, the implementation of structured metadata schemas for AI usage—encompassing tool identity, configuration parameters, usage environment, and the affected content segments—has been strongly advocated as essential for advancing transparent and reproducible science. Incorporating such metadata not only enables large-scale automated assessments of LLM adoption but also streamlines the construction of domain-specific corpora and supports the development of robust evaluation pipelines, especially in fields where algorithmic disclosure and traceability are imperative [20]. Together, these practices promote the responsible, effective, and transparent integration of prompt engineering, model optimization, and specialization strategies across disciplinary frontiers [110][20].

6 Quality Assurance, Feedback, and Oversight

6.1 AI and Agent-Based Quality Assurance

Recent advancements in artificial intelligence (AI) and agent-based methodologies have fundamentally transformed quality assurance (QA) paradigms across scientific and applied domains, particularly within complex, data-centric environments. Agent-based QA systems, leveraging autonomous or semi-autonomous modules, now provide scalable mechanisms for real-time monitoring, anomaly detection, and iterative process optimization. In water quality monitoring, for example, autonomous multi-agent fleets powered by local Gaussian processes and deep reinforcement learning have demonstrated a marked reduction in estimation errors versus traditional centralized methods, thereby enhancing both spatial coverage and responsiveness to environmental variability [107]. Analogous agentic paradigms have gained traction in fields ranging from modular robotics to decentralized on-demand transport, wherein localized sensing and action capacity foster a level of adaptive quality control unattainable through static procedural or equation-based approaches [108][47][11][56].

Despite the scalability and efficiency of AI-driven QA pipelines—particularly those underpinned by machine learning—algorithmic detection of problematic patterns or outliers remains insufficient when ambiguity or data sparsity is present. For instance, automated behavioral triage of survey interviews achieves performance approaching expert reliability in flagging high-risk segments and question types across multiple languages [47][11]. Nonetheless, these systems are prone to both false positives and context-dependent errors, especially when encountering nuanced or infrequently observed phenomena, thereby necessitating human expert review as an essential complement [107][4]. Hybrid frameworks have emerged in response: by routing clearly classifiable instances to computational algorithms and reserving ambiguous cases for peer adjudication, such systems have realized significant efficiency gains (enabling 54%-80% of open-ended responses to be coded automatically) while maintaining overall data validity [93][52][50][107][11].

The role of human peer review thus remains indispensable, particularly for tasks demanding methodological supervision, domain-specific discernment, or resolution of coder disagreement. Empirical analyses in both survey automation and academic writing workflows consistently reveal that expert intervention is critical for resolving ambiguity and validating methodological soundness [93][52][50][107][4][11][113]. Approaches such as double coding followed by expert adjudication maximize classification accuracy in the presence of disagreement or uncertainty, while single coding may be justified in contexts of low anticipated error to optimize resource utilization [52][50]. Furthermore, AI-driven feedback mechanisms—when integrated with established peer review practices in academic settings—facilitate a dual process: machine learning systems rapidly handle routine assessment tasks, while peer reviewers provide nuanced evaluations of methodology and argumentation, ensuring comprehensive formative and summative feedback [113].

A detailed comparison of agentic and traditional QA approaches reveals distinctive strengths and limitations. While automation confers unparalleled scalability and consistency, expert-driven peer review excels in accommodating context sensitivity, navigating ethical complexity, and promoting critical thinking [17][42][71][56]. For example, AI-assisted feedback in academic writing can efficiently identify linguistic or stylistic issues at scale but remains inadequate for evaluating originality, theoretical coherence, or upholding academic integrity [17][42]. Similarly, in broader system validation and oversight, optimal outcomes are achieved when agent-based models operate in concert with human expertise, thereby synergistically leveraging the efficiencies of automation and the discernment of expert judgment.

This relationship is summarized in Table 8, which contrasts the principal attributes of agent-based and traditional human-driven QA approaches.

Table 7: Comparison of Discrete (Manual) and Continuous (P-Tuning) Prompt Engineering Methods

Property	Discrete Prompts (Manual)	Continuous Prompts (P-Tuning)
Stability	Highly sensitive to minor changes; performance can vary greatly	Smooth response to perturbations; improved stability
Reproducibility	Often unreliable; requires repeated prompt search	High; stable convergence is more easily achieved
Design Effort	Labor-intensive; depends on domain and expertise	Largely automated via optimization
Generalizability	Typically low; transferability across tasks or domains is limited	High; generalizes across tasks, supervision, and models
Performance	Varies; brittle across benchmarks	Consistently strong; demonstrates gains on NLU tasks

Table 8: Comparison of Agent-Based and Human-Driven Quality Assurance Approaches

Dimension	Agent-Based QA	Human Peer/Expert QA
Scalability	High (enables large-scale, real-time coverage)	Limited (resource- and time-intensive)
Consistency	Algorithmic, objective	Variable, subjective
Context Sensitivity	Reasonable for well-specified use cases; limited adaptability to nuance	High; expert judgment handles ambiguity and novel cases
Ethical Considerations	Relies on pre-specified rules and data; lacks autonomous ethical discernment	Capable of ethical reasoning and integrity assessment
Critical Thinking	Constrained by algorithmic context	Supports critical and creative evaluation
Efficiency in Routine Tasks	Excellent for pattern recognition and triage	Less efficient, better suited to complex tasks
Handling Ambiguity	Limited; ambiguity often escalated to human review	Strength in resolving ambiguity and disagreement

Consequently, contemporary quality assurance is characterized by hybridization: machine learning or agent-driven systems furnish rapid formative feedback, while human reviewers exercise summative oversight, arbitrating ambiguous cases, validating methodological soundness, and remediating algorithmic failures [93][52][50][107][4][11][113]. This synergistic approach not only accelerates quality monitoring but also establishes a robust foundation for continual improvement, transparency, and adaptability within data-intensive disciplines.

6.2 Data Quality in Survey Automation

Automated surveys, especially those administered via messaging platforms, have revolutionized large-scale data collection by offering unparalleled reach, cost efficiency, and rapid deployment. Such systems—exemplified by WhatsApp-driven survey platforms—demonstrate improved completion rates and lower costs per respondent, especially among mobile and transient populations. Notwithstanding these advantages, new challenges emerge around technical reliability, privacy adaptation, and the risk of response bias [28].

Ensuring high data quality in automated survey contexts requires a dual focus: mitigating sources of technical failure and participant attrition, and actively monitoring for threats to data integrity, participant engagement, and bias. Dynamic adaptation of survey logistics to domain-specific attributes—such as participant language, literacy levels, and privacy sensitivities—proves essential for fostering respondent trust and obtaining valid data [28]. To counteract attrition, strategies including intelligent reminders, engaging user interfaces, and adaptive branching logic are employed. In parallel, ongoing bias mitigation remains crucial: this involves frequent calibration of models to promote equitable performance across demographic subgroups, as well as transparent accommodation of language and modality translations. A continuous QA process, supported by AI-based tools for automated anomaly detection and complemented by human peer review for the adjudication of nuanced events, forms the backbone of best-practice in automated survey deployment.

Together, these advances underscore the imperative for dynamic, hybrid quality assurance frameworks in AI-augmented research workflows, where the complementary strengths of agentic automation and expert human oversight are actively integrated to uphold scientific rigor and ethical responsibility [4, 11, 17, 28, 42, 47, 50, 52, 56, 71, 93, 107, 108, 113].

7 Ethics, Integrity, Transparency, and Regulation

7.1 Ethics and Academic Integrity

The widespread adoption of artificial intelligence (AI) in research and education has introduced complex ethical challenges that demand urgent attention regarding academic integrity, transparency, and equitable practice. While AI-powered tools have significant potential to automate knowledge synthesis and enhance learning outcomes, they also create new challenges—ranging from plagiarism and misinformation to unreliable outputs and sophisticated forms of research fraud. Such developments raise fundamental concerns for all stakeholders in the academic community [4, 5, 8, 10–15, 18, 24, 38–40, 45, 47, 50–54, 57–59, 67, 72, 73, 79, 82, 84–86, 89, 91, 97, 98, 101, 104, 105, 107, 108, 114].

Risks and Threats. AI amplifies not only longstanding issues such as plagiarism—now exacerbated by advanced large language models (LLMs) capable of producing human-like text [8, 47, 82, 84, 85, 89, 104, 107]—but also introduces new risks including factual inaccuracies, hallucinations, biased or discriminatory outputs, and elaborate fraud schemes encompassing ghostwriting, data fabrication, and undetectable falsification [5, 12, 13, 18, 24, 39, 45, 50, 59, 67, 91, 97, 101, 108, 114]. As LLM-generated text becomes virtually indistinguishable from authentic human writing, both identification of academic misconduct and validation of information become increasingly challenging [12, 13, 47, 58, 73, 104, 107]. The proliferation of AI-powered misinformation, especially when propagated through literature reviews, publication channels, and social media, poses an acute threat to scholarly rigor and public trust [11, 38, 39, 72, 73, 86].

Detection Mechanisms and Multimodality. To address these risks, a comprehensive and multimodal set of detection strategies is required. Manual expert review remains crucial for sophisticated judgment and contextual assessment [47, 57, 79, 84, 91]. However, the growing scale and subtlety of AI-enabled misconduct have surpassed the capacity of human reviewers alone. Accordingly, a range of machine learning-driven detection tools have gained prominence—including watermarking [8, 47, 50, 91, 105, 107], stylometric and statistical analyses [11, 13, 59, 73, 82], and advanced ensemble classifiers with anomaly detection capabilities [4, 5, 10, 11, 13, 38, 58, 82]. These systems increasingly leverage hybrid and multi-factorial methodologies to better counteract adversarial evasion, linguistic diversity, and mixed-authorship scenarios [40, 47, 59, 73, 82]. Yet, empirical analyses indicate that current detection methods are imperfect: their efficacy diminishes as models grow more sophisticated, and with the introduction of paraphrased or multimodal data [4, 12, 13, 40, 47, 73, 82]. This challenge underscores the importance of continuously refining detection benchmarks and employing robust cross-validation protocols [11, 13, 38, 40, 73, 107].

Crucially, the global nature of scholarly communication necessitates that equity and fairness be systematically embedded in detection frameworks. For example, insufficiently tuned detectors may exhibit bias against non-native English writers, thereby perpetuating linguistic and cultural inequities [11, 13, 18, 38, 40, 47, 82]. The literature advocates for AI-driven evaluative systems that are transparent, explainable, and adaptable, so as not to exacerbate existing structural disparities [12, 38, 45, 84, 85, 94].

As highlighted in Table 9, each core detection strategy offers unique advantages and limitations. This diversity necessitates a hybrid and continually evolving approach for effective AI-authorship identification.

Emerging Threats and Countermeasures. Generative AI continues to introduce threats that outstrip traditional academic boundaries. The automated generation of convincingly fabricated research articles—including spurious citations and falsified data—poses pronounced risks to the reliability of the scientific record [11, 39, 47, 82, 84, 107, 108]. Addressing these risks requires both technical interventions (such as AI-authorship watermarking, enhanced metadata standards, and robust traceability protocols) [8, 11, 40, 47, 50, 91] and institutional reforms (updated publisher and funding guidelines, strengthened IRB practices, and inter-institutional intelligence sharing) [38, 58, 59, 67, 82, 97, 101]. Additionally, the rise of highly automated, self-modifying agentic systems introduces novel questions about accountability and potential harms [20, 26, 31, 39, 67, 82, 100, 101]. To ensure ethical operation and oversight, system designs must support contestability, transparency, and human interruptibility as foundational principles [31, 58, 82, 86].

7.2 Regulation and Standardization

The evolving landscape of AI-assisted research, publishing, and education underscores the need for harmonized regulation and standardization. As various AI-driven and agentic workflows proliferate, there is a growing imperative for unified guidelines that ensure transparent and responsible AI adoption across the academic spectrum [4, 6, 11, 12, 14, 16, 19–21, 25, 26, 31, 44, 52, 67–69, 81, 84, 88, 89, 91, 93, 100, 102, 105, 112].

Metadata, Traceability, and Publisher Guidelines. One persistent challenge is the lack of widely adopted, standardized metadata that describes AI usage throughout the research process [20, 23, 40, 82]. The systematic integration of structured metadata—specifying tool names, versions, purposes, relevant manuscript sections, and parameterization—into scholarly workflows would enable enhanced analysis, greater transparency, and more reliable detection of AI-generated content [20, 23, 82]. The embedding of such metadata at both the author submission and editorial review stages is crucial for facilitating downstream research on the linguistic, ethical, and disciplinary impacts of AI-assisted writing [20, 23, 40, 82].

Watermarking, Detection, and Transparency Techniques. Technically, watermarking methods remain promising for document-level traceability. However, approaches such as white-box, black-box, and neural watermarking present significant trade-offs in terms of robustness, interpretability, and vulnerability to adversarial attacks or paraphrasing [8, 11, 40, 47, 50, 91]. To improve overall transparency and fairness—especially in high-stakes or regulated environments—ensemble and hybrid detection pipelines are recommended, integrating statistical, behavioral, and metadata-based signals. Open benchmarking against advancing threat models and model architectures is essential for credible, independent system validation [11, 12, 20, 21, 25, 73, 84, 91].

Unified Reporting and Regulatory Calls. Due to global variations in institutional and publisher policies, calls for unified reporting standards and harmonized regulatory frameworks have intensified [16, 19–21, 26, 44, 52, 67, 68, 81, 88, 89, 93, 102, 105, 112]. Existing international guidelines (e.g., EU Ethical AI, OECD, IEEE) articulate foundational principles—such as accountability, explicability, justice, and beneficence—yet their translation into operational standards remains inconsistent [4, 14, 16, 19, 20, 26, 67, 69, 89, 91, 93, 102, 105]. Robust systems for impact assessment, independent audits, and ongoing monitoring must therefore be embedded within organizational and academic workflows [4, 11, 12, 16, 20, 21, 31, 40, 52, 59, 68, 69, 73, 82, 88, 107].

Living review systems and open science initiatives are further highlighted as essential complements to conventional policy infrastructures, supporting rapid, evidence-driven responses to emergent ethical issues and sustaining trust in AI-augmented research practices [18, 20, 23, 31, 40, 58, 59, 68, 82, 91].

Open Challenges and Future Directions. Despite notable advancements in technical and policy safeguards, important gaps remain: no single detection or regulatory framework currently affords comprehensive or future-proof protection against the array of evolving threats introduced by AI [13, 47, 73, 82, 104, 107]. As a result, the field must pursue adaptive, multi-layered strategies—encompassing multimodal detection, rigorous metadata governance, harmonized international standards, and ongoing interdisciplinary collaboration [8, 12, 18, 20, 21, 23, 31, 38, 58, 67, 82, 85, 105]. Only through such sustained, coordinated effort can the unprecedented opportunities of AI in research, education, and society be realized, while at the same time responsibly mitigating its attendant risks.

Table 9: Core Detection Approaches for AI-Generated Academic Content

Method	Description	Key Limitations	
Manual Expert Review	Contextual judgment by domain specialists	Scalability, subjectivity	
Watermarking (white/black box)	Hidden markers embedded in AI output for traceability	Evasion, robustness, explainability	
Stylometric/Statistical Analysis	Quantitative features (e.g., word frequency, syntax patterns) used for source attribution	Vulnerable to paraphrasing, false positives	
Anomaly	Ensemble Classifiers	Combination of behavioral, linguistic, and metadata inputs for detection	Model drift, adversarial adaptation

8 Cross-Domain Applications and Educational Impact

8.1 Cross-Sector Applications

The integration of agentic and artificial intelligence (AI) systems across diverse sectors has catalyzed a paradigm shift in the conduct of knowledge-intensive work, notably within science, environmental management, healthcare, law, commerce, and education. Central to this transformation is the deployment of AI-enabled solutions facilitating tasks such as automated literature survey generation, systematic review, text and video analysis, real-time monitoring, and the development of domain-specific resources [4, 5, 9, 11–13, 15–17, 24, 29, 33, 35, 37, 40, 42, 43, 45, 47, 50, 54, 56, 58, 60, 67, 71, 76, 82, 91, 93, 102, 103, 105, 107, 108, 110, 112, 114].

The automation of scholarly workflows—exemplified by techniques such as AutoSurvey and LLM-powered systematic review platforms—addresses the accelerating volume of research outputs, providing scalable mechanisms for knowledge synthesis and curation. Deep learning and natural language processing tools have notably streamlined labor-intensive processes such as record screening and open-ended text categorization. These technologies yield significant reductions in manual effort and demonstrate high sensitivity, yet non-trivial challenges persist, including citation misalignment, incomplete automation, and the potential exclusion of relevant information [4, 9, 11, 12, 15, 16, 24, 29, 33, 40, 43, 45, 47, 50, 60, 82, 91, 93, 102, 103, 107, 108, 110, 112]. Such limitations underscore the necessity of robust human oversight and the ongoing advancement of more rigorous benchmarks and evaluation metrics [16, 24, 40, 103, 110, 112].

Urban management represents a particularly salient domain for agentic systems. The deployment of Internet of Things (IoT)-linked sensors and multi-agent frameworks is fundamentally reshaping resource orchestration. For instance, smart parking systems utilizing multi-agent architectures leverage real-time IoT sensor data to optimize urban parking availability, mitigate congestion, and advance sustainable city environments [96]. These implementations both minimize resource waste and exemplify the broader potential for agentic orchestration in urban infrastructure. Notwithstanding these successes, challenges concerning scalability, integration into real-world environments, and the heterogeneity of data and actors remain prominent [42, 71, 76, 96, 114]. Agent-based and multi-agent paradigms have demonstrated robust decision-making, adaptability, and enhanced performance under uncertainty in areas such as transportation, hydrological monitoring, and logistics [24, 42, 50, 71, 76, 107, 108, 114]. However, real-world deployment frequently encounters barriers related to standardization, ethical compliance, and distributed coordination, necessitating development of advanced frameworks integrating automated reasoning with human-in-the-loop oversight [4, 15, 24, 33, 42, 50, 56, 71, 82, 107, 108, 112, 114].

The application of agentic systems in the social and health sciences has yielded substantial improvements in the analysis and monitoring of large-scale surveys and interviews. Machine learning pipelines embedded in Computer-Assisted Recorded Interviewing (CARI) tools can swiftly process multilingual audio data, flag problematic survey items, and detect interviewer effects with performance approaching that of human experts while offering significantly heightened efficiency [11, 16, 17, 40, 47, 60, 102, 103]. Hybrid human-AI workflows further optimize open-ended survey coding by algorithmically triaging cases for automation versus those requiring expert involvement [60, 102]. This synergy delivers process efficiency while preserving methodological rigor, though ongoing challenges include modeling rare responses and achieving coder consensus in ambiguous cases [16, 47, 60].

Furthermore, the increasing integration of agentic approaches in commerce and legal domains has accelerated data classification and information retrieval, as seen in economic census NLP applications [43]. These advancements concurrently heighten the salience of transparency, bias, and compliance considerations. The embedding of ethical and regulatory modules within agentic systems has become increasingly necessary, with recent progress integrating compliance and ethical reasoning alongside memory, reasoning, and autonomous tool execution [5, 76, 114]. Notwithstanding these innovations, persistent challenges remain regarding AI bias, opacity of system outputs, and the need for robust mechanisms to ensure auditability and recourse [5, 17, 42, 56, 58, 71].

As summarized in Table 10, while agentic AI systems have generated substantial benefits across diverse domains, each application area faces persistent technical, ethical, and organizational challenges that underscore the need for ongoing innovation and robust oversight.

8.2 Educational Impact

The educational domain stands at the forefront of both the transformative promise and complex tensions engendered by agentic and AI automation. These technologies have introduced multi-faceted impacts at the levels of curriculum development, educational policy, and research competency building [5, 17, 42, 56, 58, 71, 76]. AI-driven tutoring systems, multimodal learning assistants, and open student models (OSM) are now central to the personalization of learning and formative assessment, leveraging data analytics to dynamically tailor instruction and foster self-regulated learning. Large language models (LLMs), such as Gemini and ChatGPT, have proven particularly effective for generating individualized feedback, differentiating instruction, and facilitating language learning [5, 58, 76]. Empirical evidence affirms that adaptive OSM platforms, when integrated within smart classroom contexts and coupled with actionable, visualized feedback, yield tangible improvements in both student engagement and knowledge retention [58, 76].

Table 10: Representative Cross-Sector Agentic AI Applications and Key Challenges

Sector/Domain	Exemplar Agentic AI Use Cases	Demonstrated Benefits	Persistent Challenges
Scholarly Publishing	Automated literature surveys, systematic review automation	Scalability, time savings, improved sensitivity	Citation misalignment, incomplete automation, risk of relevant information omission
Urban Management	Smart parking, resource orchestration with IoT-enabled agents	Efficiency, congestion reduction, sustainability	Scalability, integration, data/actor heterogeneity
Social/Health Sciences	Automated audio/text survey analysis, CARI, hybrid coding	High throughput, near-expert performance, process efficiency	Modeling rare responses, coder consensus, explainability
Commerce and Law	NLP-enabled document sorting, compliance checking	Speed, accuracy of classification/retrieval, regulatory compliance	Bias, transparency, auditability, ethical integration

However, the expanded deployment of AI in educational settings simultaneously introduces significant challenges, notably in upholding academic integrity, promoting equity, and preserving essential skills such as critical thinking [5, 17, 42, 56, 58, 71]. While AI-based writing assistance platforms deliver support ranging from grammar correction to personalized research feedback, concerns persist regarding overreliance, diminished originality and argumentation, and the exacerbation of ethical dilemmas like plagiarism [5, 17, 42, 56, 58, 71]. Institutional responses to these issues are highly heterogeneous: some, such as Stanford University, have updated academic integrity policies, whereas others, such as Middlebury College, have chosen to prohibit AI tool usage in classroom environments, reflecting the ongoing debate concerning appropriate integration of AI in academic curricula [5, 42, 56, 58, 71].

Despite these complexities, the prevailing view in the literature is that AI cannot and should not replace the foundational pedagogical value of traditional writing and research training. Instead, a balanced paradigm is advocated, wherein AI is leveraged as an assistive resource while safeguarding essential competencies in critical thinking, creativity, and ethical practice [5, 42, 56, 58, 71]. Ensuring the positive impact of AI on education thus rests upon sound policy development, sustained human oversight, and vigilant attention to issues of transparency, bias, accessibility, and equity [42, 56, 58, 71, 76].

8.3 Integrated Workflows

The ongoing convergence of automation paradigms is driving the emergence of next-generation academic and research workflows, marked by comprehensive integration of survey automation, agent-based video and text recognition, and sophisticated orchestration of scholarly tasks [9, 16, 24, 40, 42, 43, 58, 60, 76, 91, 103, 110, 112]. Hierarchical frameworks, which capitalize on the cognitive capabilities of domain-specialized AI agents, facilitate rapid synthesis of literature, dynamic assessment of evidence quality, and coordinated analysis over multimodal data streams.

Despite these substantial advances, several enduring needs remain, including the enhancement of explainability, improved domain adaptation, collaborative human-AI evaluation paradigms, and continuous validation in light of evolving standards for scientific rigor and ethical practice [24, 40, 42, 56, 58, 71, 76, 91, 103, 110, 112]. The trajectory of cross-domain agentic systems in science, education, and beyond will ultimately be determined by the intersection of technical innovation, sound architectural design, and interdisciplinary cooperation grounded in alignment with societal and ethical imperatives.

9 Explainability, Human-Centric AI, and Inclusive Systems

9.1 Explainability and Transparency

Recent advancements in artificial intelligence (AI)—particularly in large language models (LLMs) and survey automation—have significantly expanded the reach and utility of these systems, while simultaneously intensifying prevailing concerns regarding explainability and transparency. The intrinsic opacity of deep neural architectures, which underpins notable progress in fields such as natural language processing (NLP), medical diagnostics, and literature review automation, restricts users' capacity to interpret, audit, or contest outputs in high-stakes contexts [3, 10, 12, 20, 26, 39, 45, 48, 49, 51, 69, 70, 79, 81, 86, 88, 94]. To respond to these challenges, contemporary research adopts a multifaceted approach that integrates technical interpretability, rigorous evaluation, and user-centered explanation.

One key area involves the rationalization of model outputs through the expression of predictions in natural language rationales accessible to non-experts. Extractive rationalization methods, which identify salient input segments as justifications, facilitate transparency and reliability, whereas abstractive approaches seek to generate flexible, user-facing explanations [10, 51, 86, 88]. Surveys of recent work demonstrate that extractive techniques, while transparent, often necessitate substantial human annotation. Conversely, abstractive rationalizations may be susceptible to hallucination or unfaithful justification, especially in the absence of robust, end-to-end evaluation protocols [10, 20, 39, 86]. This underscores the importance of matching explanation modalities to the application context, carefully balancing the demands of fidelity and user accessibility.

The evaluation of explainability now increasingly employs standardized frameworks and metrics. Tools supporting rationale annotation, together with automatic metrics—including ROUGE, BLEU, METEOR, and BERTScore—provide systematic means for comparing generated explanations against reference summaries and human judgments [3, 10, 12, 26, 49, 69, 70, 79]. Novel divergence-based measures, such as Mauve, have been introduced to align more closely with human judgments and reveal disparities between system and reference outputs [81]. However, methodological challenges endure: many current benchmarks lack linguistic and genre diversity, which undermines their representativeness, and excessive dependence on automated metrics risks reinforcing superficial or misleading explanations [20, 48]. As summarized in Table 11, the field remains dynamic, with interpretability advancing particularly in high-stakes, multilingual, or cross-domain tasks [20, 39, 48, 79, 88].

Transparency is further promoted through the adoption of explicit explanation protocols within automated academic workflows. LLM-powered literature survey tools, for instance, utilize hierarchical outline-driven synthesis and iterative refinement, explicitly

Table 11: Overview of major explainability evaluation methods and their key characteristics

Method	Type	Key Advantages	Limitations
ROUGE/BLEU/METEOR	Automatic metric	Quantitative, reproducible	Surface-level, limited semantics
BERTScore	Embedding-based	Semantic similarity tracking	Sensitive to pretrained models
Mauve	Divergence-based	Closer to human judgment	Requires reference distributions
Manual Annotation	Human evaluation	Rich qualitative insight	Expensive, low scalability

tracking citations, coverage, and content relevance using multimetric evaluation frameworks [12, 69, 79, 94]. In clinical review automation, LLMs are increasingly designed to generate both decision rationales and, upon request, transparent model revisions, thereby fostering user trust and supporting expert human review, though not replacing it [12, 70, 94]. Ongoing challenges such as citation misalignment, insufficient benchmarking, and inadequate user-facing outputs emphasize the continued necessity for community-driven standardization and comprehensive human evaluation [10, 12, 20, 39, 69, 86].

9.2 Equity, Bias, and Fairness

As automated and AI-augmented systems become embedded in survey research and academic workflows, critical questions of equity, algorithmic bias, and inclusivity have gained prominence. The literature reveals both the democratizing potential and the notable risks of machine learning in addressing the needs of diverse and historically marginalized communities, particularly with respect to underrepresented languages, marginalized user groups, and the construction of inclusive benchmarks [84][97][54][67][105][53][13][72][57][104][39][114][38][24][14][8][86][51][101][18][45][12][85][79][10][59][73][52][89][50][

Algorithmic bias, originating from both training data and model architecture, remains an enduring concern. In automated survey analysis and open-ended response classification, models predominantly trained on English or high-resource languages often exhibit substantial performance degradation—or even systematic bias—when applied to low-resource or non-English data [84][67][53][39][86][51][18][85][10][82][47][11]. Research on models for Amharic news summarization or bilingual survey evaluation demonstrates that targeted fine-tuning and rule-based postprocessing can yield notable improvements for low-resource languages. Nevertheless, these gains do not overcome persistent obstacles, namely data scarcity, inconsistent annotation, and difficulties in cross-linguistic generalization [53][85][47][11]. The literature thus calls for not only technical adaptation, but also the development of datasets that more equitably capture linguistic and demographic diversity [39][79][10][47][11].

Bias detection in text generation and classification, especially concerning AI-generated text (AIGT) and authorship attribution, remains contested. Watermarking, statistical, and ensemble classifier techniques have demonstrated some efficacy in distinguishing AI-generated outputs. However, their reliability can be undermined by model scale, adversarial tactics, and the inadvertent marginalization or misclassification of non-native language users [72][114][38][101][89][108][4]. The literature consequently recommends fairness-aware, explainable detection protocols and the construction of equitable evaluation suites spanning language, genre, and demographic variables [114][18][89][50][47][11]. Table 12 synthesizes several prevalent bias detection methods and their evaluation considerations.

Equity also fundamentally involves privacy and ethical stewardship of sensitive data. Automated survey analysis tools often process personally identifiable or medical information, which heightens the stakes concerning privacy breaches, misuse, and inadvertent harm [54][105][13][24][8][12][79][52][40]. Multimodal platforms, leveraging modalities such as audio, geolocation, and behavioral traces, further amplify concerns around user consent, data minimization, and the differentiated impact of surveillance on vulnerable groups [54][67][13][57][104][8][59][73][52][89][50][40]. Technical safeguards such as privacy-preserving computation, transparent anonymization, and data governance must therefore be coupled with regulatory compliance and active stakeholder engagement [13][39][38][89][50][40][27][20].

Inclusive benchmarking and contestability are advancing as fundamental principles. Increasingly, models are evaluated not solely on canonical datasets, but also across resource-limited, domain-diverse, and demographically varied corpora, thereby surfacing disparities and reducing the risk of systematic exclusion [97][105][53][14][8][101][18]. Achieving equitable AI goes beyond technical solutions; it requires ongoing interdisciplinary collaboration, community-driven standards, and iterative, impact-driven improvement [39][86][18][45][12][10][50][108].

9.3 Human-Centered and Contestable Systems

A prevailing theme in contemporary research is the imperative to design AI-powered survey and academic automation systems that are not only transparent and equitable but are also fundamentally human-centered and contestable. Contestability is conceptualized as the capacity for operators, stakeholders, and end-users to interrogate, challenge, and, where warranted, override or correct algorithmic outputs. This safeguard is particularly critical in environments demanding rapid, high-stakes decisions [39][45][12][10][73][107][15][4][40][27][2

Technical progress includes embedding explainability and uncertainty quantification directly into automated survey systems and developing interfaces that promote collaborative, human-in-the-loop review and annotation [39][12][10][73][40][27][20]. For instance, systems that display ranked explanations, highlight low-confidence cases for human attention, and maintain transparent audit logs of decision-making processes promote both accountability and continuous learning [45][12][73][40][27]. Automated literature review and agentic academic workflow platforms are increasingly equipped with audit and contestability features, such as revision tracking, explicit operator feedback, and transparent explanatory logic [45][12][94][70][27].

Table 12: Common bias detection approaches in AI-generated text and classifications

Approach	Technique	Strengths	Limitations
Watermarking	Output token patterns	Robust to simple evasion	Adversarially brittle
Statistical Classifiers	Model comparison	Broad applicability	Sensitive to shift
Ensemble Detection	Multi-model voting	Improved robustness	Complexity, ambiguity
Fairness-Aware Evaluation	Demographic testing	Surfaces group-specific issues	Requires diverse datasets

From a methodological perspective, contestability is advanced by design frameworks like "Design for Defeaters," which articulate both direct and indirect mechanisms through which users can question and contest model outputs or their underlying justifications [73][107][15]. Case studies spanning domains—including survey research and medical triage—demonstrate that absent robust contestability structures, authority and public trust deteriorate, accountability gaps widen, and opportunities for timely error correction are diminished [39][12][73][107][15].

Community-driven strategies, rooted in stakeholder engagement, iterative co-design, and open-source tool development, further empower users and democratize oversight [86][18][45][12][40][27][20]. The literature stresses the importance of transparent documentation, open evaluation protocols, and shared data resources, in tandem with continual monitoring and field testing post-deployment to detect emergent behaviors and unintended effects [45][12][73][40][27]. The integration of perspectives spanning technical, regulatory, and user communities is consistently highlighted as essential to the sustainable, trustworthy automation of academic and survey processes [39][45][10][50][108][103][40][20].

In summary, the current frontier in explainability, equity, and human-centric design for automated survey and academic systems is defined by the intricate interplay of technical, methodological, and ethical innovations. Ensuring that these systems are interpretable, inclusive, and contestable is not merely an aspirational technical goal, but a pressing socio-technical imperative—one that requires sustained, collaborative, and principled engagement across disciplines and communities.

10 Standardization, Interoperability, and Collaborative Benchmarking

10.1 Protocols and Systems Standards

The accelerating scope of academic and survey automation necessitates rigorous standardization and interoperability at both technical and methodological levels. Persistent fragmentation—arising from disjointed tools, disparate datasets, and heterogeneous evaluation protocols—continues to impede reproducibility, scalability, and the comparability of research outcomes across studies and domains. This issue is particularly pronounced in systematic literature review (SLR) automation, where critical steps such as study selection, data extraction, and synthesis vary markedly in both their implementation and reporting practices [4, 6, 11, 12, 16, 19, 20, 26, 31, 52, 68, 69, 81, 88, 89, 91, 93, 100].

Recent efforts to enhance harmonization have underscored the urgent need for standardized reporting frameworks and interoperable toolchains. The absence of universally accepted benchmarks and consistent terminologies across systematic review (SR) automation methods has complicated the direct comparison and objective assessment of different systems in realistic research settings [20, 88, 91, 100]. Many software solutions focus narrowly on isolated segments of the SLR pipeline—such as the screening stage—without supporting seamless integration or transition between components. This compartmentalized approach has received critical attention, as it restricts potential savings in time and labour and complicates the integration of automated systems with both manual processes and other technologies [12, 16, 26, 100].

A comprehensive analysis of over 20 prominent web-based SR tools demonstrates that, while capabilities for collaboration and screening are relatively mature, automation in subsequent stages—including data extraction and synthesis—remains underdeveloped. Furthermore, there is limited adherence to unified protocols governing feature support and reporting [6, 16, 69, 81, 88]. These gaps emphasize the need for systematic frameworks that enable robust, interoperable automation throughout the complete SLR workflow.

A parallel challenge arises in the context of AI-powered question generation and automated assessment systems: the lack of reproducibility standards and transparent documentation hampers broader adoption and critical appraisal. The evolution of large language model (LLM)-based and semantic similarity-driven approaches is often constrained by divergent dataset formats and inconsistent evaluation metrics. As a result, many solutions function as 'local maxima'—achieving strong performance in narrow contexts, but lacking the generalizability required for robust deployment in educational or survey automation settings [25, 44, 67, 84, 105]. Moreover, insufficient documentation concerning provenance, parameterization, and workflow design further impedes critical evaluation and independent validation [4, 11, 31, 67, 68, 112].

To address these limitations, several initiatives within the AI for SLR community—as well as in related fields such as economic and official statistics—advocate for the explicit adoption of protocol-driven and modularized systems standards as prerequisites for credible automation [14, 31, 52, 68, 100, 112]. These recommendations encompass, for example, the use of standardized metadata fields (including AI tool name, version, application context, and usage parameters), thereby facilitating transparency and enabling meta-analyses of tool efficacy and linguistic impact, particularly as generative AI tools proliferate within academic workflows [11]. Best-practice guides and reporting taxonomies—exemplified by the PRISMA and AMSTAR-2 frameworks for evidence synthesis—are increasingly promoted. These distinguish between methodological and reporting standards, underscoring that checklist adherence should be informed by substantive methodological rigor [20, 68, 88, 91].

Table 13 summarizes key gaps in current systems standards, mapping them to their implications for SLR and survey automation.

In summary, while significant advances have facilitated the development of modular and partially automatable pipelines for survey data collection and SLRs, persistent deficiencies in reproducibility, interoperability, and harmonization protocols remain. These deficiencies reaffirm the necessity of ongoing community-wide efforts to unify protocols, standardize reporting formats, and enhance tool interoperability—objectives that are foundational to the continued development of transparent, reliable, and scalable academic and survey automation [4, 6, 11, 12, 16, 19, 20, 25, 31, 52, 67–69, 81, 84, 88, 89, 91, 93, 100, 105].

10.2 Open Datasets and Collaborative Practices

Central to resolving the challenges of fragmented automation ecosystems is the sustained development of open datasets and the institutionalization of collaborative benchmarking practices. The trajectory of recent machine learning advances—particularly in domains such as question generation, automated answer assessment, and literature review synthesis—has been intimately tied to the availability of large, high-quality, and openly accessible datasets [6, 11, 12, 20, 21, 40, 43, 46, 59–61, 74, 77, 81, 94, 110]. Datasets like SQuAD, MS MARCO, RACE, and specialized resources such as SciReviewGen and BigSurvey have provided crucial shared foundations for reproducible experimentation and collaborative algorithm development [11, 21, 25, 59, 77].

Despite these successes, systematic analyses reveal that publicly available datasets remain disproportionately concentrated in specific fields, particularly biomedicine and computer science. Other critical domains—such as multimedia question generation, cross-disciplinary SLR automation, or large-scale survey automation—are comparatively underrepresented [40, 43, 46, 60, 74]. The limited breadth and diversity of datasets thus constrain the generalizability of existing systems and hinder the transferability of learned models to novel or underserved research areas [21, 40, 46, 59, 61, 110]. As a result, there is an intensifying call for the creation of cross-domain, multilingual, and multimodal dataset initiatives, in concert with open-source code and workflow sharing, to enable more comprehensive evaluation and accelerate the translation of innovative methods across diverse fields [20, 61, 77, 94].

Collaborative benchmarking serves as a pivotal mechanism to both advance technical progress and ensure transparency. Systematic, community-driven comparison of tools and algorithms—using common, curated benchmarks—enables nuanced understanding of strengths, weaknesses, and context-dependent performance variabilities [6, 12, 20, 21, 46, 77, 94, 110]. Current best practices recommend multi-dimensional evaluation suites (including citation recall/precision, coverage of relevant content, and standardized metrics such as ROUGE or Mauve for generative tasks) as well as the use of community-led challenge platforms to support comprehensive comparison. Such structures are essential not only for fostering innovation but also for addressing emerging ethical, fairness, and explainability requirements associated with AI-powered literature synthesis and assessment [11, 12, 20, 77, 110].

Of particular significance, major consortia and infrastructure efforts are increasingly promoting the systematization of "living reviews"—continuously updated evaluations—and the seamless integration of shared APIs and harmonized metadata. These initiatives are designed to ensure the persistence and relevance of benchmarks even as fields rapidly evolve, while also facilitating ongoing evaluation of new tools against state-of-the-art baselines [6, 20, 74, 81].

Table 14 presents a condensed comparison of prominent challenges faced by the open dataset and benchmarking ecosystem, together with emerging collaborative responses.

As the complexity and heterogeneity of academic content continue to increase, the future of robust, trustworthy, and reproducible automation in survey and literature review tasks will depend decisively on the ongoing growth and collective stewardship of open, collaborative, and cross-domain resources. These foundational efforts remain vital to sustaining scientific integrity, enabling transparent comparisons, and ensuring that new methodologies achieve meaningful impact across disciplines [6, 11, 12, 20, 21, 40, 43, 46, 59–61, 74, 77, 81, 94, 110].

11 Limitations, Challenges, and Future Prospects

11.1 Barriers and Open Challenges

Despite rapid progress in the automation of systematic reviews and academic workflows, substantive barriers remain across technical, methodological, and ethical domains. Chief among these challenges is the **variable quality and accessibility of research data**. Limitations such as insufficient metadata, incomplete or domain-biased corpora, and restricted access due to paywalls or non-standardized sources continue to impede both the efficacy and generalizability of automation efforts [1, 4–6, 10–12, 15, 17–20, 22, 27, 28, 31, 32, 38, 40, 42, 45, 47, 50–52, 56, 58, 59, 63, 68–71, 73, 76, 78–82, 85, 86, 88, 89, 91–93, 98, 99, 103, 106–108, 110, 113, 115]. Many advanced AI tools, particularly large language models (LLMs) and domain-specific classifiers, are trained or validated on narrowly scoped datasets, raising concerns regarding replicability and their capacity for cross-domain integration. This problem is especially pronounced in survey automation, where **dataset scarcity, attrition, and technical dependency on proprietary APIs** persist as substantial obstacles [28].

In addition to data-centric issues, **workflow opacity and lack of interpretability** constitute enduring limitations. While transformer-based and other neural architectures have driven significant breakthroughs, their internal complexity often hinders the transparency of decision-making processes, particularly for non-expert end-users. This opacity impairs adoption within critically scrutinized fields such as medicine, social sciences, and official statistics [15, 17, 18, 27, 28, 31, 50, 58, 59, 69, 71, 82, 86, 89, 91, 93, 107, 108, 110, 113]. The lack of accessible, open evaluation benchmarks and user-friendly interfaces further exacerbates the difficulty for practitioners who need to trust and meaningfully interact with automated outputs [4–6, 12, 17–19, 27, 40, 42, 47, 50, 52, 58, 70, 76, 81, 82, 86, 88, 89, 91, 93, 98, 107, 108, 110, 113].

A further significant constraint is imposed by **computational demands and scalability limitations**. Deep learning and graph-based methods, while offering considerable power, often require substantial compute resources during both training and inference phases [5, 6, 12, 15, 18, 19, 31, 38, 40, 42, 45, 50, 52, 56, 58, 59, 69, 71,

Table 13: Key Gaps in System Standards and Their Implications for SLR Automation

Gap Area	Current Limitation	Implication
Protocol Heterogeneity	Incompatible toolchains and ad hoc implementation	Hinders integration and reproducibility across studies
Lack of Unified Benchmarks	Inconsistent metrics and dataset use	Impedes objective comparison and meta-analyses
Feature Reporting Variability	Absence of standardized metadata and reporting frameworks	Obstructs transparency and critical appraisal
Incomplete Automation Coverage	Focus on screening, limited support for data extraction/synthesis	Restricts end-to-end automation potential
Opaque Algorithm Documentation	Insufficient provenance, parameter, and workflow disclosure	Reduces trust and impedes independent validation

Table 14: Key Challenges and Collaborative Responses in Open Datasets and Benchmarking

Challenge	**Domain Status**	**Collaborative Response**
Dataset Domain Skew	Predominance of biomedicine and computer science	Calls for domain expansion and cross-domain dataset curation
Limited Diversity/Multimodality	Scarcity of multilingual and multimedia datasets	Initiatives for multilingual/multimodal benchmark creation
Fragmented Evaluation	Disparate metrics and bespoke tasks	Adoption of multi-metric benchmarking suites and leaderboards
Tool Comparability	Inconsistent reporting and transparency	Community-led review and challenge platforms integrating standard protocols
Benchmark Stagnation	Static datasets lagging behind field developments	Living datasets, APIs, and continuous integration pipelines

76, 79, 81, 82, 85, 86, 88, 89, 91, 93, 103, 107, 108]. This resource intensity disproportionately affects institutions with limited access to advanced hardware, thereby amplifying inequities within the global research community. Notably, scalable and lightweight solutions remain underexplored and underutilized [56, 94, 103].

Tensions between **openness and privacy** persist, particularly in contexts involving sensitive processes such as peer review automation, survey response classification, or author attribution. Such processes raise complex concerns regarding participant confidentiality, fairness, and potential algorithmic biases [11, 15, 20, 28, 42, 45, 47, 56, 58, 71, 73, 85, 98, 107]. Empirical studies reveal that automated systems frequently encode or exacerbate existing linguistic, demographic, or geographic biases present in training data, thereby impacting the fairness and reliability of outputs—especially when models struggle with non-dominant languages or diverse author profiles [15, 17, 27, 31, 42, 56, 58, 71, 78, 82, 89, 107, 110].

The rapid proliferation of AI-generated content demands robust systems for **oversight and fraud prevention**. Standard detection techniques, such as watermarking and stylometric analysis, increasingly fail to keep pace with adversarial tactics and the sophistication of content generation models, thus posing risks to both academic integrity and information reliability [12, 28, 45, 56, 71, 73, 110]. These detection methods often exhibit inconsistent effectiveness across domains and languages and may result in disproportionate penalties for non-native users, highlighting the urgent need for equitable and adaptive detection frameworks [17, 42, 47, 71, 89, 107, 110]. Structured comparison of these detection approaches and their limitations is presented in Table 15.

Another pronounced gap lies in the **long-term and cross-cultural validation** of AI-powered academic workflows. The overwhelming majority of empirical assessments are confined to Anglophone or biomedical contexts, with minimal focus on under-resourced languages, diverse academic cultures, or longitudinal impacts of automation on scholarly ecosystems [28, 58]. Furthermore, there is still a paucity of rigorous empirical evaluations, especially regarding the downstream effects of automation on research quality, user behaviors, and knowledge dissemination [28, 58, 98]. This shortfall undermines transparency, generalizability, and evidence-based development in automated tools.

Finally, **survey automation presents its own set of challenges**: high attrition rates, complexity of technical setup, dependency on proprietary systems, and inconsistent data completeness all impair the reliability and representativeness of automated survey pipelines [28, 85]. Empirical results suggest that, although these tools may significantly reduce manual workload, they often do so at the expense of sensitivity to rare but critical signals, reinforcing the continued necessity of human oversight [28, 58].

11.2 Opportunities and Future Directions

Amidst the current landscape of challenges, compelling opportunities are emerging that could enable the development of **innovative and empowering solutions** for systematic reviews and academic knowledge production. Advances in **interpretable and intelligent agents**—incorporating multi-modal, multi-lingual, and cross-lingual functionality—have the potential to create more robust and inclusive tools, accessible to researchers around the world [5, 6, 12, 15, 17–19, 26–28, 31, 38, 40, 42, 45, 50, 52, 56, 58, 59, 69–71, 76, 79, 81, 82, 85, 86, 88, 89, 91, 93, 94, 103, 107, 108, 110, 113]. Interdisciplinary innovations in explainable AI—especially those integrating symbolic reasoning, knowledge graphs, and user-centric feedback—offer promising avenues for bridging the prevailing gaps in workflow transparency and interpretability, thereby enlarging participation and fostering greater trust [15, 18, 19, 27, 28, 31, 42, 45, 52, 58, 71, 76, 81, 89, 93, 110, 113].

Significant progress can also be realized through the **expansion of collaborative benchmarking and open data resources**. Shared and standardized corpora, open access APIs, and community-driven evaluation platforms are fundamental for ensuring methodological rigor, reproducibility, and equity. Such resources facilitate systematic auditing across languages, modalities, and disciplines, and underpin the development of meta-research on automation in academic communication [5, 6, 12, 15, 17, 19, 27, 28, 38, 40, 42, 50, 52, 56, 59, 69, 70, 76, 81, 85, 88, 89, 91, 93, 103, 110, 113]. Structured AI usage metadata in publications further enhances transparency

Table 15: Comparative Overview of Automated Fraud Detection Approaches in Academic Workflows

Approach	Typical Use Case	Key Limitations	Equity and Generalizability Issues
Watermarking	Detection of AI-generated text	Vulnerable to paraphrasing, often language-specific, easily removed	Low robustness across languages and domains; false positives for non-native speakers
Stylometric Analysis	Author verification and attribution	Sensitive to text length and domain, challenged by adversarial writing	Bias against less-represented writing styles; unfair penalties for non-dominant language users
Metadata-based Forensics	Peer review and provenance tracking	Reliant on data completeness, susceptible to spoofing	Limited cross-cultural applicability; issues with privacy and consent
Source Attribution (Plagiarism Detection)	Academic misconduct detection	Ineffective when facing sophisticated paraphrasing or mixed sources	May penalize legitimate re-use, particularly in multilingual settings

through machine-readable documentation, supporting downstream tool development and policy-making [56, 58].

There is increasing emphasis on the creation of **integrative and holistic academic workflows**, wherein AI methods can support the entire research life cycle—from literature search and planning, through data extraction and synthesis, to reporting and peer evaluation [5, 28, 31, 52, 56, 58, 70, 81, 103, 110, 113]. Achieving this vision will require robust multidisciplinary collaboration across computer science, information science, linguistics, sociology, and ethics in order to harmonize efficiency, robustness, and fairness [58, 113].

Emerging research highlights the promise of **scalable and lightweight architectures**—including distilled models, multilingual transformers, and agent-based distributed systems—to facilitate democratized automation across diverse resource environments and application domains [5, 26, 42, 56, 58, 69, 70, 76, 81, 85, 86, 88, 94, 103]. When these models are complemented by inclusive interface design and participatory user evaluation, they hold the potential to reduce inequities tied to computational constraints, thereby fostering a more pluralistic research ecosystem.

Looking to the future, the development of **living review systems**—dynamic, AI-enabled frameworks that provide ongoing, open updates to research synthesis—represents a transformative vision for open science. Such systems would be responsive to emerging evidence, evolving domains, and shifting priorities [6, 17, 27, 42, 47, 58, 59, 70, 71, 110, 113]. The integration of privacy-preserving computation, verifiable audit trails, and adaptive fraud detection into these workflows could further ensure the technical and ethical scalability necessary for next-generation knowledge production [12, 15, 28, 45, 56, 73, 85, 98, 107, 110].

In summary, while obstacles related to data quality, transparency, equity, computational demands, privacy, and empirical validation currently circumscribe the reach of automation in systematic reviews and academic workflows, emerging research across disciplines has begun to chart a roadmap for overcoming these barriers. Sustained innovation will depend on advancing interpretability, inclusiveness, reproducibility, and strong multidisciplinary collaboration, guaranteeing that automated systems augment—rather than replace—the essential human judgments intrinsic to scholarly inquiry.

12 Synthesis, Best Practices, and Conclusion

12.1 Responsible Integration and Adoption

The accelerating convergence of artificial intelligence (AI), agentic architectures, and advanced survey automation technologies is fundamentally reshaping standards and practices across academia, healthcare, and policymaking. This transformation imposes a heightened obligation for ethical, transparent, and community-aligned integration strategies. Foundational guidelines stress the necessity of embedding core ethical values—including transparency, explainability, and robust accountability—throughout each stage of the AI lifecycle. Influential frameworks from organizations such as the IEEE, EU, and OECD prescribe beneficence, autonomy, explicability, and justice as indispensable pillars [4, 5, 20, 40, 58, 82]. Nevertheless, the operationalization of these principles, particularly within evolving multimodal and agent-driven systems, remains beset by significant challenges. The presence of complex interactants, opaque model decisions, and rapid deployment cycles frequently strains the limits of oversight and erodes societal trust [11, 20, 40].

To address these challenges, emergent best practices advocate multiple, interdependent strategies. Foremost, system design must prioritize contestability by ensuring that human operators retain both the epistemic access and institutional authority to interrogate, challenge, and override AI-driven outputs. This serves as a critical safeguard against responsibility gaps that might otherwise undermine public trust [11, 40, 47]. Realizing contestability requires the technical integration of explainability modules, confidence scoring mechanisms, and traceable decision histories. Organizationally, it demands structural reforms that empower human operator intervention, especially within high-stakes or time-sensitive decision-making environments [4, 20, 40, 47].

Additionally, the institutionalization of continuous improvement cycles is paramount. This process should be grounded in robust, harmonized standards governing automated writing, monitoring, and survey quality assurance [11, 12, 20, 22, 40, 59, 78, 94, 110]. Tools such as semi- and fully-automated systematic review platforms exemplify the efficiency and rigor gains that can be achieved through transparent, iterative development. Key elements include judicious use of human-in-the-loop oversight, cross-tool and multi-LLM validation strategies, and explicit thresholding mechanisms—all of which facilitate automation while mitigating complacency and preserving methodological soundness [12, 40, 59, 78, 94, 110]. Despite this progress, significant limitations persist; inconsistent framework adoption, inadequate support for "living" (continuously updated) reviews, and barriers to transparency and usability highlight areas requiring persistent collaborative refinement and open science initiatives [12, 20, 22, 110].

Societal alignment further compels participatory and multidisciplinary approaches for the governance of AI and agentic systems at scale [1, 2, 4–6, 8, 9, 11, 12, 14–17, 19–21, 23–25, 27, 28, 32, 40, 42, 43, 45, 47–54, 56–58, 60, 61, 63, 64, 68, 70–72, 74–76, 80, 82, 83, 86, 89–94, 104–108, 110, 111, 115]. The integration of AI into academic, clinical, and policy domains encounters both sector-specific and cross-sectoral obstacles, including data privacy risks, institutional and algorithmic bias, and challenges in equitable access [4, 15, 20, 40, 47, 89]. Addressing such challenges necessitates policy-level mechanisms that balance scalability with ethical sustainability: public registries for AI tool usage, standardized metadata reporting, and formalized avenues for surfacing community concerns

within development and oversight cycles [20, 42, 58, 82]. Prominent models utilizing distributed agent interactions, such as those in multi-agent literature synthesis and swarm-based optimization for resource allocation, demonstrate the critical importance of transparency, resilience, and adaptive governance for enduring societal benefit [21, 48, 61, 74, 92, 111].

A resilient trajectory for the scalable and ethical adoption of AI centers on four interlocking tenets: embedding ethical principles by design; ensuring contestability and traceability of system outputs; harmonizing continuous improvement through cross-domain standards; and instituting participatory, transparent, and accountable governance at every level [4, 5, 11, 12, 20, 22, 40, 47, 58, 59, 78, 82, 94, 110]. Only through deliberate institutional strategies, buttressed by rigorous impact assessments and adaptive regulatory frameworks, can the promise of transformative AI be captured in alignment with societal values and the imperative to sustain public trust.

12.2 Summary of Advances and Open Issues

Recent years have marked significant advances in the interoperability, autonomy, and scalability of AI and agentic systems. However, these technological leaps also bring forth persisting challenges in ethics, scientific integrity, and compliance. At the technological frontier, the deployment of multimodal large language models (LLMs), hierarchical agent-based systems, and automated survey tools has expanded operational capacity—from conducting hundreds of literature reviews per hour to enabling real-time optimization in complex distributed environments [12, 21, 40, 48, 59, 78, 94, 110, 111]. Noteworthy innovations, such as automated design of agentic systems (ADAS), represent the next stage of meta-automation, in which meta-agents autonomously create robust and adaptive agents, catalyzing advances across science, education, and organizational learning [21, 48, 61, 74, 111]. Methodological progress is further evident in new strategies for automatic question generation, advances in rationalization for explainability, and frameworks supporting continuous-updating ("living") academic and clinical reviews [40, 59, 78, 94, 110].

As delineated in Table 16, while the technological capabilities of AI and agentic systems have expanded dramatically, persistent open issues remain central to future research and deployment efforts. Chief among these challenges is the increasing opacity and unpredictability inherent in highly autonomous, multimodal systems, whose interactions often give rise to emergent behaviors—including bias, misalignment, and unforeseen societal impacts—that defy straightforward technical remediation [4, 40, 58, 82]. Although mechanisms such as operator contestability, independent audits, and transparent system design are widely advocated as compensatory safeguards, their practical implementation remains inconsistent, and ongoing tensions persist between the speed of technical innovation and the pace of regulatory adaptation [11, 20, 40, 47].

Furthermore, automation in evidence synthesis can alleviate human workload but simultaneously introduces new vulnerabilities: overreliance on algorithmic processes, incomplete data coverage, and the risk of error propagation at scale [12, 59, 78, 94, 110]. Automated survey and agentic research tools, when inadequately monitored, may amplify data contamination and fraud, or erode trust in scientific outputs [4, 11, 15, 20, 23, 27, 40, 42, 47, 50, 52, 56, 71, 82, 89, 107, 108].

The incorporation of AI into academic writing, assessment, and survey research also prompts unresolved questions regarding critical thinking, equity, and the preservation of creative autonomy [15, 19, 23, 28, 50, 52, 56, 71, 89, 91, 93, 107, 108]. Although empirical studies suggest that AI tools can enhance productivity and personalize feedback, they are not substitutes for the holistic educational and evaluative roles fulfilled by human instructors, nor can they uphold academic integrity absent rigorous curricular and policy interventions [15, 19, 28, 52, 56, 71, 107].

Looking forward, several priorities for future research and implementation emerge. These include the creation of harmonized benchmarks and validation protocols for agentic systems and AI-augmented reviews; scalable governance frameworks encompassing technical, societal, and regulatory stakeholders; intensified investment in explainable, contestable, and resilient system architectures; and the institutionalization of living, continually updated review systems with explicit, transparent reporting standards [5, 11, 12, 20, 22, 40, 42, 47, 58, 59, 71, 78, 82, 94, 110]. Progress in these domains is contingent upon enhanced cross-sectoral collaborations, which are essential to synchronize technological innovation with the imperatives of ethical stewardship and public trust. By anchoring the integration of AI and agentic technologies in values that serve the broadest possible societal interest, the field can fully realize the transformative potential of these advances.

13 Appendices (Exemplary Artifacts and Case Details)

13.1 Benchmark Tables and Datasets

A robust foundation for evaluating survey, review, monitoring, and detection systems hinges on access to comprehensive benchmarks and datasets. Recent years have seen significant advancements in compiling benchmark datasets that encapsulate domain complexity, linguistic diversity, and evolving operational demands. Large-scale open-domain datasets—including SQuAD, MS MARCO, RACE, NewsQA, TriviaQA, and LearningQ—originally developed for tasks such as question generation and answer assessment, have become instrumental for constructing and evaluating automatic survey and review systems. Nevertheless, their adoption has highlighted persistent shortcomings in subjective and multimedia domains [6, 7, 12].

Advanced multi-domain benchmarks, such as SurveyBench and BigSurvey, provide structured, long-form, multi-document summarization references essential for rigorous performance assessment in literature review automation [94, 110]. In parallel, detection and monitoring frameworks utilize curated corpora dedicated to AI-generated text (AIGT) detection—such as GPABench2, OUTFOX, and LLMFake—which facilitate systematic evaluation of fraud detection and integrity monitoring workflows [15, 50, 82].

Despite these advances, multi-lingual and domain-specific datasets remain insufficiently represented, limiting both generalizability and fairness in assessment [42, 107]. Both commercial and open-source solutions—from DistillerSR and LiteRev to experimental prototypes like SurveyX and SurveyForge—are commonly benchmarked on

Table 16: Key Technological Advances and Persistent Open Issues in AI and Agentic Systems

Technological Advances	Persistent Open Issues
Multimodal large language models (LLMs) and hierarchical agent-based systems enabling large-scale and real-time tasks	Opacity and unpredictability of autonomous, multimodal systems leading to emergent risks (e.g., bias, misalignment, unintended consequences)
Automated design of agentic systems (ADAS) empowering meta-agents to create adaptive agents	Inconsistent implementation of contestability, transparency, and audit mechanisms; regulatory lag
Iterative, automated systematic review tools and question-generation frameworks enhancing research and survey workflows	Overreliance on automation, potential error propagation, data completeness challenges, and risks of data contamination or scientific fraud
Cross-tool and multi-LLM validation increasing methodological rigor and efficiency	Threats to academic integrity, critical thinking, and creative autonomy in writing and assessment

recall, precision, workload reduction, and scalability. However, considerable heterogeneity persists in evaluation standards, underscoring the ongoing need for consensus methodologies and greater transparency in reporting [17, 42, 50, 56, 71, 76, 89, 93, 107].

Table 17 provides a comparative view of prominent datasets and systems, emphasizing their respective evaluation domains and key limitations.

13.2 Example Code Listings and Pipelines

The operational landscape of literature review automation and monitoring is underpinned by diverse computational pipelines, often illustrated through code artifacts encompassing style transfer, prompt tuning, survey automation, agent-based monitoring, topic modeling, and AIGT detection. For instance, inverse style transfer for authorship leverages large language models (LLMs) to generate paired (neutral, stylized) instances, thereby enabling robust replication of author style even under conditions of data scarcity [7]. Similarly, prompt tuning—utilizing techniques such as P-Tuning and continuous prompt embeddings—has been shown to stabilize LLM performance, effectively reducing reliance on volatile hand-crafted prompt compositions [6][70].

Survey automation pipelines typically capitalize on modular architectures for data extraction, ranking, iterative classification, and integration with external interfaces. Notable examples include LiteRev's open-source modules for literature clustering and automated WhatsApp survey administration frameworks, which employ the WhatsApp Business API in conjunction with cloud-based data integration to support scalable, interactive surveys [59][91][15][110]. Agent-based monitoring frameworks combine scriptable simulation environments with structured decision logic—using either BDI agents or reinforcement learning strategies—thereby ensuring reproducibility and extensibility across diverse monitoring scenarios [93][50][107][40][94][17][42][71][56][28].

Topic modeling, through methods such as Latent Dirichlet Allocation (LDA) and supervised LDA (sLDA), facilitates explainable, large-scale survey analysis. Accompanying codebases routinely provide annotated notebooks for preprocessing, model fitting, and result interpretation, bolstering the academic transparency and extensibility of these tools [52][108][42]. In the realm of AI-generated text and fraud detection, ensemble detectors, neural watermarking, and task-specific classifiers have become standard; these approaches benefit from public code releases and accessible API platforms, enabling proactive research and operational deployment [50][15][110][28].

While improvements in code portability and reusability are evident, harmonizing cross-platform standards and robust benchmarking remain ongoing challenges, particularly within multi-agent and multilingual contexts [107][94][42][56].

13.3 Computational Roadmaps and Reproducibility Workflows

Ensuring reproducibility and methodological transparency is increasingly central to the advancement of automated survey, synthesis, writing, and monitoring pipelines. Contemporary frameworks such as AutoSurvey, SurveyX, and SurveyForge outline modular roadmaps comprising reference retrieval, outline generation, section drafting, and integrative review—each phase subjected to stringent evaluation, version-controlled script sharing, structured data splits, and detailed configuration manifests [6, 19, 40, 42, 56, 71, 94, 110].

Emergent writing automation practices favor reflective, hierarchical designs that prioritize citation tracking, content coverage, and iterative revision, aligning computational artifacts with established standards and evidentiary rigor. Such workflows include clear provenance annotations and comprehensive artifact logs to facilitate transparency [6, 11, 17, 19, 27, 40, 42, 50, 56, 71, 94, 107]. Similarly, monitoring and agent-based modeling workflows are regularly distributed along with reproducible simulation environments, scenario scripts, and model checkpoints to enable consistent benchmarking and support real-time auditability [11, 27, 42, 50, 56, 71, 107, 110].

Recent advances reinforce the imperative of documenting not only the code and datasets, but also parameter settings, initialization seeds, and environment specifications—especially in contexts where empirical outcomes are sensitive to such configurations [17, 40, 107]. Despite these strides, reproducibility efforts are frequently constrained by inconsistent open-source licensing, incomplete documentation, or narrow data access. Consequently, there is a pressing need for further harmonization of workflow standards, dataset versioning, and open science practices to foster replicable, scalable research in review automation [42, 56, 71].

13.4 Case Studies and Exemplary Systems

A fertile landscape of agentic platforms, domain-specialized and multilingual models, and novel architectures offers a window into the practical realities of automated survey, review, monitoring, and detection systems. Prominent agentic systems harnessing LLMs—including vertical domain agents and multi-agent simulators—demonstrate full-cycle automation across contexts such as healthcare and finance, accommodating real-time adaptability and compliance requirements [15, 19, 50, 73, 82, 89, 91, 107, 108]. Next-generation automated review frameworks, for instance SurveyForge and SurveyX, exemplify the shift from static, monolithic SaaS modalities to orchestrated, multi-agent architectures, resulting in improvements in accuracy, content quality, and cost-efficiency under diverse benchmark scenarios [11, 19, 27, 40, 103, 110].

Adaptations for multilingual and domain-specific deployment—like MindLLM and medical informatics-focused surveillance—underscore critical challenges and solutions for extending automation into

Table 17: Overview of representative benchmark datasets and systems, highlighting domain focus, core evaluation roles, and current gaps.

Dataset/System	Domain Coverage	Benchmark Purpose	Key Gaps
SQuAD, MS MARCO	Open-Domain	QG/QA/Summarization	Subjective/MM domains
SurveyBench	Multi-Disciplinary	Structured Survey Eval.	Multilingual span
GPABench2, OUTFOX	AIGT Detection	Fraud/Integrity Monitoring	Domain diversity
LiteRev, DistillerSR	Biomedical / Systematic	Review Automation Bench.	Standardization
SurveyX, SurveyForge	Research Prototypes	Modular Automation	Scalability, reporting

low-resource or highly specialized environments [15, 40, 82, 107]. Mobile-first systems leveraging WhatsApp's API illustrate innovative strategies for survey delivery among hard-to-reach populations, featuring adaptive branching, automated engagement, and seamless data flow for longitudinal research [17, 28, 42, 56, 71, 110].

Comprehensive evaluations of exemplary deployments span both technical and empirical axes (e.g., accuracy, efficiency, engagement metrics), emphasizing the vital synergy between automation, ethical safeguards, and human oversight. Many leading systems specifically report on the centrality of human-mediated intervention at critical decision points, particularly in fraud detection or sensitive participant engagement—functions that currently exceed the reliable reach of automation alone [17, 28, 42, 56, 71, 108]. This trend underscores the effectiveness of human-in-the-loop architectures in handling complexity, uncertainty, and adversarial dynamics [15, 28, 50, 73, 82, 91].

13.5 Metadata Proposals and Policy Guidance

In pursuit of greater transparency, trust, and research continuity, the field is moving toward systematic standardization of AI and agentic tool metadata, structured survey system proposals, and policy frameworks tailored to automated scholarly communication [20][76][28]. Metadata proposals advocate formalized, machine-readable formats comprising tool name, version, parameters, use context, and targeted manuscript sections. Such standards are envisioned to facilitate large-scale trend analysis, linguistic evolution tracking, and impact assessment of AI integration within academic workflows [20].

Implementation best practices prioritize interoperable formats such as JSON or XML and in-depth collaboration among publishers, databases, and researchers to enforce consistent reporting and tool traceability [20][28]. Policy guidance increasingly foregrounds real-time monitoring, responsive design, trust, and contestability, advocating for cost-quality optimizations, operator empowerment, and safeguarded intervention capabilities [76]. These recommendations aim to proactively counteract risks like data contamination, automation bias, and ambiguous accountability, calling for a blend of technical and institutional reforms [76][28].

Longitudinal case studies and participatory design initiatives further highlight the necessity of aligning computational methods with evolving regulatory, ethical, and community guidelines, thereby upholding both scientific rigor and broader societal trust [20][76][28].

13.6 Research Summary Tables

Synthesis across longitudinal, cross-cultural, and multi-domain studies reveals persistent gaps and forward-looking recommendations for the field. Principal challenges include the scarcity of representative multilingual and multimodal datasets [58, 113], the demand for standardized benchmarking, and the necessity for transparent performance reporting across system functionalities [42, 71]. Additionally, there is an emergent consensus surrounding robust, dynamically updatable review systems capable of evolving in tandem with continuously shifting disciplinary landscapes and data streams [28, 56, 71].

Strategic research summaries do more than catalog system features or highlight deficiencies. They map actionable pathways for integrating agentic modeling into survey workflows, establish best practices for open science and reproducibility, and articulate strategies to harmonize human and algorithmic judgment amidst uncertainty or adversarial actors. When published in concert with corresponding code, data, and policy artifacts, these tables provide a durable and actionable foundation for propelling future research, development, and deployment efforts across domains [28, 42, 56, 71].

Table 18 presents a synthesized overview of central research gaps with mapped recommendations and their potential to drive progress across the automation domain.

References

[1] F. Adobbati and Ł. Mikulski. 2025. Asynchronous Multi-Agent Systems with Petri nets. *arXiv preprint arXiv:2504.00602* (2025). https://arxiv.org/abs/2504.00602

[2] M. Afzaal, J. Nouri, A. Zia, P. Papapetrou, U. Fors, Y. Wu, X. Li, and R. Weegar. 2021. Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation. *Frontiers in Artificial Intelligence* 4 (2021), 723447. doi:10.3389/frai.2021.723447

[3] N. F. Ali, M. M. Mohtasim, S. Mosharrof, and T. G. Krishna. 2024. Automated Literature Review Using NLP Techniques and LLM-Based Retrieval-Augmented Generation. *arXiv preprint arXiv:2411.18583* (2024). https://arxiv.org/abs/2411.18583

[4] H. Aljuaid. 2024. The Impact of Artificial Intelligence Tools on Academic Writing Instruction in Higher Education: A Systematic Review. *Arab World English Journal (AWEJ) Special Issue on ChatGPT* (April 2024), 1–30. https://ssrn.com/abstract=4814342

[5] H. Aljuaid. 2024. The Impact of Artificial Intelligence Tools on Academic Writing Instruction in Higher Education: A Systematic Review. Online. https://osf.io/ph24v/download/ Accessed: 2024-06-09.

[6] D. Antons, C. F. Breidbach, A. M. Joshi, and T. O. Salge. 2023. Computational Literature Reviews: Method, Algorithms, and Roadmap. *Organizational Research Methods* 26, 1 (2023), 107–138. doi:10.1177/1094428121991230

[7] C. F. Atkinson. 2024. Cheap, Quick, and Rigorous: Artificial Intelligence and the Systematic Literature Review. *Social Science Computer Review* 42, 2 (2024), 376–393. doi:10.1177/08944393231196281

[8] Michael Belfrage, Emil Johansson, Fabian Lorig, and Paul Davidsson. 2024. [In]Credible Models – Verification, Validation & Accreditation of Agent-Based Models to Support Policy-Making. *Journal of Artificial Societies and Social Simulation* 27, 4 (2024), 4. doi:10.18564/jasss.5505

Table 18: Summary of key research gaps, recommended actions, and projected impacts on the future of automated survey, review, and monitoring systems.

Gap / Challenge	Recommendation	Anticipated Impact
Insufficient multilingual/multimodal datasets	Invest in open, equitable data expansion, prioritizing low-resource and diverse modalities	Enhanced fairness, global applicability, mitigation of language/data silo effects
Lack of standardized benchmarking/reporting	Develop consensus frameworks, mandatory feature/performance reporting	Transparent, comparable assessments, accelerated field-wide progress
Limited reproducibility and workflow harmonization	Promote open science, comprehensive version control/documentation	Increased replicability, community trust, scaled collaboration
Overreliance on automation for sensitive tasks	Integrate adaptive human-in-the-loop and contestable mechanisms	Greater resilience, contextual intelligence, and ethical assurance

[9] N. V. Blamah, A. A. Oluyinka, G. Wajiga, and Y. B. Baha. 2020. MAPSOFT: A Multi-Agent based Particle Swarm Optimization Framework for Travelling Salesman Problem. *Journal of Intelligent Systems* 30, 1 (2020), 413–428. doi:10.1515/jisys-2020-0080

[10] S. Blaschke. 2024. Publication authorship: A new approach to the bibliometric study of scientific work and beyond. *PLOS ONE* 19, 4 (2024), e0297005. doi:10.1371/journal.pone.0297005

[11] M. M. Boillos and N. Idoiaga. 2025. Student perspectives on the use of AI-based language tools in academic writing. *Journal of Writing Research* (2025). https://www.jowr.org/jowr/article/view/1518 Early view, published Jan. 24, 2025.

[12] F. Bolaños, A. Salatino, F. Osborne, and E. Motta. 2024. Artificial intelligence for literature reviews: opportunities and challenges. *Artificial Intelligence Review* 57, 259 (2024), 1–59. doi:10.1007/s10462-024-10902-3

[13] F. Bousetouane. 2025. Agentic Systems: A Guide to Transforming Industries with Vertical AI Agents. *arXiv preprint arXiv:2501.00881* (2025). https://arxiv.org/abs/2501.00881

[14] L. Buess, M. Keicher, N. Navab, A. Maier, and S. Tayebi Arasteh. 2025. From large language models to multimodal AI: A scoping review on the potential of generative AI in medicine. *arXiv preprint arXiv:2502.09242* (2025). https://arxiv.org/abs/2502.09242

[15] Christine P. Chai. 2019. Text Mining in Survey Data. *Survey Practice* 12, 1 (2019). doi:10.29115/SP-2018-0035

[16] K. E. K. Chai, R. L. J. Lines, D. F. Gucciardi, and L. Ng. 2021. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic Reviews* 10 (2021), Article no. 93. doi:10.1186/s13643-021-01635-3

[17] L. Chen, P. Chen, and Z. Lin. 2020. Artificial Intelligence in Education: A Review. *IEEE Access* 8 (2020), 75264–75278. doi:10.1109/ACCESS.2020.2988510

[18] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113. https://www.jmlr.org/papers/volume24/22-1144/22-1144.pdf

[19] Asaph Young Chun, Steven G. Heeringa, and Barry Schouten. 2018. Responsive and Adaptive Design for Survey Optimization. *Journal of Official Statistics* 34, 3 (2018), 581–597. https://sciendo.com/article/10.2478/jos-2018-0028

[20] J. Conde, P. Reviriego, J. Salvachúa, G. Martínez, J. A. Hernández, and F. Lombardi. 2024. Understanding the Impact of Artificial Intelligence in Academic Writing: Metadata to the Rescue. *Computer* 57, 1 (2024), 85–88. https://arxiv.org/abs/2502.16713

[21] K. Cowie, A. Rahmatullah, N. Hardy, K. Holub, and K. Kallmes. 2022. Web-Based Software Tools for Systematic Literature Review in Medicine: Systematic Search and Feature Analysis. *JMIR Medical Informatics* 10, 5 (2022), e33219. doi:10.2196/33219

[22] K. Cowie, A. Rahmatullah, N. Hardy, K. Holub, and K. Kallmes. 2022. Web-Based Software Tools for Systematic Literature Review in Medicine: Systematic Search and Feature Analysis. *JMIR Medical Informatics* 10, 5 (2022), e33219. https://medinform.jmir.org/2022/5/e33219/

[23] A. Dafoe, R. Sandbrink, C. O'Brien, S. Cotton-Barratt, J. Drexler, F. Flynn, J. Hannon, P. Kwon, L. Maynard, K. Redei, R. Salvatier, and M. Scharre. 2018. When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research* 62 (2018), 729–754. doi:10.1613/jair.1.11222

[24] C. Dalvi, M. Rathod, S. Patil, S. Gite, and K. Kotecha. 2021. A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets and Future Directions. *IEEE Access* 9 (2021), 165806–165840. doi:10.1109/ACCESS.2021.3137226

[25] B. Das, M. Majumder, S. Phadikar, and S. A. Ahmed. 2021. Automatic question generation and answer assessment: a survey. *Research and Practice in Technology Enhanced Learning* 16, 5 (2021), 1–42. https://rptel.apsce.net/index.php/RPTEL/article/download/2021-16005/29/62

[26] J. de la Torre-López, A. Ramírez, and J. R. Romero. 2023. Artificial intelligence to automate the systematic review of scientific literature. *Computing* 105 (2023), 2171–2194. doi:10.1007/s00607-023-01181-x

[27] M. Z. Degu and M. Meshesha. 2024. Fine-Tuned Pretrained Transformer for Amharic News Headline Generation. *Applied AI Letters* 5, 4 (2024), 1–10. doi:10.1002/ail2.98

[28] C. Delcea and N. Chirita. 2023. Exploring the Applications of Agent-Based Modeling in Transportation. *Applied Sciences* 13, 17 (2023), 9815. doi:10.3390/app13179815

[29] C. Delcea, R. J. Milne, and L.-A. Cotfas. 2022. Evaluating Classical Airplane Boarding Methods for Passenger Health during Normal Times. *Applied Sciences* 12, 7 (2022), 3235. doi:10.3390/app12073235

[30] D. Dell'Anna, N. Alechina, F. Dalpiaz, M. Dastani, and B. Logan. 2022. Data-Driven Revision of Conditional Norms in Multi-Agent Systems. *Journal of Artificial Intelligence Research* 75 (2022), 1377–1418. doi:10.1613/jair.1.13683

[31] S. Demir, U. Mutlu, and Ö. Özdemir. 2019. Neural Academic Paper Generation. *arXiv preprint arXiv:1912.01982* (2019). https://arxiv.org/abs/1912.01982

[32] D. Deplano, N. Bastianello, M. Franceschelli, and K. H. Johansson. 2025. Optimization and Learning in Open Multi-Agent Systems. *arXiv preprint arXiv:2501.16847* (2025). https://arxiv.org/abs/2501.16847

[33] A. Domenteanu, C. Delcea, N. Chiriță, and C. Ioanăș. 2023. From Data to Insights: A Bibliometric Assessment of Agent-Based Modeling Applications in Transportation. *Applied Sciences* 13, 23 (2023), 12693. doi:10.3390/app132312693

[34] O. Erin, X. Liu, J. Ge, J. Opfermann, Y. Barnoy, L. O. Mair, J. U. Kang, and Y. Diaz-Mercado. 2022. Comparative Analysis of Sensors in Rigid and Deformable Modular Robots for Shape Estimation. *Advanced Intelligent Systems* 4, 6 (2022), 2200072. doi:10.1002/aisy.202200072

[35] Mark Esposito, Saman Sarbazvatan, Terence Tse, and Gabriel Silva-Atencio. 2024. The use of artificial intelligence for automatic analysis and reporting of software defects. *Frontiers in Artificial Intelligence* 7 (2024), 1443956. doi:10.3389/frai.2024.1443956

[36] W. Fedus, B. Zoph, and N. Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39. https://www.jmlr.org/papers/volume23/21-0998/21-0998.pdf

[37] Hiran Ferreira, Guilherme P. de Oliveira, Rafael Araújo, Fabiano Dorça, and Renan Cattelan. 2019. Technology-enhanced assessment visualization for smart learning environments. *Smart Learning Environments* 6 (2019), 14. doi:10.1186/s40561-019-0096-z

[38] Giorgio Franceschelli and Mirco Musolesi. 2023. Reinforcement Learning for Generative AI: State of the Art, Opportunities and Open Research Challenges. *Journal of Artificial Intelligence Research* 78 (2023), 859–899. https://jair.org/index.php/jair/article/view/14369

[39] Kathleen C. Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2023. Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods. *Journal of Artificial Intelligence Research* 79 (2023), 1–52. https://jair.org/index.php/jair/article/view/14438

[40] K. C. Fraser, H. Dawkins, and S. Kiritchenko. 2025. Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods. *Journal of Artificial Intelligence Research* 82 (2025), 1145–1187. doi:10.1613/jair.1.16665

[41] J. Fu, A. Tacchetti, J. Perolat, and Y. Bachrach. 2021. Evaluating Strategic Structures in Multi-Agent Inverse Reinforcement Learning. *Journal of Artificial Intelligence Research* 71 (2021), 953–993. doi:10.1613/jair.1.12594

[42] A. Golda, S. Debnath, N. Gupta, S. Mondal, B. Sikdar, and C. Kim. 2024. Privacy and Security Concerns in Generative AI: A Comprehensive Survey. *IEEE Access* 12 (2024), 48126–48144. doi:10.1109/ACCESS.2024.3381611

[43] Shuanglei Gong. 2024. Transition from machine intelligence to knowledge intelligence: A multi-agent simulation approach to technology transfer. *Journal of Intelligent Systems* 33, 1 (2024), 20230320. doi:10.1515/jisys-2023-0320

[44] Cristobal Rodolfo Guerra-Tamez, Keila Kraul Flores, Gabriela Mariah Serna-Mendiburu, David Chavelas Robles, and Jorge Ibarra Cortés. 2024. Decoding Gen Z: AI's influence on brand trust and purchasing behavior. *Frontiers in Artificial Intelligence* 7, Article 1323512 (2024). doi:10.3389/frai.2024.1323512

[45] Eddie Guo, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Michael Paget, and Christopher Naugler. 2024. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *Journal of Medical Internet Research* 26 (2024), e48996. doi:10.2196/48996

[46] E. Guo, M. Gupta, J. Deng, Y.-J. Park, M. Paget, and C. Naugler. 2024. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *Journal of Medical Internet Research* 26 (2024), e48996. doi:10.2196/48996

[47] H. Guo and S. H. Zaini. 2024. Artificial Intelligence in Academic Writing: A Literature Review. *Asian Pendidikan* 4, 2 (2024), 46–55. doi:10.53797/aspen.v4i2.6.2024

[48] S. Gurrapu, A. Kulkarni, L. Huang, I. Lourentzou, and F. A. Batarseh. 2023. Rationalization for explainable NLP: a survey. *Frontiers in Artificial Intelligence* 6 (2023), Article 1225093. https://www.frontiersin.org/articles/10.3389/frai.2023.1225093/full

[49] E. Mendez Guzman, V. Schlegel, and R. Batista-Navarro. 2024. From outputs to insights: a survey of rationalization approaches for explainable text classification. *Frontiers in Artificial Intelligence* 7 (2024), Article 1363531. https://www.frontiersin.org/articles/10.3389/frai.2024.1363531/full

[50] H. Gweon, M. Schonlau, and M. Wenemark. 2020. Semi-automated classification for multi-label open-ended questions. *Survey Methodology* 46, 2 (2020), 265–282. https://www150.statcan.gc.ca/n1/pub/12-001-x/2020002/article/00005-eng.htm

[51] H. Hassani and E. S. Silva. 2023. The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field. *Big Data and Cognitive Computing* 7, 2 (2023), 45. https://www.mdpi.com/2504-2289/7/2/45

[52] Z. He and M. Schonlau. 2020. Automatic Coding of Open-ended Questions into Multiple Classes: Whether and How to Use Double Coded Data. *Survey Research Methods* 14, 3 (2020), 267–287. doi:10.18148/srm/2020.v14i3.7639

[53] S. Hu, C. Lu, and J. Clune. 2025. Automated Design of Agentic Systems. *arXiv preprint arXiv:2408.08435 [cs.AI]* (2025). https://arxiv.org/abs/2408.08435

[54] M. Imran and N. Almusharraf. 2024. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. *Smart Learning Environments* 11 (2024), 22. doi:10.1186/s40561-024-00310-z

[55] Ebru Yilmaz Ince and Akif Kutlu. 2021. Web-Based Turkish Automatic Short-Answer Grading System. *Natural Language Processing Research* 1, 3-4 (2021), 46–55. https://www.atlantis-press.com/journals/nlpr

[56] M. Jafari, R. Kazemi, A. Mohammadpour, and M. Saif. 2020. A neurobiologically-inspired intelligent trajectory tracking control for unmanned aircraft systems in presence of uncertain system and dynamic environment. *Advanced Intelligent Systems* 2, 12 (2020), 2000140. doi:10.1002/aisy.202000140

[57] Anetta Jedličková. 2024. Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development. *AI & Society* (2024). doi:10.1007/s00146-024-02040-9

[58] S. L. Jen and A. R. Salam. [n. d.]. Using Artificial Intelligence for Essay Writing. Online. https://osf.io/vtcz9/download/?format=pdf Accessed: 2024-06-09.

[59] T. Kasanishi, M. Isonuma, J. Mori, and I. Sakata. 2023. SciReviewGen: A Large-scale Dataset for Automatic Literature Review Generation. *arXiv preprint arXiv:2305.15186* (2023). https://arxiv.org/abs/2305.15186

[60] S. Ariffin Kashinath, S. A. Mostafa, D. Lim, A. Mustapha, H. Hafit, and R. Darman. 2021. A general framework of multiple coordinative data fusion modules for real-time and heterogeneous data sources. *Journal of Intelligent Systems* 30, 1 (2021), 947–965. doi:10.1515/jisys-2021-0120

[61] K. Kolaski, L. Romeiser Logan, and J. P. A. Ioannidis. 2023. Guidance to best tools and practices for systematic reviews. *JBI Evidence Synthesis* 21, 9 (2023), 1699–1731. doi:10.11124/JBIES-23-00139

[62] K. Lannelongue, M. de Milly, R. Marcucci, S. Selevarangame, A. Supizet, and A. Grincourt. 2019. Compositional grounded language for agent communication in reinforcement learning environment. *Journal of Autonomous Intelligence* 2, 1 (2019), 22–44. https://jai.front-sci.com/index.php/jai/article/view/56

[63] Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open* 3 (2022), 106–110. doi:10.1016/j.aiopen.2022.03.001

[64] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* 1 (2024), 1–57. doi:10.1007/s44336-024-00009-2

[65] Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu Li. 2025. SurveyX: Academic Survey Automation via Large Language Models. *arXiv preprint arXiv:2502.14776* (2025). https://arxiv.org/abs/2502.14776

[66] D. J. Liebling, M. Kane, M. Grunde-Mclaughlin, I. J. Lang, S. Venugopalan, and M. P. Brenner. 2025. Towards AI-assisted Academic Writing. In *Proceedings of the NAACL 2025 Workshop on AI for Scientific Discovery*. https://arxiv.org/abs/2503.13771

[67] Chien-Chang Lin, Anna Y. Q. Huang, and Owen H. T. Lu. 2023. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments* 10 (2023), 41. doi:10.1186/s40561-023-00260-y

[68] T. Lin, Y. Wang, X. Liu, and X. Qiu. 2022. A survey of transformers. *AI Open* 3 (2022), 111–132. https://www.sciencedirect.com/science/article/pii/S2666651022000163

[69] S. Liu, J. Cao, R. Yang, and Z. Wen. 2023. Generating a Structured Summary of Numerous Academic Papers: Dataset and Method. *arXiv preprint arXiv:2302.04580* (2023). arXiv:2302.04580 https://arxiv.org/abs/2302.04580

[70] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang. 2023. GPT understands, too. *AI Open* 4 (2023), 30–36. https://www.sciencedirect.com/science/article/pii/S2666651023000141

[71] S. Yanes Luis, H. Li, S. Pan, X. Sun, X. Wei, and J. Yu. 2024. Deep Reinforcement Multiagent Learning Framework for Information Contamination Event Detection. *Advanced Intelligent Systems* 6, 2 (2024), 2300059. doi:10.1002/aisy.202300850

[72] E. Miehling, K. Natesan Ramamurthy, K. R. Varshney, M. Riemer, D. Bouneffouf, J. T. Richards, A. Dhurandhar, E. M. Daly, M. Hind, P. Sattigeri, D. Wei, A. Rawat, J. Gajcin, and W. Geyer. 2025. Agentic AI Needs a Systems Theory. *arXiv preprint arXiv:2503.00237* (2025). https://arxiv.org/abs/2503.00237

[73] Joeri Minnen, Sven Rymenants, Ignace Glorieux, and Theun Pieter van Tienoven. 2023. Answering Current Challenges of and Changes in Producing Official Time Use Statistics Using the Data Collection Platform MOTUS. *Journal of Official Statistics* 39, 4 (2023), 489–505. doi:10.2478/jos-2023-0023

[74] Z. Munn. 2016. Software to support the systematic review process: the Joanna Briggs Institute System for the Unified Management, Assessment and Review of Information (JBI-SUMARI). *JBI Evidence Synthesis* 14, 10 (2016), 1. doi:10.11124/JBISRIR-2016-002421

[75] R. Nakamoto, A. Iwasawa, N. Takahashi, and R. Oka. 2024. Unsupervised techniques for generating a standard explanatory sentence for self-explanatory mathematics. *Research and Practice in Technology Enhanced Learning* 19, 16 (2024), 1–15. https://rptel.apsce.net/index.php/RPTEL/article/download/2024-19016/2024-19016

[76] M. H. Nguyen. [n. d.]. Academic writing and AI: Day-1 experiment. Online. https://osf.io/xgqu5/download Accessed: 2024-06-09.

[77] E. Orel, I. Ciglenecki, A. Thiabaud, A. Temerev, A. Calmy, O. Keiser, and A. Merzouki. 2023. An Automated Literature Review Tool (LiteRev) for Streamlining and Accelerating Research Using Natural Language Processing and Machine Learning: Descriptive Performance Evaluation Study. *Journal of Medical Internet Research* 25 (2023), e39736. doi:10.2196/39736

[78] E. Orel, I. Ciglenecki, A. Thiabaud, A. Temerev, A. Calmy, O. Keiser, and A. Merzouki. 2023. An Automated Literature Review Tool (LiteRev) for Streamlining and Accelerating Research Using Natural Language Processing and Machine Learning: Descriptive Performance Evaluation Study. *Journal of Medical Internet Research* 25 (2023), e39736. https://www.jmir.org/2023/1/e39736/

[79] B. Ozek, Z. Lu, F. Pouromran, S. Radhakrishnan, and S. Kamarthi. 2023. Analysis of pain research literature through keyword Co-occurrence networks. *PLOS Digital Health* 2, 9 (2023), e0000331. doi:10.1371/journal.pdig.0000331

[80] K. Padur, H. Borrion, and S. Hailes. 2025. Using Agent-Based Modelling and Reinforcement Learning to Study Hybrid Threats. *Journal of Artificial Societies and Social Simulation* 28, 1 (2025), 1. https://www.jasss.org/28/1/1.html

[81] K. Pillutla, L. Liu, J. Thickstun, S. Welleck, S. Swayamdipta, R. Zellers, S. Oh, Y. Choi, and Z. Harchaoui. 2023. MAUVE Scores for Generative Models: Theory and Practice. *Journal of Machine Learning Research* 24, 356 (2023), 1–92. https://www.jmlr.org/papers/volume24/23-0023/23-0023.pdf

[82] N. Pinzón, M. M. Mathur, A. H. Liu, D. L. Redmiles, W. D. Bowen, M. A. Rodriguez, J. S. Jones, J. W. Deem, and P. Grabowicz. [n. d.]. AI-powered fraud and the erosion of online survey integrity: An analysis of 31 fraud detection strategies. Online. https://osf.io/95tka/ Accessed: 2024-06-11.

[83] J. Prather, J. Leinonen, N. Kiesler, J. G. Benario, S. Lau, S. MacNeil, N. Norouzi, S. Opel, V. Pettit, L. Porter, B. N. Reeves, J. Savelka, D. H. Smith IV, S. Strickroth, and D. Zingaro. 2024. Beyond the Hype: A Comprehensive Review of Current Trends in Generative AI Research, Teaching Practices, and Tools. arXiv preprint arXiv:2412.14732, to appear in Proceedings of the 2024 Working Group Reports on Innovation and Technology in Computer Science Education (ITiCSE-WGR 2024). https://arxiv.org/abs/2412.14732 Accessed: 2024-06-13.

[84] D. Pugh, M. D. O'Reilly, and C. Jasper. 2020. Can automated item generation be used to develop high-quality multiple-choice questions for medical education? A comparative study. *Research and Practice in Technology Enhanced Learning* 15, 12 (2020), 1–22. https://rptel.apsce.net/index.php/RPTEL/article/download/2020-15012/68/141

[85] M. N. Quang, T. Rogers, J. Hofman, and A. B. Lanham. 2019. New framework for automated article selection applied to a literature review of Enhanced Biological Phosphorus Removal. *PLOS ONE* 14, 5 (2019), e0216126. doi:10.1371/journal.pone.0216126

[86] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam. 2024. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access* 12 (2024), 26839–26874. doi:10.1109/ACCESS.2024.3365742

[87] Archana Rani, Naresh Grover, N. Deepa, and C. Prajitha. 2024. A smart agent-based approach for privacy preservation and threat mitigation to enhance security in the Internet of Medical Things. *Journal of Autonomous Intelligence* 7, 5 (2024), 1–17. https://jai.front-sci.com/index.php/jai/article/view/1629

[88] T. Rehman, D. K. Sanyal, S. Chattopadhyay, P. K. Bhowmick, and P. P. Das. 2023. Generation of Highlights From Research Papers Using Pointer-Generator Networks and SciBERT Embeddings. *IEEE Access* 11 (2023), 91358–91374. doi:10.1109/ACCESS.2023.3292300

[89] M. Revilla. 2022. How to enhance web survey data using metered, geolocation, visual and voice data? *Survey Research Methods* 16, 1 (2022), 1–12. doi:10.18148/srm/2022.v16i1.8013

[90] S. L. Rhodes, S. A. Crabtree, and J. Freeman. 2024. An Agent-Based Model of Hierarchical Information-Sharing Organizations in Asynchronous Environments. *Journal of Artificial Societies and Social Simulation* 27, 2 (2024), 2. https://www.jasss.org/27/2/2.html

[91] Andrea Roberson. 2021. Applying Machine Learning for Automatic Product Categorization. *Journal of Official Statistics* 37, 2 (2021), 395–410. doi:10.2478/jos-2021-0017

[92] D. J. Rosenkrantz, M. V. Marathe, Z. Qiu, S. S. Ravi, and R. E. Stearns. 2025. On Some Fundamental Problems for Multi-Agent Systems Over Multilayer Networks. In *Proceedings of the 24th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2025)*. Detroit, MI. https://arxiv.org/abs/2503.12684

[93] M. Schonlau and M. P. Couper. 2016. Semi-automated categorization of open-ended questions. *Survey Research Methods* 10, 2 (2016), 143–152. doi:10.18148/srm/2016.v10i2.6213

[94] Zhonghui Shao, Jing Zhang, Haoyang Li, Xinmei Huang, Chao Zhou, Yuanchun Wang, Jibing Gong, Cuiping Li, and Hong Chen. 2024. Authorship style transfer with inverse transfer data augmentation. *AI Open* 5 (2024), 94–103. doi:10.1016/j.aiopen.2024.08.003

[95] K. Sowa and A. Przegalinska. 2025. From Expert Systems to Generative Artificial Experts: A New Concept for Human-AI Collaboration in Knowledge Work. *Journal of Artificial Intelligence Research* 82 (2025). doi:10.1613/jair.1.17175

[96] Ahmed Srhir, Tomader Mazri, and Manale Boughanja. 2024. Smart parking: Multi-agent approach, architecture, and workflow. *Journal of Autonomous Intelligence* 7, 4 (2024), 1–16. https://jai.front-sci.com/index.php/jai/article/view/1376

[97] Tim Steuer, Anna Filighera, Thomas Tregel, and André Miede. 2022. Educational Automatic Question Generation Improves Reading Comprehension in Non-native Speakers: A Learner-Centric Case Study. *Frontiers in Artificial Intelligence* 5 (2022), 900304. doi:10.3389/frai.2022.900304

[98] Q. Su, M. Wan, X. Liu, and C.-R. Huang. 2021. Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective. *Natural Language Processing Research* 1, 1 (2021), 1–13. https://www.atlantis-press.com/journals/nlpr

[99] B. Sun and K. Li. 2021. Neural Dialogue Generation Methods in Open Domain: A Survey. *Natural Language Processing Research* 1, 3-4 (2021), 56–70. https://www.atlantis-press.com/journals/nlpr

[100] G. Sundaram and D. Berleant. 2023. Automating Systematic Literature Reviews with Natural Language Processing and Text Mining: a Systematic Literature Review. *arXiv preprint arXiv:2211.15397* (2023). https://arxiv.org/abs/2211.15397

[101] V. Taecharungroj. 2023. 'What Can ChatGPT Do?': Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing* 7, 1 (2023), Article number 20. https://www.mdpi.com/2504-2289/7/1/20

[102] B. Tóth, L. Berek, L. Gulácsi, M. Péntek, and Z. Zrubka. 2024. Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed. *Systematic Reviews* 13, 1 (2024), Article no. 174. https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-024-02592-3

[103] I. van Heerden and A. Bas. 2021. Viewpoint: AI as Author – Bridging the Gap Between Machine Learning and Literary Theory. *Journal of Artificial Intelligence Research* 71 (2021), 1269–1277. doi:10.1613/jair.1.12593

[104] Herman Veluwenkamp and Stefan Buijsman. 2025. Design for operator contestability: control over autonomous systems by introducing defeaters. *AI and Ethics* (2025). doi:10.1007/s43681-025-00657-0

[105] Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. AutoSurvey: Large Language Models Can Automatically Write Surveys. *arXiv preprint arXiv:2406.10252* (2024). https://arxiv.org/abs/2406.10252

[106] Jianan Xu, Jiajin Huang, Jian Yang, and Ning Zhong. 2023. M2GCF: A multi-mixing strategy for graph neural network based collaborative filtering. *Web Intelligence and Agent Systems: An International Journal* 21, 2 (2023), 149–166. doi:10.3233/WEB-220054

[107] Ting Yan, Hanyu Sun, and Anil Battalahalli. 2024. Applying Machine Learning to Survey Question Assessment. *Survey Practice* 17 (2024). doi:10.29115/SP-2024-0006

[108] Ting Yan, Hanyu Sun, and Anil Battalahalli. 2025. Using Machine Learning to Evaluate Questions in a Multilingual Survey. *Survey Practice* 19, Special Issue (mar 2025). doi:10.29115/SP-2024-0021

[109] Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. 2025. SurveyForge: On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing. *arXiv preprint arXiv:2503.04629 [cs.CL]* (2025). https://arxiv.org/abs/2503.04629

[110] Y. Yang, H. Sun, J. Li, R. Liu, Y. Li, Y. Liu, Y. Gao, and H. Huang. 2024. MindLLM: Lightweight large language model pre-training, evaluation and domain application. *AI Open* 5 (2024), 1–26. https://www.sciencedirect.com/science/article/pii/S2666651024000111

[111] A. Yehudai, L. Eden, A. Li, G. Uziel, Y. Zhao, R. Bar-Haim, A. Cohan, and M. Shmueli-Scheuer. 2025. Survey on Evaluation of LLM-based Agents. *arXiv preprint arXiv:2503.16416* (mar 2025). https://arxiv.org/abs/2503.16416

[112] Yuelun Zhang, Siyu Liang, Yunying Feng, Qing Wang, Feng Sun, Shi Chen, Yiying Yang, Xin He, Huijuan Zhu, and Hui Pan. 2022. Automation of literature screening using machine learning in medical evidence synthesis: a diagnostic test accuracy systematic review protocol. *Systematic Reviews* 11 (2022), 11. doi:10.1186/s13643-021-01881-5

[113] R. Zheldibayeva. 2025. The impact of AI and peer feedback on research writing skills: a study using the CGScholar platform among Kazakhstani scholars. *arXiv preprint arXiv:2503.05820* (2025). https://arxiv.org/abs/2503.05820

[114] Tammy Zhong, Yang Song, Raynaldio Limarga, and Maurice Pagnucco. 2023. Computational Machine Ethics: A Survey. *Journal of Artificial Intelligence Research* 77 (2023), 795–841. doi:10.1613/jair.1.14302

[115] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Chengyang Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81. doi:10.1016/j.aiopen.2021.01.001