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, 38 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

Objectives and Section Scope: This subsection aims to clarify the evolving landscape of academic automation by synthesizing trajectories, delineating current frameworks and barriers, and explicitly mapping open research challenges. The focus is to articulate both opportunities and persistent tensions, in alignment with the survey's overarching goal of providing actionable insights and critical perspectives on AI-driven, agentic, and distributed paradigms in scholarly research.

The convergence of traditional academic methodologies with emergent AI-driven, agentic, and distributed paradigms is reshaping foundational scientific standards, workflows, and labor distribution. This transformation brings opportunities for efficiency but also engenders significant tensions, particularly in maintaining transparency, reliability, and privacy as automation scales [1, 9, 12, 31, 36, 43, 86, 90, 114]. The community's adoption of automation typically proceeds incrementally: first targeting high-volume, reproducible tasks, then expanding toward broader integration. This staged approach, coupled with continuous evaluation, helps balance gains in scale and productivity with demands for methodological rigor and adaptability [6, 14, 21, 102].

Hybrid workflow models—where critical human judgment complements powerful automated systems—are particularly promising. These dynamic models allow task allocation to shift based on model confidence, risk, and ethical implications, exemplifying the interplay between automation and human oversight identified as a best practice for reliability and ethical stewardship [4, 10, 14, 47, 86]. However, the path to trustworthy automation remains complex: full substitution of human expertise is not yet practical, and most tools optimize only subcomponents of the research process [4, 12, 21, 47, 102].

As platforms become more intelligent, decentralized, and cloud-based, handling sensitive, proprietary, and high-dimensional personal data grows increasingly challenging. Stringent privacy-by-design practices, robust access controls, and vigilant bias monitoring are required, especially as distributed and agentic systems proliferate [12, 17, 31, 42, 69, 94]. Usability, transparency, and adoption gaps persist, as many tools fall short of providing standardized, open evaluation frameworks and reproducibility [12, 21].

To foster robust, reproducible, and impactful automation, the field must prioritize clear software documentation, open evaluation protocols, living reviews, and cross-disciplinary collaboration [6, 12, 14, 18, 21, 102]. Explicit integration of advanced AI—such as large language models with explainable outputs and knowledge-graph logic—remains an open research challenge [12, 86]. Furthermore, the practical realities of interoperability, model bias, human-centric evaluation, and trust in agentic research assistants continue to demand attention [4, 12, 14, 47, 86].

Open Research Challenges: Formally distinguishing open research challenges and their implications, the most pressing problems include: (1) Integration of advanced AI elements (e.g., LLMs, explainable models, and knowledge graphs) into dynamic, distributed academic workflows for greater transparency and trust [12, 86]. (2) Standardization of reporting, evaluation metrics, and reproducibility protocols to overcome fragmented tool development and inconsistent adoption [12, 21, 102]. (3) Ensuring privacy, security, and robust bias mitigation amid scaling and distribution of agentic research ecosystems [12, 17, 42, 94]. (4) Addressing ethical and practical limitations of fully-automated systems and affirming the enduring value of human critical judgment in atypical scenarios [4, 14, 47]. (5) Bridging usability gaps to close the adoption loop between technically-advanced tools and real-world researchers, particularly in fields with highly diverse workflows and regulatory demands [12, 21].

Synthesis and Takeaways: Achieving the benefits of automation in academic research will require a multidisciplinary roadmap that emphasizes transparency, adaptability, and responsible stewardship. Critical gaps—especially those in explainable automation, privacy, evaluation frameworks, and effective human–AI collaboration—remain substantial but addressable through cross-field cooperation and evidence-driven refinement. Realizing the full promise of AI-driven scholarship will depend less on technological determinism and more on fostering a culture of continuous critical evaluation and open, inclusive innovation.

2.4 Automated and Hybrid Survey Systems

This subsection provides a focused examination of automated and hybrid systems developed for AI-driven literature surveys. The goal is to clarify how these methods enhance scalability, efficiency, and reproducibility in survey creation, and to delineate their current technological limitations and open research challenges. By doing so, we aim to align the discussion with the overall survey objectives of providing actionable insights and benchmarking state-of-the-art methodologies.

Automated systems for literature surveys leverage algorithmic approaches—primarily utilizing natural language processing (NLP), information retrieval, and machine learning techniques—to streamline stages such as literature search, article screening, data extraction, and synthesis. While these systems offer substantial efficiency gains, they often face challenges related to semantic understanding, context-awareness, and the heterogeneity of scientific writing. Hybrid systems seek to combine automated components with expert-in-the-loop mechanisms, relying on human judgment for nuanced tasks where automation is prone to error or bias.

A key quantitative goal in this domain is to reduce the manual labor required for comprehensive literature reviews by a measurable margin (e.g., percentage of articles screened automatically with high recall and precision), without sacrificing analytical depth or robustness. Effective systems also aim to minimize false positives and negatives, particularly during article selection and classification.

Recent comparative studies and frameworks highlight that most existing automated solutions excel at initial information retrieval but encounter diminishing returns when required to perform advanced synthesis or critical appraisal. Hybrid approaches partly mitigate this bottleneck by creating feedback loops between automation and expert review, but integrating and calibrating these systems remains an ongoing challenge.

Explicit limitations of current systems include limited generalizability across domains, struggles with ambiguous or multidisciplinary terminology, and a lack of transparent decision-making processes. Many frameworks operate as black boxes, which hampers adoption in settings demanding interpretability or auditability.

Open Research Challenges: 1. Enhancing contextual comprehension within automated pipelines to accurately interpret complex, nuanced scientific arguments across disciplines. 2. Developing robust evaluation metrics for benchmarking hybrid systems, including quantitative measures of time-savings, recall, precision, and quality of final survey synthesis. 3. Ensuring transparency and interpretability in automated decision-making processes to foster user trust and facilitate integration into expert workflows. 4. Addressing biases inherent in algorithmically curated literature corpora, relating to publication source, language, or disciplinary scope. 5. Designing adaptive systems that can be efficiently tuned or trained for emerging domains with few available high-quality exemplars.

Summary and Synthesis:
Automated and hybrid survey systems continue to redefine the landscape of scientific literature reviews by improving speed, repeatability, and scale. Despite significant advancements, several challenges—including generalizability, transparency, and integration with human expertise—remain open. Addressing these will not only enable more effective and trustworthy reviews, but will also support the broader objectives of accelerating discovery and promoting evidence-based synthesis across ever-expanding scientific domains.

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

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

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

Section Objectives and Scope. This subsection surveys recent AI- and NLP-driven advances in scientific knowledge recognition and the automation of literature review generation, with a focus on scalability, reproducibility, and domain adaptation. It seeks to clarify: (i) the state of the art in automated survey and systematic review tools; (ii) key technical and usability limitations; (iii) the degree of integration with established reporting standards; and (iv) unresolved challenges and actionable research directions most relevant for broader adoption.

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 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—including AutoSurvey, SurveyX, and Survey-Forge—represent the current state of the art. These platforms leverage large language models (LLMs), retrieval-augmented generation (RAG), and advanced knowledge base structuring (such as Attribute-Tree) 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].

Comparison with Existing Surveys. Recent work [12, 21, 26, 100] provides broader overviews of AI-based systematic review automation, highlighting the progressive integration of machine learning classifiers, active learning, and semi-automated workflows. Unlike these efforts, systems such as AutoSurvey and SurveyX emphasize LLM-centric retrieval and generation pipelines coupled with semantically structured knowledge bases, enabling more sophisticated outline management and citation validation. However, as with prior surveys, real-world applicability remains bounded by gaps in usability, inconsistent terminology, and a lack of robust open datasets and benchmarks [12, 21].

As illustrated in Table 4, these systems embody diverse methodological choices and design trade-offs:

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 [105]. 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 closer to human quality [65]. However, greater breadth of coverage often occurs at the expense of granularity or interpretability in synthesized outputs [65]. SurveyForge further advances through outline heuristics and memory-driven content writing but remains constrained by evaluation heterogeneity and the challenge of achieving deep knowledge synthesis across references [109].

Automated scientific knowledge recognition extends beyond literature review to encompass 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 remains insufficiently addressed within contemporary automated frameworks [76]. Preliminary features such as glossary extraction and entity annotation have emerged, but systems that reliably model temporal and cross-disciplinary semantic drift are notably lacking, motivating future research on adaptive, dynamic ontologies [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—via iterative relevance feedback or retrospective validation—yields higher accuracy and trust than fully autonomous solutions [11, 12, 15, 50, 52, 91, 93, 113]. The establishment of open benchmarks and standardized reporting frameworks, such as PRISMA and GRADE, is essential; these standards are foundational for reproducibility, comparability, and fair evaluation—especially as AI-driven tools diversify in architecture and domain application [6, 12, 61, 94, 110]. Despite such initiatives, adoption and harmonization remain stymied by inconsistent terminology, fragmented evaluation metrics, and the lack of universally accessible annotated datasets [11, 12, 91, 110].

Leading NLP and AI techniques—including neural text classification, active or weak supervision, and embedding-based retrieval—have achieved impressive technical results but fall short in usability, transparency, and accessibility for non-technical end users [14, 26, 49, 91, 111, 113]. Many tools are optimized for technically proficient researchers and lack interpretability features such as rationale extraction or explainable AI components, hindering adoption in critical fields like medicine and interdisciplinary research [4, 15, 48, 111]. In summary, while contemporary systems provide substantial progress in automating evidence synthesis and content organization, broader adoption requires increased focus on interpretability, strict adherence to open standards, and robust human oversight. These recalibrations are imperative to counteract risks of bias, misclassification, or erosion of methodological rigor [11, 14, 15, 26, 49, 74, 76, 94].

Summary of Section Takeaways. In sum, AI-driven automation has advanced literature review scalability and citation accuracy—but faces persistent challenges: (1) The tension between breadth, granularity, and interpretability; (2) Limited semantic understanding of scientific terminology evolution; (3) Usability and transparency gaps for end users; (4) Fragmented standards and benchmarks limiting fair comparison; and (5) Dependency on human expertise for robust quality assurance. Addressing these will require: a) hybrid systems explicitly integrating human validation, b) enhanced dynamic knowledge modeling for terminology and concept drift, c) universal benchmarks and harmonized evaluation criteria, and d) improved, explainable interfaces adapted for diverse research communities.

Open Research Challenges. Remaining research avenues and actionable challenges include: - Developing scalable, fine-grained reasoning systems that can synthesize, organize, and contextualize knowledge at human-expert granularity while maintaining traceable citation logic. - Constructing adaptive ontologies and benchmarking platforms that model semantic and terminological evolution across both time and disciplines. - Advancing explainability capabilities (e.g., rationale extraction, transparent entity linking) accessible to non-technical users, with particular emphasis on high-stakes domains. - Establishing universally adopted, open benchmarking datasets and reporting frameworks for AI-driven survey tools, fostering comparability and trust. - Integrating robust human-in-the-loop mechanisms at all stages of automatic evidence synthesis to mitigate error propagation, foster accountability, and ensure interpretability and methodological integrity.

Explicitly addressing these challenges is crucial for realizing the promise of AI-powered literature synthesis and ensuring its trustworthiness, transparency, and broad societal impact.

3.2 Agent-Based Recognition Systems in Video and Academic Applications

Section Objectives. This subsection reviews the deployment of agent-based frameworks across video behavioral recognition and distributed academic applications, with particular emphasis on: (1) the architectural components underlying agentic autonomy, adaptability, and collaboration; (2) concrete advances and metrics reported in recent systems; and (3) open research challenges that restrict scalable, trustworthy deployment in data-intensive and high-stakes domains.

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

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 [105]	Citation misalignment; domain adaptation issues (pre-prints oriented) [105]
SurveyX	Hybrid retrieval pipeline; structured KB	Improved reference organization, multi-phase processing [65]	Trade-off between breadth and granularity of content [65]
SurveyForge	LLM with citation validation, open benchmarks	Content coverage optimization, ties to reporting frameworks [109]	Development dependent on inconsistent evaluation standards [109]

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. Notably, SAPPTMF obtains high levels of performance: accuracy 94.5%, precision 91.0%, recall 93.4%, F-score 92.4%, and mean squared error (MSE) 0.09. This attests 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]. While significant potential exists, previous literature highlights persistent challenges including data privacy, robustness against distributed attacks, and the need to produce interpretable outputs understandable by human decision-makers, thus maintaining trust and accountability [11, 15, 50, 110].

Open Research Challenges To accelerate the deployment and trustworthiness of agent-based recognition across video, academic, and healthcare settings, several open research challenges are identified: - *Privacy and Security*: Existing frameworks must ensure privacy preservation and resilience to adversarial threats, especially in medical and survey applications. Adaptive agent communication protocols, robust encryption, and strict access policies are ongoing needs. - *Interpretability and Transparency*: Outputs must be made interpretable for human stakeholders. Current systems often focus on performance metrics, but standardizing explanatory models and feedback processes remains an open question, as emphasized across survey and academic literature [11, 12, 15]. - *Scalability and Modality Integration*: Seamless orchestration of multi-agent workflows over rapidly evolving and large-scale, multimodal datasets remains restricted by both computational and algorithmic bottlenecks. Quantitative evaluation benchmarks and common frameworks are needed to scope progression here. - *Human-AI Collaboration*: Effective integration of human oversight—via transparent suggestions, ambiguity flagging, and intuitive control interfaces—requires further development, particularly for complex multi-label or free-text analysis tasks [50]. - *Ethical and Societal Impact*: Broader adoption necessitates explicit frameworks for ensuring ethical AI behavior, accountability, and fair treatment across diverse user populations.

Summary and Synthesis. Agent-based recognition systems, exemplified by FMER and SAPPTMF, demonstrate substantial promise in both behavioral video analysis and distributed academic/medical applications, achieving strong quantitative results and advancing modular, collaborative intelligence. However, realizing the full benefits of agentic paradigms demands focused research on privacy, interpretability, scalability, and ethical integration. Progress in these directions will be pivotal for trustworthy, scalable, and impactful deployment in both high-stakes and large-scale data environments.

3.3 Intelligent Agent-Based Survey Delivery

Objectives and Contributions: This subsection aims to (1) delineate how intelligent, agent-based architectures improve the automation and adaptability of survey delivery and data management; (2) provide a critical overview of contemporary agentic approaches—including hybrid and decentralized systems—for large-scale, longitudinal, or hard-to-reach populations; and (3) identify and discuss open technical challenges and future directions in the field.

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. Functional examples include WhatsApp Business API-driven surveys, where the architecture leverages message flow automation, branching logic, reminders, and integration with data management tools such as Google Sheets, notably facilitating longitudinal engagement with mobile or hard-to-reach participants [28]. These frameworks support flexible deployment, continuous longitudinal engagement, and context-sensitive adjustments, thereby enhancing data quality and response rates while easing traditional administrative burdens [17][42]. Empirical evidence suggests that automated chat platforms yield higher completion rates and lower operational costs compared to legacy SMS or IVR modalities in refugee and geographically dispersed samples [28].

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, several open challenges remain, which can be further specified as follows. *System integration and interoperability* are persistent obstacles, particularly given the heterogeneity of survey platforms, mobile APIs, and evolving messaging protocols [28]. *Sample attrition and engagement decline* present practical hurdles in longitudinal deployments, necessitating improved retention algorithms, personalized reminder strategies, and multi-modal outreach, as highlighted in deployments with mobile populations [28]. *Technical robustness* remains a critical issue, specifically with regard to the detection and amelioration of message delivery failures, API constraints, and workflow interruptions. On the analytical front, more nuanced *benchmarking versus traditional survey modalities* is needed to rigorously quantify added value across different domains and populations.

Further research should address: (a) the design of robust threshold protocols for amalgamated human/AI coding that maintain data reliability at scale; (b) automated detection and mitigation of technical biases, such as model drift or differential attrition; and (c) the development of interpretable, explainable modules to improve the transparency of agentic decision-making [110][42]. Enhanced real-time monitoring, adaptive sampling, and secure, privacy-aware data management will continue to shape the future landscape of intelligent agent-based survey delivery.

Key Takeaway: Intelligent agent-based survey systems, particularly those leveraging hybrid and decentralized architectures, are demonstrably superior in adaptability and scalability but require ongoing innovation in integration, data integrity assurance, and transparency to fulfill their potential across diverse research applications.

4 Advanced Agent-Based Modeling and Multi-Agent Systems

At the outset, this section aims to (i) systematically review the foundations and recent advances in agent-based modeling (ABM) and multi-agent system (MAS) paradigms, focusing on architectural innovations, coordination approaches, and application domains; (ii) highlight hybrid and decentralized agentic architectures, elucidating their unique contributions and limitations; and (iii) identify and analyze open research challenges, providing detailed insights into current technical barriers and emerging future directions.

4.1 Overview and Scope

Agent-based modeling has become a core methodology for simulating and understanding complex, distributed, and adaptive systems. Recent developments in multi-agent systems have expanded both the theoretical and practical horizons of the field, with specialized agent architectures enabling new forms of interaction, autonomy, and emergent behavior. This section synthesizes literature trends and categorizes the main modeling frameworks, with a particular emphasis on the impact of hybrid and decentralized paradigms.

4.2 Hybrid and Decentralized Agentic Architectures

Hybrid architectures integrate multiple agent reasoning schemes, combining reactive and deliberative capabilities to address the tradeoffs between responsiveness and planning depth. Decentralized MAS introduce coordination mechanisms—such as distributed consensus or auction-based negotiation—to allocate tasks and resources without requiring centralized control. These architectural patterns have found application in domains including swarm robotics, smart grids, and large-scale simulation environments.

Key Takeaway: Hybrid and decentralized agentic architectures offer robust solutions for dynamic and uncertain environments by balancing adaptability, scalability, and fault tolerance. However, they introduce new design complexities, such as in agent interoperability and emergent behavior predictability.

4.3 Open Research Challenges

Despite recent progress, several granular technical and application-specific challenges persist. These open gaps include:

Scalable Coordination Protocols: As MAS scale, achieving consensus or distributed scheduling remains computationally intensive, especially in settings with unreliable communication or heterogeneous agent capabilities.

Interoperability and Standardization: Lack of standard interfaces and semantics across agent platforms hampers integration, both for hybrid architectures that combine distinct reasoning schemes and for decentralized deployments spanning multiple organizations or systems.

Robustness to Adversarial and Non-IID Environments: Many current ABM and MAS implementations assume relatively benign operating contexts, limiting their reliability when exposed to adversarial manipulation, noisy observations, or out-of-distribution tasks.

Evaluation Methodologies: There is a need for systematic benchmark environments and evaluative measures that reflect real-world complexity, enabling reproducible comparison across architectures and approaches.

Cross-Domain Transfer: Effective methods for transferring learned behaviors or coordination policies across domains or agent populations remain nascent, particularly for highly heterogeneous or dynamic environments.

Future research should address these challenges by developing new coordination algorithms with provable scalability guarantees, establishing interoperability standards, enhancing fault-tolerant reasoning for non-stationary contexts, refining evaluation protocols, and exploring advanced transfer learning strategies in MAS.

Key Takeaway: Addressing these open research gaps requires a combination of theoretical innovation and practical benchmark development, with particular emphasis on scalability, robustness, and transferability across heterogeneous agent environments.

4.4 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.

4.5 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

At the outset, this section aims to clarify the crucial workflows, automation paradigms, and AI-driven tools that underpin modern agentic systems. Our objectives are to: (1) provide a structured overview of automation in typical AI writing and agent workflows; (2) highlight the integration and evolution of hybrid and decentralized agent architectures; (3) analyze the spectrum of AI writing assistance, from rule-based systems to fully agentic models; and (4) articulate key open research gaps and future challenges in the domain. By distilling current practices, identifying remaining bottlenecks, and tracing emerging patterns, this section seeks to inform both academic and applied audiences working with or developing sophisticated AI-driven writing workflows.

5.1 Technical Workflows in Automated Writing Systems

Efficient workflow design is foundational to scalable AI writing assistance. Such workflows typically comprise prompt engineering, iterative feedback cycles, content evaluation, post-processing, and system adaptation. Automation is realized at multiple levels, ranging from template-driven approaches to orchestration via multi-agent pipelines. Notably, hybrid configurations—combining rule-based components with learning-based modules—offer practical flexibility but introduce new complexity in coordination and maintenance.

5.2 Architectural Paradigms: Centralized, Hybrid, and Decentralized Agents

AI writing assistance architectures span centralized models, in which a primary agent orchestrates all stages, to more decentralized or hybrid designs that distribute responsibilities across loosely coupled agents. Hybrid architectures may, for example, fuse deterministic controllers with generative language models, balancing predictability and creative flexibility. Decentralized approaches enable scalable division of labor but currently face open challenges in relation to communication overhead, consensus, and robust error propagation mitigation.

5.3 Open Challenges and Future Directions

For widespread adoption, several technical and application-specific open challenges remain. These include: (1) Reliable measurement frameworks for workflow quality and robustness under automation; (2) Fine-grained coordination mechanisms for hybrid teams of symbolic and neural agents, including adaptive delegation and error recovery; (3) Scalable, privacy-preserving data flow protocols in decentralized agentic settings; (4) Task adaptation in fluid multi-agent pipelines, particularly under distributional shift and user-driven customization.

Advancing AI writing workflows will require nuanced frameworks for benchmarking, deeper models of inter-agent negotiation, and robust adaptation strategies spanning both infrastructure and interface layers.

Key Takeaway: Modern AI writing assistance is evolving from monolithic, template-based automation toward more modular, hybrid, and decentralized workflows. To unlock the full potential of agentic systems, future work must address granular technical bottlenecks in coordination, robustness, and application-tailored adaptation.

5.4 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.

5.5 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 demonstrate 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, 27, 47, 58, 108, 110]. For instance, in Amharic headline generation, the combination of model fine-tuning, data cleaning, specialized tokenization, and rule-based post-processing have yielded marked improvements over non-fine-tuned baselines and prior systems [27], while the ML-based triage of Spanish survey item performance efficiently flags problematic items akin to English evaluations [108]. AI-based multilingual modules—encompassing cross-lingual screening, citation recommendation, and structured data extraction—increasingly 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, potentially exacerbating exclusion or misinterpretation [11, 27, 40, 58, 89, 108]. The availability of high-quality, domain-relevant datasets remains crucial, as detection systems trained predominantly on English data can exhibit fairness and explainability issues, such as bias against non-native users [40]. Additionally, as AI-based writing tools become more prominent in education, especially for language learners and non-English speakers, there are emerging concerns around academic integrity and the need for responsible, inclusive tool deployment [11, 47, 58]. Collectively, these findings accentuate the ongoing imperative for improved datasets, domain adaptation, and policy interventions aimed at both technical robustness and ethical inclusivity in multilingual AI research tools.

5.6 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.

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].

5.7 Generative Tools and Policies

Section Objective: This subsection aims to synthesize the current landscape of generative AI writing tools as they pertain to academic writing, with a focus on institutional policy developments, ethical safeguards, and emerging open challenges.

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]. Notably, academic writing instruction, as evidenced by reviews of university practices (e.g., Stanford updating AI usage policies, Middlebury College banning ChatGPT to protect critical thinking, and University of California issuing targeted guidance), demonstrates varied and evolving institutional responses [5]. These policies often reflect a nuanced approach—permitting AI to support non-native speakers and reduce language barriers, while enforcing rules concerning disclosure, restricting AI in summative assessments, and reinforcing academic integrity through 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. However, empirical reviews consistently note that, although AI can enhance aspects such as grammar and style, it does not substitute for comprehensive academic writing education, which is essential for cultivating creativity and critical reasoning [5, 58]. A responsible, balanced integration—rather than blanket restriction or unchecked adoption—is thus increasingly seen as optimal for academic communities' diverse and evolving needs.

Open Challenges and Future Research Gaps:

- Determining the long-term impacts of generative AI tools on learners' critical thinking, creativity, and independent research skills, given current evidence that such skills remain inadequately supported by automation [5, 58]. - Establishing consensus on disclosure and authorship frameworks—particularly as models grow more sophisticated in co-authoring and feedback provision [5]. - Developing robust, enforceable institutional policies that strike an appropriate balance between supporting accessibility and mitigating risks to academic integrity, with greater cross-institutional comparison and effectiveness analysis still needed [5, 58]. - Advancing technical safeguards and ethical assurance mechanisms—such as watermarking or attribution—whose practical effectiveness and user acceptance remain open research questions [110]. - Supporting under-explored research contexts: most current studies and policy exemplars are concentrated in North American and European settings, with limited investigation of needs and outcomes in diverse global educational environments [58].

5.8 Synthesis

Section Objective Recap: This synthesis aims to explicitly connect the thematic survey threads—automated and hybrid workflows, multilingual and inclusive design, advanced citation/evaluation tools, and generative writing assistance—with concrete open research gaps, methodological challenges, and promising future directions in AI-powered research and academic writing.

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. Technological advances enable significant gains in efficiency and inclusivity; however, substantial challenges and nuanced trade-offs persist.

Positive Trends and Case Examples: Automated review tools now reduce systematic review time by up to 96% in some medical domains and accelerate large-scale survey writing[16, 21, 45, 105], while fine-tuned LLMs and inclusive pipelines are supporting applications in under-resourced languages such as Amharic and growing cross-lingual survey evaluation[27, 108]. Integrated platforms that combine AI-driven assistance with peer review, as exemplified in recent cross-cultural studies, foster skill development and higher writing standards without replacing essential human roles[5, 58, 113].

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.

Ethical Safeguards and Institutional Responses: The proliferation of AI tools has prompted a taxonomy of institutional responses: some universities (e.g., Stanford) revise integrity policies to accommodate tool usage, others such as Middlebury ban certain AI applications, and a balanced approach is advocated in systematic reviews to preserve creativity and critical thinking[4, 5]. Current writing courses emphasize human judgement and ethics, with AI positioned as an enhancement, not a replacement.

Open Challenges and Future Research Gaps: - Automation remains uneven: screening and selection see robust AI deployment, but stages such as data extraction, synthesis, and reporting remain bottlenecks due to limited tool usability and transparency, heavy reliance on English, and inconsistent standards[12, 21, 100, 102]. - Detection of AI-generated text is a moving target: effectiveness fluctuates across domains and languages, with bias and generalizability as persistent concerns and a need for more representation in datasets and benchmarks[39, 40]. - Quality assurance and evaluation frameworks for LLM-assisted writing and research agents need broader, more fine-grained, and standardized methodologies across diverse settings and languages[14, 69, 111]. - Multilingual and underrepresented domains require concerted research to close quality gaps in inclusivity, dataset coverage, and adaptation of methods[27, 108]. - Ensuring equitable access, mitigating over-reliance, and providing transparent rationales for AI-aided decisions are critical themes for stakeholder trust and adoption[11, 47–49].

Emerging Directions: Sustained progress depends on: - Standardizing metrics and benchmarks for automation tools and inclusive evaluation. - Developing explainable, robust, and user-friendly systems tailored to non-expert users. - Integrating human oversight and ethical frameworks within both academic and institutional practice. - Expanding foundational research and case examples across less-studied languages and disciplines.

In summary, the continued evolution of both technological frameworks and policy mechanisms is essential to meet the requirements of scholars, educators, and institutions. Sustained attention to oversight, equitable access, inclusive design, and ethical stewardship will shape the productive incorporation of AI in academic writing and research workflows [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].

5.9 Prompt Engineering, Model Optimization, and Specialized Agents

This subsection aims to systematically review advances in prompt engineering, model optimization, and the development of specialized agents. We orient the discussion around: (1) prevailing methodologies and foundational techniques; (2) emergent research directions, including hybrid and peer-review paradigms; and (3) critical knowledge gaps and future challenges within each area. The overarching objective is to provide the reader with an integrated and comparative perspective that clarifies both state-of-the-art practices and pressing open problems.

Prompt engineering has rapidly evolved, with current strategies ranging from few-shot design to automated prompt search. Despite impressive downstream performance gains, there remain substantial challenges in understanding prompting robustness, transferability across tasks, and managing prompt-induced biases. Model optimization efforts encompass algorithmic advances for efficiency, scaling, and specialization, yet the field must address trade-offs between generality, resource consumption, and emergent failure modes. Specialized agents encode domain knowledge or task-specific heuristics to achieve superior performance, but questions persist regarding explainability, safe deployment boundaries, and the complexity-cost balance inherent in agent architectures.

A notable institutional response to these technological advances is the emergence of ethical safeguards at multiple layers, from data governance frameworks to evaluation guidelines with embedded bias and fairness audits. For example, several organizations have instituted red-teaming and post-deployment monitoring practices as case studies in responsible AI development. However, standards vary considerably, and persistent challenges remain regarding transparency, user agency, and global harmonization of safeguards.

While hybrid approaches (e.g., combining learned and rule-based systems) and peer-review inspired evaluation pipelines have been proposed to address reliability and transparency, they introduce drawbacks, including increased system complexity, potential integration bottlenecks, and the risk of human-in-the-loop biases influencing outcome validity. Assessing the trade-offs of these paradigms is essential for future progress.

Boxed summary of open challenges and future research gaps: The current landscape in prompt engineering, model optimization, and specialized agent design reveals several concrete open research questions: What theoretical frameworks can provide robust guarantees or predictability for prompt transfer and robustness across domains? How can model optimization techniques balance efficiency with generalization without exacerbating existing fairness or bias concerns? What principled approaches enable interpretable, scalable, and safe specialized agent architectures in high-stakes domains? Which evaluation protocols or taxonomies of institutional safeguards best standardize transparency, ethics, and accountability for rapidly advancing agent systems? What are the trade-offs between adopting hybrid learning paradigms and increased system complexity, and how can these be mitigated in real-world deployments?

5.9.1 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.

5.9.2 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

This section aims to provide a comprehensive overview of the mechanisms, methodologies, and emerging paradigms underpinning quality assurance, feedback integration, and institutional oversight in AI systems. The objectives are to (1) articulate the current state of quality assurance practices, (2) evaluate dominant peer and hybrid review frameworks, (3) survey ethical safeguards and institutional responses, and (4) delineate open challenges and research gaps that may guide future work.

6.1 Quality Assurance Mechanisms

We review established and emerging quality assurance mechanisms, with special attention to their methodological underpinnings and effectiveness. Conventional validation protocols—such as systematic benchmarking, cross-validation, and adversarial stress testing—are compared against newer automated and hybrid approaches. Hybrid methodologies, which combine automated metrics with structured human review, have gained prominence for their potential to balance efficiency with nuanced assessment. However, counterpoints arise regarding scalability and subjectivity, especially in contexts where human input introduces bias or limits reproducibility.

Summary of Open Challenges and Future Research Gaps: **1. Scalability of Hybrid Approaches** remains an open area, especially as system complexity increases. **2. Mitigating Human Bias** in qualitative reviews demands further methodological innovation. **3. Automation Limits** must be systematically investigated for coverage and false positive rates. **4. Standard Benchmarks and Protocols** are needed for comparative assessment across domains.

6.2 Feedback Integration Strategies

Effective feedback mechanisms are vital for continuous improvement and accountability. Approaches range from end-user feedback loops and crowdsourced assessment to formal peer review and post-deployment monitoring. While hybrid/peer review paradigms are

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

widely adopted, possible drawbacks include reviewer fatigue, inconsistent application of standards, and feedback loop latency. Further, a lack of well-defined taxonomy for integrating disparate feedback sources can hinder actionable insights.

Summary of Open Challenges and Future Research Gaps: 1. Taxonomy Development for feedback mechanisms remains incomplete. **2. Reviewer Fatigue** and inconsistent standards suggest a need for incentivization and process redesign. **3. Integration of Diverse Feedback** into a coherent, actionable framework is an unresolved challenge. **4. Closing the Feedback Loop** in a timely and robust manner continues to be a bottleneck.

6.3 Ethical Safeguards and Oversight

The role of ethical safeguards and institutional oversight has become increasingly central. Numerous institutions have adopted governance frameworks, most commonly through dedicated review boards, risk assessment committees, or compliance auditing bodies. For instance, some case studies highlight proactive revision of policies in response to unintended model behaviors, while broader taxonomies categorize institutional approaches as either prescriptive, advisory, or adaptive. Despite these efforts, gaps persist regarding oversight scope, long-term enforcement, and adaptation to evolving technological landscapes.

Summary of Open Challenges and Future Research Gaps: 1. Enforcement Mechanisms for ethical safeguards require rigorous evaluation. **2. Comparative Studies of Institutional Models** are needed for evidence-based best practices. **3. Dynamic Adaptation** of oversight to new risks and regulatory environments remains an active research area. **4. Comprehensive Taxonomies** for mapping institutional responses are yet to be standardized.

Section Summary: This section has outlined a structured framework for understanding quality assurance, feedback integration, and oversight in AI. By categorizing mechanisms and surfacing open research challenges, the survey aims to guide both foundational and applied research agendas.

6.4 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][17]. 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][42][71][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 [107][108][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 [4][47][11][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 [4][47][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 [4][47][11]. 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.

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.

6.5 Data Quality in Survey Automation

This section examines data quality within the overarching framework of hybrid AI–human survey automation, a core survey objective being to identify how automated and hybrid methodologies shape data integrity, reliability, and equity in large-scale, remote, or rapidly iterated research workflows. The review synthesizes recent advances on dynamic, agentic quality assurance, with a focus on both technical and ethical challenges that emerge when deploying automation at scale.

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.

However, proponents of hybrid and peer review dominant paradigms must also contend with counterpoints: purely automated processing offers unmatched efficiency and scalability, but may struggle to achieve required interpretive nuance or accuracy in open-ended, multilingual, or demographically diverse samples [50, 52, 93, 107, 108]. Conversely, excessive reliance on human review can reintroduce bottlenecks, inconsistency, and higher costs, and may itself be limited by annotator availability or domain expertise [50, 52]. Hybrid frameworks seek to optimize this trade-off by channeling routine or high-confidence tasks to algorithms, while allocating complex or ambiguous cases to human experts. Yet, questions remain around how to set or adapt thresholds for human intervention, best structure reviewer collaboration, and continuously audit for emergent biases or procedural drift [50, 108].

Concrete open research questions in this methodological area include: How can hybrid models dynamically calibrate the allocation between algorithmic and human review to adaptively respond to variations in demographics, language, or response patterns? What composite QA metrics or anomaly detection strategies are most effective for identifying subtle threats to data quality—including annotation bias, technical artifacts, or systematic exclusion? How might feedback from downstream analysis or model deployment inform upstream QA strategies for continuous improvement? What are the best-practices for integrating participant-centered privacy controls without degrading core data quality or introducing new participation biases? How can survey automation approaches be made robust to evolving technical constraints (e.g., API limitations, messaging platform policies) and maintain scientific rigor across diverse research settings [28]?

Together, these advances and ongoing debates 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

This section aims to provide a comprehensive overview of key contemporary issues involving ethical considerations, research integrity, transparency mechanisms, and regulatory frameworks within AI, with a focus on identifying unresolved challenges and highlighting areas for future research.

7.1 Ethical Safeguards in AI

A variety of ethical safeguards are being developed and deployed to mitigate risks associated with AI deployment. These include privacy-preserving methodologies, fairness audits, and bias detection toolkits. Institutions have responded with diverse approaches, such as the development of ethics boards, regulatory advisory groups, and public guideline frameworks. For instance, universities and major technology companies have established internal committees to oversee AI research activities, while governmental agencies continue to evolve policy guidelines to ensure societal benefit and reduce harm.

Boxed Summary: Ongoing open challenges include establishing universally-accepted guidelines that remain adaptive to rapidly developing technologies, ensuring accountability for both developers and users, evaluating the long-term impacts of deployed systems,

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

and closing the gap between aspirational ethical codes and their consistent real-world enforcement.

7.2 Research Integrity and Transparency

Transparency in AI model development is vital for enabling reproducibility, external audit, and public trust. Current open access initiatives and model card standards represent progress, yet underlying model architectures, datasets, and training procedures are not always disclosed. Integrity in the field also requires that conflicts of interest be declared and that negative results be reported more routinely. Existing institutional responses range from the encouragement of open data repositories to the formulation of publication standards in top-tier venues.

Boxed Summary: Key unresolved issues include incentivizing full transparency despite proprietary interests, standardizing disclosure norms across industry and academia, ensuring adequate peer review of complex deep learning systems, and addressing privacy concerns without sacrificing reproducibility or accountability.

7.3 Regulatory Frameworks

Regulation of AI technologies encompasses national and international initiatives aiming to balance innovation and societal risk. Recent regulatory proposals adopt multi-stakeholder perspectives and seek to introduce adaptive, risk-based requirements. Institutional responses span from voluntary compliance programs to binding data governance legislation.

Boxed Summary: Core future research gaps involve the harmonization of diverging regulatory approaches across jurisdictions, assessing the effectiveness of different enforcement mechanisms, calibrating regulatory interventions to specific domains (such as healthcare or finance), and anticipating unintended consequences of stringent or weak regulation.

7.4 Ethics and Academic Integrity

This section surveys the evolving landscape of ethics and academic integrity amid the integration of artificial intelligence (AI) in research and education, focusing on the survey's core objectives: to provide a methodological overview of detection approaches for AI-enabled misconduct, highlight their benefits and drawbacks, and enumerate concrete open research questions for safeguarding scholarly trust. Our framework is structured around three methodological pillars: risks and threats, detection mechanisms, and emerging challenges, with explicit attention to equity, transparency, and adaptability. We further aim to synthesize counterpoints on hybrid detection paradigms and foreground open, actionable research directions.

Risks and Threats. AI technologies intensify pre-existing integrity issues such as plagiarism, now sharply amplified by large language models (LLMs) generating highly convincing academic text [8, 47, 82, 84, 85, 89, 104, 107]. Newer risks include misinformation, biased outputs, hallucinations, and forms of research fraud (e.g., ghostwriting, data fabrication) that are increasingly difficult to detect and remediate [5, 12, 13, 18, 24, 39, 45, 50, 59, 67, 91, 97, 101, 108, 114]. As LLM-authored submissions become indistinguishable from human ones, the effort required to ascertain authorship and content veracity escalates [12, 13, 47, 58, 73, 104, 107]. Widespread propagation of AI-produced misinformation threatens the rigor of scholarship and undermines public trust, across both formal literature and broader dissemination channels [11, 38, 39, 72, 73, 86].

Open Research Questions (Risks & Threats): What new forms of academic fraud may arise as LLMs improve multimodal and cross-lingual output generation? How can integrity protocols adapt to rapidly shifting threat surfaces, especially in interdisciplinary and global contexts? What are the long-term consequences of undetectable AI authorship on the evolution of trust and knowledge validation in academia?

Detection Mechanisms and Multimodality. Addressing these threats, current detection methodologies span three primary modalities: manual expert review for context-rich, nuanced evaluation [47, 57, 79, 84, 91]; algorithmic detection via watermarking, stylometry, and anomaly detection [4, 5, 8, 10, 11, 13, 38, 40, 47, 50, 58, 59, 73, 82, 91, 105, 107]; as well as hybrid peer review—integrating human oversight with automated, multi-factor systems [40, 47, 59, 73, 82]. Hybrid approaches combine the complementary strengths of interpretability and scalability, yet also bring forth trade-offs in bias, explainability, and procedural consistency [12, 13, 38, 40, 47, 73, 107].

Detection mechanisms, as summarized in Table 9, have distinct limitations. For instance, stylometric analyses are vulnerable to paraphrasing, watermark signals can often be obfuscated or removed, and ensemble classifiers may drift as attack and evasion strategies evolve. Notably, recent empirical studies reveal that detection effectiveness degrades against advanced LLMs, especially when facing mixed-authorship, paraphrased, or multimodal content [4, 13, 38, 40, 47, 59, 73, 82]. Ongoing refinement of benchmarks, as well as transparent and standardized cross-validation protocols, is essential for robust performance and comparability [11, 13, 38, 40, 73, 107].

A further concern is the risk of bias and inadvertent harm in detection: many tools disproportionately misattribute AI-writing to

Table 9: Core Detection Approaches for AI-Generated Academic Content: Methods, Descriptions, and Limitations. Effective detection requires hybridization, continual refinement, and attention to fairness across linguistic and disciplinary contexts.

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

non-native English speakers, reinforcing linguistic inequities [11, 13, 18, 38, 40, 47, 82]. Thus, research must prioritize transparent, explainable, and culturally aware detection systems [12, 38, 45, 84, 85, 94].

Open Research Questions (Detection): How can detection methodologies be reliably applied in low-resource languages, and what adaptations are needed for cross-lingual and multimodal academic outputs? What frameworks will balance explainability, performance, and bias mitigation to ensure fair assessments across author demographics? How can detection benchmarks remain relevant as LLM synthesis and evasion tactics evolve in complexity?

Drawbacks and Counterpoints on Hybrid/Peer Review Paradigms. While hybrid approaches address scalability and leverage both algorithmic and expert insight, they risk institutionalizing systemic bias (e.g., against non-native language patterns), false positives, and burden-shifting without due process [13, 38, 40, 47]. Peer review remains susceptible to subtle or sophisticated LLM-generated text, often lacking tools or guidance for assessing AI-influenced submissions [12, 13, 47, 58, 73, 104, 107]. Thus, simply layering detection may not address false-negative risks or equity concerns, highlighting the need for contestability, procedural transparency, and accountability mechanisms [12, 38, 40, 104].

Open Research Questions (Hybrid Paradigms): What are the best procedural safeguards to ensure just, contestable AI-misconduct determinations? How can detection systems be stress-tested for process fairness, with feedback loops from affected communities? What oversight frameworks can harmonize automated, peer, and community review to sustain research integrity?

Emerging Threats and Countermeasures. Innovations in generative AI, especially self-modifying or highly automated agentic systems, enable fabrication of full articles, spurious citations, and entire research workflows [11, 13, 26, 31, 39, 47, 53, 82, 84, 100, 107, 108]. Addressing these novel threats entails both technical and institutional measures: embedding standardized metadata on AI use for provenance, advancing robust traceability protocols, enforcing enhanced publisher and funder policies, and promoting inter-institutional intelligence sharing [5, 8, 11, 20, 39, 40, 47, 50, 58, 59, 67, 82, 91, 97, 101]. Furthermore, contestability, transparency, and interruptibility must be designed as foundational principles for all future academic and research systems [31, 58, 82, 86, 104].

Open Research Questions (Emerging Threats): How can metadata-driven transparency protocols be standardized across publishers, institutions, and disciplines to enable traceable, auditable scholarly communication? What technical and policy approaches will maintain contestability and human oversight, even as agentic and self-modifying systems gain adoption? How can international and cross-institutional frameworks support intelligence sharing, rapid adaptation, and harmonized standards for academic AI governance?

Novel Contributions and Future Outlook. This survey explicitly enumerates core open research questions by methodological area, contextualizes detection strategies with their core limitations, and foregrounds counterpoints and procedural safeguards for prevailing paradigms. We advocate for a new, equity-oriented taxonomy of AI-authorship detection, blending technical innovation, institutional reform, and community engagement as mutually reinforcing pillars of academic integrity.

Consistent, flawless reference formatting, as required for scholarly rigor, is applied throughout.

7.5 Regulation and Standardization

This subsection aims to synthesize the current landscape and future directions of regulation and standardization in AI-assisted research, publishing, and education, with a focus on practical pathways toward responsible, harmonized adoption. We critically compare technical and policy-oriented approaches, situate emerging proposals in the context of real-world complexity, and propose a framework highlighting interoperability, transparency, and adaptive governance as core objectives. Our discussion serves the survey's broader goal of clarifying actionable strategies for balancing the unprecedented opportunities of AI with the sustained integrity and trust essential in academic and educational domains.

Metadata, Traceability, and Publisher Guidelines. A coherent foundation for regulation is the development and adoption of standardized metadata protocols that transparently disclose AI involvement throughout the research workflow [20, 23, 40, 82]. This includes not only specifying tool names, versions, and purposes, but also the sections of manuscripts and operational parameters affected. As highlighted in recent proposals [20], such structured metadata empowers systematic downstream analysis—enabling nuanced investigations into linguistic shifts, disciplinary trends, and correlations with research outcomes or citation impact. Wider adoption by publishers and inclusion at both the author submission and editorial review stages are imperative for building trustworthy datasets that facilitate robust, reproducible studies of AI's influence across contexts. Compared to current, often manual, disclosure practices, this critical technical advance offers standardization and automation but faces challenges in cross-publisher interoperability and the burden it might place on authors and editors.

Watermarking, Detection, and Transparency Techniques. Technical safeguards, such as watermarking (white-box, black-box, and neural variants), support document-level traceability and detection of AI-generated content [8, 11, 40, 47, 50, 91], but each approach bears inherent trade-offs. White-box methods offer clarity yet may be easily bypassed; black-box and neural watermarking improve robustness but can be susceptible to adversarial or paraphrasing attacks. The literature [40] indicates that detector performance depends on model architecture, text domain, and the presence of mixed or paraphrased content, with existing tools often demonstrating variable generalizability and fairness, particularly across different user demographics and languages. Ensemble and hybrid pipelines that combine statistical, behavioral, and metadata-based signals are recommended to enhance detection reliability. Technical advances should be paralleled by open, transparent benchmarking against evolving threat models to ensure continuous, credible validation [11, 12, 20, 21, 25, 73, 84, 91]. However, detection alone remains insufficient without regulatory mechanisms for redress, as adversarial techniques continue to erode single-point safeguards.

Unified Reporting and Regulatory Calls. Amid global variation in institutional and publisher policies, there is strong advocacy for harmonized reporting standards and international cooperation [16, 19–21, 26, 44, 52, 67, 68, 81, 88, 89, 93, 102, 105, 112]. While foundational frameworks exist—notably the EU's Ethical AI, OECD guidelines, and IEEE standards—their operationalization varies widely, leading to substantial regulatory gaps and inconsistencies [4, 14, 16, 19, 20, 26, 67, 69, 89, 91, 93, 102, 105]. A critical comparison with technical approaches suggests that institutional initiatives often lack enforcement and remain aspirational, while developer-led standards can fragment into incompatible silos. To move beyond complexity, practical recommendations include embedding robust impact assessments, routine independent audits, and persistent monitoring as integral workflow steps for both organizations and academic communities [4, 11, 12, 16, 20, 21, 31, 40, 52, 59, 68, 69, 73, 82, 88, 107]. Living review systems and open science initiatives [18, 20, 23, 31, 40, 58, 59, 68, 82, 91] are highlighted not merely as supplements but as necessary adaptive mechanisms—ensuring that regulatory regimes remain current, evidence-driven, and responsive to rapid technological and ethical change.

Proposed Integrative Framework. To structure future progress, we outline a conceptual taxonomy drawing together the survey's findings. First, *transparency infrastructure* encompasses standardized metadata and audit trails, enabling interpretability and accountability across research lifecycles [20, 23, 40, 82]. Second, *technical safeguards* comprise watermarking, detection systems, and adversarial robustness, situated within ongoing open benchmarking regimes [8, 11, 40, 47, 50, 91]. Third, *policy and organizational practices* cover harmonized reporting, regulatory mandates, and continuous impact monitoring [4, 12, 16, 20, 21, 31, 40, 59, 68, 82, 88]. This framework emphasizes that interoperability among these domains—rather than piecemeal improvement—is key to both effective governance and the realization of AI's scholarly and educational benefits.

Open Challenges and Future Directions. Despite notable safeguards, no unified technical or regulatory solution currently provides comprehensive, future-proof protection against evolving threats [13, 47, 73, 82, 104, 107]. Synthesis across both technical and policy proposals reveals that adaptive, multi-layered strategies—incorporating multimodal detection, rigorous metadata governance, harmonized and enforceable international standards, and interdisciplinary collaboration—are urgently required [8, 12, 18, 20, 21, 23, 31, 38, 58, 67, 82, 85, 105]. Practical pathways forward include: (1) establishing baseline metadata and reporting requirements for all AI-involved submissions; (2) mandating periodic independent audits and living review updates; (3) incentivizing transparent tool development and benchmarking; and (4) strengthening global alliances for AI policy and standards development. Only through sustained, integrated action can the opportunities of AI in research, education, and society be realized, while proactively mitigating future risks and upholding the trust on which academic advancement is built.

8 Cross-Domain Applications and Educational Impact

This section aims to examine the diverse applications of AI techniques across multiple domains and to analyze their specific impact within educational contexts. By outlining the breadth and depth of cross-domain influences, we clarify how the surveyed approaches contribute to both theoretical advances and practical developments. Our focus in this section is to provide readers with a structured overview of the key intersections between cross-domain AI innovations and education, thus reinforcing the survey's overarching objective of mapping state-of-the-art capabilities and challenges.

In what follows, we first present a concise framework for understanding cross-domain applications, highlighting major themes and notable examples. We then transition to a focused discussion on the ways these applications have shaped, and continue to shape, educational practices, opportunities, and research priorities. This dual perspective offers both a panoramic and granular understanding, foregrounding the distinct contribution of this survey.

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

This subsection examines the multi-dimensional impact of agentic and AI automation in education, with a particular focus on how these technologies are reshaping curriculum development, educational policy, and research competency building. Our goal is to critically evaluate both the benefits and challenges that AI-driven systems introduce in academic environments, and to articulate how these developments intersect with the broader objectives of the survey: understanding the opportunities and tensions that arise from integrating advanced AI in diverse domains.

Recent advancements have placed the educational domain at the forefront of the transformative promise and complex tensions brought by AI automation. AI-driven tutoring systems, multimodal learning assistants, and open student models (OSM) now play a pivotal role in the personalization of learning and formative assessment. These systems utilize 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 studies indicate that adaptive OSM platforms, when integrated within smart classroom settings and coupled with actionable, visualized feedback, lead to measurable improvements in student engagement and knowledge retention [58, 76].

Nonetheless, widespread adoption of AI in educational contexts raises significant challenges related to academic integrity, equity, and preservation of essential cognitive skills such as critical thinking [5, 17, 42, 56, 58, 71]. While AI-based writing assistance platforms offer automated support ranging from grammar correction to personalized research feedback, growing concerns exist regarding overreliance, diminished originality and argumentation, and ethical dilemmas such as plagiarism [5, 17, 42, 56, 58, 71]. Institutional responses to these challenges vary substantially; for example, Stanford University has updated academic integrity policies, whereas Middlebury College has prohibited AI tool usage in classroom environments, demonstrating the ongoing debate about appropriate AI integration in curricula [5, 42, 56, 58, 71].

To position these developments within a broader conceptual framework, we propose viewing the role of AI in education through a balanced paradigm. Rather than envisioning AI as a replacement for foundational pedagogical practices, the literature consistently advocates for leveraging AI as an assistive resource while safeguarding critical skills in reasoning, creativity, and ethics [5, 42, 56, 58, 71]. Ensuring a positive impact thus depends on sound policy design, continual human oversight, and proactive engagement with issues of transparency, bias, accessibility, and equity [42, 56, 58, 71, 76].

This discussion connects with the survey's overall aims by underscoring the interplay between technological innovation and human-centered educational values. Articulating these goals reinforces the narrative unity of the paper, situating educational impact as both a site of transformative opportunity and ongoing debate in the age of agentic and AI-augmented systems.

8.3 Integrated Workflows

This section aims to elucidate the objectives and core developments underlying the integration of automation paradigms in next-generation academic and research workflows. Specifically, it investigates how technical advances and policy considerations shape the comprehensive unification of survey automation, multimodal

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

AI-based recognition, and orchestration of scholarly activities [9, 16, 24, 40, 42, 43, 58, 60, 76, 91, 103, 110, 112]. The intent is to clarify the benefits, ongoing challenges, and regulatory pathways associated with these integrated workflows, providing a critical perspective for readers new to the topic.

Hierarchical frameworks leveraging domain-specialized AI agents are now used to facilitate extensive automation in literature synthesis, dynamic evidence appraisal, and coordination across multimodal data sources. Technical implementations span agent-based optimization for specific tasks as in MAPSOFT [9], multi-agent simulation for knowledge integration [43], real-time data fusion architectures [60], and semi-automated abstract screening tools like Research Screener, which demonstrably lower systematic review workloads without sacrificing scientific rigor [16]. In the medical domain, machine learning methods have been systematically evaluated for literature screening, enabling large-scale and efficient evidence synthesis, though variability in specificity and comprehensive test benchmarks remain policy-relevant obstacles [112]. Similarly, advanced frameworks support facial emotion recognition across demographics [24], product categorization in official statistics [91], and text generation or detection for academic integrity [40, 58, 76, 103, 110].

Despite these substantial advances, enduring needs must be addressed to realize the full potential of integrated workflows. Technical approaches increasingly focus on improving explainability, adaptive domain transfer, and collaborative evaluation paradigms involving both humans and AI [24, 40, 42, 58, 76, 91, 103, 110, 112]. Regulatory and policy proposals, as highlighted in the literature, stress the practical importance of fairness—including bias mitigation in detection algorithms [40]—as well as transparency, uncertainty calibration, and continuous, in-domain, real-world validation. Notably, progress is being made beyond merely exposing the complexity of these concerns: For example, ensemble detection and human-in-the-loop evaluation strategies are actively recommended as feasible pathways to address detectability, robustness, and equity [40]. In the context of agent-based modeling, concrete guidance is evolving to ensure not only modularity and autonomy but also ethical application and scalable, trustworthy oversight [56, 71].

Thus, the trajectory for cross-domain, agentic systems in science and education will be determined not only by technical and architectural innovations but equally by interdisciplinary regulatory synthesis. This synthesis should move beyond acknowledging complexity to explicit consideration of practical implementation, operational accountability, and sustained alignment with societal and ethical imperatives.

9 Explainability, Human-Centric AI, and Inclusive Systems

9.1 Explainability and Transparency

This subsection examines the current landscape of explainability and transparency within AI-driven survey automation and literature review workflows, focusing on both model-level interpretability and evaluation practices. The objectives here are twofold: first, to synthesize how explainability is approached and evaluated in contemporary systems; and second, to clarify the role of transparency as both a technical and user-centered requirement in the context of automated academic tasks. These considerations align with the overarching aim of this survey to identify not only technological progress but also the specific gaps and needs for trustworthy deployment in high-stakes and interdisciplinary academic settings.

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].

To clarify this evolving landscape, we adopt a taxonomy that distinguishes explainability efforts along three axes: (1) explanation modality (extractive versus abstractive rationalization); (2) evaluation paradigm (automatic metrics, embedding-based similarity, divergence-based approaches, and human annotation); and (3) application context (general academic summarization, clinical review automation, bibliometric analysis). This structure, reflected in recent surveys [12, 48, 49], underscores the interplay between technical progress and practical deployment requirements across diverse tasks.

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 tracking citations, coverage, and content relevance using multi-metric 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

Section Objectives. This section aims to synthesize major challenges and current approaches regarding equity, bias, and fairness as automated and AI-driven systems become foundational to survey research and academic workflows. Our goals are to (a) clarify how algorithmic and data-related disparities arise and persist in these contexts, (b) survey both technical and organizational measures intended to mitigate them, and (c) propose a conceptual lens through which related efforts—such as standardization, contestability, and fairness evaluation—can be systematized and advanced across diverse domains and populations. We specifically connect these themes to the overarching objective of the survey: advancing equitable, credible, and trustworthy AI-augmented academic and survey practices.

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] [7

Conceptual Synthesis. Research suggests that progress toward equity requires addressing three interlocking dimensions: (1) *Standardization*, such as developing transparent reporting protocols and metadata for AI use [20]; (2) *Contestability*, referring to mechanisms that allow stakeholders to challenge, review, or audit algorithmic outputs [104] [8]; and (3) *Fairness*, encompassing the proactive identification and mitigation of bias, as well as the inclusive representation of language and demographic diversity [105] [18]. A holistic approach recognizes that technical, procedural, and community-driven safeguards jointly underpin more accountable and equitable AI in both survey and academic settings.

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
Targeted interventions such as fine-tuning and rule-based postprocessing enable tangible improvements for specific under-resourced languages (e.g., news headline generation in Amharic [27]), but persistent hurdles include data scarcity, annotation inconsistency, and limited 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, underscoring the need for robust benchmarks and coverage across user communities [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
Systematic reviews highlight the limitations of English-centric datasets [40], the importance of fairness and explainability, and the call for adaptive, robust, and uncertainty-calibrated detection—especially as text generation systems evolve [39] [40]. Accordingly, the literature 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].

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

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

Inclusive benchmarking and contestability are advancing as fundamental principles. Models are no longer assessed solely on canonical datasets; recent work highlights the value of contestability and robust, standardized procedures, not only for fairness but also for system accreditation and policy trust [104] [8]. Increasingly, models are evaluated across resource-limited, domain-diverse, and demographically varied corpora, which surfaces disparities and reduces the risk of exclusion [97] [105] [53] [14] [8] [101] [18] [59] [73] [50] [107] [15] [4] [27] [20]. Achieving equitable AI thus depends on ongoing interdisciplinary collaboration, community-driven standards, and iterative, impact-driven improvement that integrates standardization, contestability, and fairness into unified assessment and governance frameworks [39] [86] [18] [45] [12] [10] [50] [108] [103] [20].

Section Recap. In summary, equity, bias, and fairness are not isolated technical issues but require a systematized approach integrating standardization, contestability, and fairness-aware evaluation. Connecting these themes to the survey's overarching objectives, we underscore that responsible AI in survey research and academia can only be advanced through technical improvements, transparent governance, and inclusive, community-driven practices.

9.3 Human-Centered and Contestable Systems

This section aims to synthesize emerging objectives and concepts in the design of AI-powered survey and academic automation systems, with a focus on explainability, contestability, human-centeredness, and equity. Our goal is to clarify how contestability and transparency can be operationalized as core pillars of responsible automation, and to establish their connections to the broader objectives of trustworthy, standardized, and fair academic and survey workflows. We recap state-of-the-art methods, highlight challenges, and propose directions for integrating these principles into unified, sustainable frameworks.

A prevailing theme in contemporary research is the imperative to design AI-powered systems that are not only transparent and equitable but 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 [4, 10, 12, 15, 20, 27, 39, 40, 45, 73, 107].

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 [10, 12, 20, 27, 39, 40, 73]. 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 [12, 27, 40, 45, 73]. 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 [12, 27, 45, 70, 94].

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 [15, 73, 107]. 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 [12, 15, 39, 73, 107].

Community-driven strategies, rooted in stakeholder engagement, iterative co-design, and open-source tool development, further empower users and democratize oversight [12, 18, 20, 27, 40, 45, 86]. 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 [12, 27, 40, 45, 73]. 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 [10, 20, 39, 40, 45, 50, 103, 108].

Synthesizing these insights, we observe that contestability in AI-powered survey and academic systems is not an isolated objective, but inherently linked to standardization, transparency, and fairness. Robust contestability frameworks, such as those leveraging standardized metadata for AI usage [20], audit trails, data-centric evaluation protocols, and open-source benchmarks [12, 40], enable not only operator oversight but also increase methodological reproducibility and fairness across applications [39, 40]. Likewise, open, standardized evaluation fosters comparability and trust [12, 73], while community-driven development ensures inclusion of diverse values and needs [12, 20, 86].

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

This section aims to clarify the essential goals and challenges surrounding standardization, interoperability, and collaborative benchmarking within AI systems, particularly as they relate to the overarching objectives of the survey: fostering trustworthy, transparent, and accountable AI deployments. Specifically, this section addresses: (1) the definition and role of standardization in enabling rigorous, reproducible AI research and development; (2) the significance of interoperability for facilitating integration and comparison across frameworks, platforms, and models; and (3) the development and adoption of collaborative benchmarking practices, which serve as critical enablers for community-driven progress and comparability in evaluation. By synthesizing these interconnected topics, we emphasize their collective necessity for building robust, transparent, and contestable AI systems.

Standardization refers to the formulation and adoption of shared protocols, interfaces, and metrics that ensure consistency and repeatability across the AI development lifecycle. The proliferation of diverse models, datasets, evaluation methodologies, and terminologies underscores the need for commonly accepted standards. Effective standardization enables systematic assessment of algorithmic performance, mitigates fragmentation, and lays the groundwork for trust through independently verifiable outcomes.

Interoperability encompasses the technical and procedural capacity for AI tools, components, and datasets to function cohesively across heterogeneous ecosystems. It is critical for enabling researchers and practitioners to compare models fairly and integrate advances without excessive reengineering or bespoke adaptations. Achieving interoperability relies on the availability of open formats, documented APIs, and adherence to community-endorsed standards, which collectively minimize barriers to entry and foster inclusive participation.

Collaborative benchmarking entails the collaborative design, curation, and maintenance of shared benchmarks, ranging from datasets to evaluation leaderboards. Robust benchmarking initiatives underpin empirical rigor by enabling transparent comparison, contestability of results, and the detection of shortcomings in real-world performance. These initiatives often require broad community engagement to ensure representativeness, fairness, and sustained relevance.

The unification of standardization, interoperability, and benchmarking not only addresses technical barriers but also strengthens the broader ecosystem's capacity for contestability and fairness. Establishing a high-level conceptual synthesis, we propose treating these processes as mutually reinforcing axes within a framework for trustworthy AI: standardization provides the foundational language and infrastructure; interoperability ensures seamless integration and iterative improvement; and collaborative benchmarking offers the transparent, empirical backbone for comparative evaluation and contestable claims. This triadic framework emphasizes that progress toward trustworthy, fair, and accountable AI depends on the concurrent advancement and integration of all three thematic pillars, which together enable reproducibility, transparency, and societal alignment across the AI landscape.

In the following sections, we elaborate on these themes, offering conceptual synthesis while explicitly connecting each thread to the central aims of contestability, fairness, and trustworthy AI. This approach ensures a cohesive narrative and provides readers with a clear understanding of both the importance and interdependence of these critical concerns.

10.1 Protocols and Systems Standards

This section examines the landscape of protocols and systems standards in academic and survey automation, situating their significance within the broader objectives of scalable, reliable, and fair automation as surveyed throughout this paper. Specifically, we synthesize evidence on how persistent fragmentation in tooling and reporting undermines reproducibility, transparency, and comparability, and advocate for unified frameworks capable of ensuring contestability and fairness in automated scholarly and survey processes.

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].

Conceptual Synthesis and Towards a Unified Framework. Drawing from cross-domain survey findings, three core dimensions underpin the development of credible, contestable, and fair automation standards for academic and survey workflows: (1) *Standardization*—encompassing consistent data formats, evaluation metrics, metadata fields, and reporting checklists; (2) *Interoperability and Modularity*—supporting seamless integration across workflow components; and (3) *Contestability and Fairness*—ensuring transparent provenance, reproducible methods, and processes amenable to independent scrutiny and iterative revision.

We propose an integrative perspective: effective protocols and system standards should unify methodological rigor (as seen in frameworks like PRISMA, AMSTAR-2, and responsive/adaptive survey design [19, 68]), technical interoperability (such as modular APIs, standardized metadata as advocated in recent studies [11, 20]), and explicit measures for contestability (providing transparent documentation of automated steps, parameters, AI tool versions, and provenance [4, 11, 20, 68]). This tripartite synthesis can act as a guiding taxonomy for evaluating and advancing automation systems, supporting not just function and efficiency, but also critical engagement, replicability, and fairness.

To address the present 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, 20]. 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, advancing toward unified protocols and system standards is central to achieving the survey's overarching objectives: trustworthiness, scalability, and fairness in academic and survey automation. Persistent deficiencies in reproducibility, interoperability, and protocol harmonization reaffirm the necessity of multi-dimensional, community-wide efforts—integrating technical, methodological, and contestability principles—to create transparent, reliable, and adaptable workflows [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

This section examines how the development of open datasets and collaborative benchmarking practices serve as foundational pillars for standardizing, contesting, and ensuring fairness in automated literature review and survey systems. Our aim is to illuminate not only the current landscape but also conceptual linkages to the overarching objectives of automation in survey research: robust reproducibility, methodological transparency, and ethical integrity across domains. Establishing communal resources and shared standards directly addresses siloed tool development and inconsistent metric adoption, and is essential for paving the way toward trustworthy, generalizable, and fair automation practices [12, 20, 61, 110].

Central to resolving challenges of fragmented automation ecosystems is the sustained growth of openly accessible datasets and the institutionalization of cross-community benchmarking. Recent advances in machine learning for question generation, automated assessment, and literature synthesis have been tightly bound to the presence of large, high-quality datasets [6, 11, 12, 20, 21, 40, 43, 46, 59–61, 74, 77, 81, 94, 110]. Datasets such as SQuAD, MS MARCO, RACE, SciReviewGen, and BigSurvey have catalyzed not only technical development but have also enabled broad participation, reproducibility, and comparative evaluation [11, 21, 25, 59, 77].

However, a conceptual synthesis reveals critical gaps: systematic reviews find that open datasets remain concentrated in a few domains (e.g., biomedicine and computer science), while areas such as multimedia question generation or large-scale cross-disciplinary automation are substantially underserved [40, 43, 46, 60, 74]. This lack of domain and modality diversity curtails fairness, transferability, and methodological contestability: learned models may exhibit performance that does not generalize outside their source domain or language [20, 21, 40, 46, 59, 61, 110]. Core survey goals—such as equitable evaluation and transparent contestation of methods—thus require not just dataset expansion but principled collaboration around multilingual, multimodal, and cross-domain resources, ideally in tandem with open methods and workflow sharing [20, 61, 77, 94].

Synthesizing observed trends, we propose a unifying framework for communal resource building in literature review automation, anchored around three main axes:

Standardization: Development of domain-independent, multilingual, and multimodal datasets, together with the adoption of

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

agreed-upon protocols for data creation and annotation, supports generalizability and mitigates bias [20, 40, 110].

Contestability: Establishment of open leaderboards, multi-metric benchmarking suites, and transparent reporting standards enables rigorous comparison and fosters critical assessment of methodological strengths and weaknesses [12, 46, 77, 81, 110]. This facilitates iterative improvements and helps avoid overfitting to idiosyncratic or static benchmarks [6, 110].

Fairness: Institutionalizing community-driven challenge platforms, incorporating metadata describing AI tool usage, and promoting fairness-aware evaluation help make systems accountable to diverse stakeholders and end-users [11, 20, 40, 110]. Such efforts address bias, transparency, and explainability—each vital for trust in automated academic tools.

Current best practices increasingly mandate multi-dimensional evaluation (recall/precision, content coverage, generative fidelity via standardized metrics such as ROUGE or Mauve), use of living datasets and APIs, and adoption of community-led benchmarking [6, 12, 20, 21, 46, 77, 81, 94, 110]. Leading consortia advocate "living reviews" and harmonized metadata standards [6, 20, 74, 81], operationalizing persistence and continuous contestability as scientific fields evolve. Critically, innovative proposals for structured metadata to capture AI tool usage in academic writing (for instance, tool name, version, and section assisted) offer a path to large-scale, automated analysis and improved fairness [20].

Table 14 provides a comparative synthesis of key conceptual challenges and coordinated collaborative responses, mapping these to the axes of standardization, contestability, and fairness that underpin progress toward automation goals.

In sum, the trajectory of robust, trustworthy, and reproducible academic automation is contingent on collective stewardship and principled synthesis of open resources. Advancing toward scalable, equitable, and critically evaluable survey and review systems requires concrete action across standardization, contestability, and fairness axes [6, 11, 12, 20, 21, 40, 43, 46, 59–61, 74, 77, 81, 94, 110]. These efforts underpin scientific integrity, transparent comparison, and meaningful cross-disciplinary impact, directly serving the core objectives of survey automation research.

11 Limitations, Challenges, and Future Prospects

This section aims to critically evaluate the boundaries and difficulties currently faced in the field, while also forecasting avenues for future inquiry. These goals align with the overarching objectives of this survey: to systematically map the current intellectual landscape, highlight unresolved issues, and provide conceptual guidance for ongoing and future research.

To move beyond descriptive recitation, we synthesize our discussion through the intersecting dimensions of *standardization*, *contestability*, and *fairness*, which jointly underpin many contemporary challenges and opportunities in the domain.

Standardization: A persistent limitation in the field is the lack of universally accepted benchmarks and protocols, which impedes both comparative evaluation and meaningful replication. Standardized datasets, clear definitions of key constructs, and shared evaluation metrics are essential for cumulative scientific progress. Variation in methodological approaches can hinder meta-analyses and the generalizability of findings, underscoring the need for community-driven standards.

Contestability: As models and systems become increasingly complex, ensuring their contestability—i.e., providing mechanisms for users and stakeholders to challenge decisions and outputs—remains technically and ethically challenging. Many existing solutions are ad hoc or context-dependent, and there remains no unified theoretical framework for contestability. Addressing this challenge calls for the development of more robust models of user interaction, transparent documentation, and responsive feedback systems.

Fairness: Equity and impartiality are recurring yet still insufficiently resolved themes within the field. Despite numerous proposed metrics and interventions, it remains difficult to achieve fairness that is both technically precise and contextually appropriate. Tensions frequently arise among different fairness criteria, exposing the need for a more nuanced understanding of trade-offs and a robust taxonomy of fairness definitions and their applications.

To synthesize these intertwined dimensions, we propose a unifying conceptual framework characterized by three foundational pillars: *Standardization as Foundation*, necessary for methodological comparability and reliability; *Contestability as Process*, facilitating accountable and user-centric systems; and *Fairness as Outcome*, ensuring equitable impacts and distributions. Positioning these elements within a single taxonomy provides a holistic view, enabling researchers and practitioners to systematically identify progress, gaps, and priorities.

Looking forward, the continued evolution of the field depends on: (a) collaborative efforts to define and adhere to standardized methodologies; (b) the development of systems enabling multi-stakeholder contestability; and (c) sustained engagement with the theoretical and practical complexities of fairness. Integrating these priorities can help advance the field from incremental improvement to transformative impact, serving both scientific and societal goals.

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	Domain expansion, cross-domain dataset curation, and standardization protocols
Limited Diversity/Multimodality	Scarcity of multilingual and multimedia datasets	Multilingual/multimodal benchmarks, open data annotation guidelines
Fragmented Evaluation	Disparate metrics and bespoke tasks	Multi-metric benchmarking suites, open leaderboards, contestability mechanisms
Tool Comparability	Inconsistent reporting and transparency	Standardized reporting, community-led challenge platforms, metadata conventions
Benchmark Stagnation	Static datasets lagging behind field developments	Living datasets, APIs, continuous integration, and persistent benchmarking
Fairness and Transparency	Underdocumented AI usage, model bias, and explainability gaps	Metadata standards, fairness-aware evaluation, transparency-focused protocols

11.1 Barriers and Open Challenges

This section outlines the main barriers and open challenges that must be navigated for the continued advancement and responsible adoption of academic workflow and survey automation. The explicit objectives here are: (1) to synthesize the major technical, methodological, and ethical impediments identified in the literature, (2) to connect these obstacles to the broader aims of standardization, contestability, and fairness in research automation, and (3) to propose a conceptual framing that underscores persistent cross-cutting themes affecting the field. In line with the overall goals of this survey, this synthesis aims to guide future research and practice towards effective, equitable, and robust automated academic systems.

Despite rapid progress in automating systematic reviews and research tasks, substantive barriers persist across several interconnected domains. Chief among these challenges is the **variable quality and accessibility of research data**. As numerous studies emphasize, insufficient metadata, non-representative corpora, and paywall-restricted or non-standardized sources constrain 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]. Existing advanced AI tools—notably large language models (LLMs) and domain-specific classifiers—are often trained or evaluated on narrowly scoped datasets, compromising replicability and hindering cross-domain integration. As the citation summaries reveal, these challenges are particularly acute in survey automation, where persistent bottlenecks include **dataset scarcity, high attrition, and technical dependencies on proprietary APIs** [28].

Alongside data-related limitations, **workflow opacity and lack of interpretability** remain significant obstacles. Although transformer-based and neural models have advanced the state of the art, their intrinsic complexity limits transparency, especially for practitioners and non-expert end-users. This opacity impedes trust and adoption, particularly in fields subjected to high scrutiny, 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 scarcity of open, standardized benchmarks and accessible interfaces further compounds the difficulty for stakeholders seeking reliable, contestable, and user-centered automation solutions [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]. The need for open science measures, transparent evaluation, and systematic documentation emerges consistently across the literature [12, 20].

A recurring theme is the **computational demands and limitations in scalability**. Deep learning and graph-based methodologies—while powerful [38, 106, 115]—require significant compute resources, particularly during large-scale training and inference [5, 6, 12, 15, 18, 19, 31, 38, 40, 42, 45, 50, 52, 56, 58, 59, 69, 71, 76, 79, 81, 82, 85, 86, 88, 89, 91, 93, 103, 107, 108]. This intensifies inequities between resource-rich and resource-limited institutions. Although lightweight, more efficient solutions are being explored [56, 94, 103, 110], their adoption remains limited, and the challenge of generalizing such models without compromising accuracy persists.

Tensions between **openness and privacy** are also unresolved, especially where sensitive information—such as peer reviews, survey responses, or author attribution—is involved. These situations raise intertwined issues of privacy, fairness, and algorithmic bias [11, 15, 20, 28, 42, 45, 47, 56, 58, 71, 73, 85, 98, 107]. Empirical evidence indicates that automated systems often reflect or exacerbate preexisting linguistic, demographic, or geographic biases, with non-dominant languages and diverse author profiles being particularly impacted [15, 17, 27, 31, 42, 56, 58, 71, 78, 82, 89, 107, 110]. The lack of standardization and contestability mechanisms for auditing these impacts remains a key research direction.

With the rapid proliferation of AI-generated content, the field now faces mounting pressures for robust **oversight and fraud prevention**. Standard detection techniques—including watermarking, stylometric analysis, and metadata-based forensics—struggle to keep pace with evolving adversarial tactics and increasingly sophisticated content generation models [12, 28, 45, 56, 71, 73, 110]. Detection effectiveness is often variable across domains and languages, and can introduce or reinforce inequities—for example, disproportionately penalizing non-native users or those working in under-resourced languages [17, 40, 42, 47, 71, 82, 89, 107, 110]. The limitations of common fraud detection approaches are comparatively summarized in Table 15, highlighting ongoing needs for adaptive, equitable, and multi-layered detection frameworks [40, 82].

A further critical gap is the **long-term and cross-cultural validation** of AI-powered academic workflows. Most empirical assessments remain concentrated in Anglophone or biomedical contexts, with little attention to under-resourced languages, varied academic cultures, or longitudinal ecosystem impacts [27, 28, 58]. Meanwhile, rigorous evaluations of downstream effects—especially on research quality, user adoption, and knowledge dissemination—are scarce [28, 58, 98], undermining transparency and evidence-based development.

Finally, **survey automation** introduces distinct barriers: high attrition rates, intricate technical setups, proprietary dependencies,

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

and inconsistent data completeness all hinder robust and representative automation pipelines [28, 85]. While these tools can substantially reduce manual burden, they frequently sacrifice sensitivity to rare or critical signals, strengthening the case for maintaining human oversight in decision-critical domains [28, 58].

Conceptual Synthesis and Unifying Perspective: The challenges outlined above converge around ongoing issues of standardization, contestability, and fairness—themes that surface repeatedly in both technical and ethical discussions. They reinforce the need for a unified research effort grounded in cross-disciplinary and cross-cultural synthesis, systematic benchmarking, and continuous attention to the downstream consequences of automation. The path forward requires building systems that can balance openness with privacy, efficiency with inclusivity, and automation with transparency—operationalizing fairness and contestability not simply as afterthoughts, but as foundational design principles.

In summary, the complex interplay of data quality, interpretability, scalability, oversight, and longitudinal validation continues to define the open challenges in automating academic and survey workflows. Addressing them in concert is critical for fulfilling the overarching goals of this survey: advancing robust, equitable, and accountable automation across diverse scholarly domains.

11.2 Opportunities and Future Directions

Objectives Recap: This survey has aimed to clarify the evolving landscape of AI-driven automation in systematic reviews and academic workflows, highlighting key methods, taxonomies, opportunities, and open research challenges across modalities, languages, and academic domains. In this section, we synthesize future directions and actionable impact metrics while explicitly linking back to these foundational objectives.

Amidst the current challenges, compelling, measurable opportunities are emerging to enable **innovative and empowering solutions** for systematic reviews and academic knowledge production. Notably, advances in **interpretable and intelligent agents**—featuring multi-modal, multi-lingual, and cross-lingual capabilities—are increasingly transcending language and domain barriers, making robust AI tools accessible to broader communities, including researchers in resource-constrained or underrepresented environments [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]. For example, the deployment of lightweight LLM architectures in under-resourced languages such as Amharic demonstrates measurable improvements in news headline generation and summarization tasks, with attainable BLEU, ROUGE-L, and Meteor scores matching or significantly surpassing prior systems and making advanced NLP accessible to new user bases [27, 110].

Interdisciplinary innovations in explainable AI are bridging transparency and participation gaps. This is realized through the integration of symbolic reasoning, knowledge graphs, and user-centric feedback, as tools like semi-automated classification and responsive design successfully balance algorithmic efficiency with human oversight—evident in multi-label open survey question coding where manual coding effort is reduced by up to 70% without loss of quality, and in adaptive feedback systems that positively impact researcher integration, especially across linguistic and cultural divides [5, 19, 50, 52, 58, 76, 93, 113].

Concrete progress is being driven by **expansion of collaborative benchmarking and open data resources**. The creation of large-scale, multi-domain corpora (such as BigSurvey and SciReviewGen [59, 69]), open access APIs, and community-driven evaluation platforms underpins rigor, reproducibility, and equity. These benchmarks, which include measurable criteria like standard evaluation metrics and public leaderboards, enable systematic auditing across languages and application domains, and support 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, machine-readable AI usage metadata in publications enhances transparency and supports downstream development and policy [56, 58].

The ambition of **integrative and holistic academic workflows** continues to advance, as highlighted by frameworks that support entire research life cycles: from literature search and planning (using automatic document clustering and keyword co-occurrence network analysis, as applied in pain research and wastewater engineering [79, 85]), through data extraction and synthesis (exhibiting generalization across biomedical, legal, and educational contexts [58, 110]), to reporting and peer evaluation. These advances require robust multidisciplinary collaboration to harmonize efficiency, robustness, and fairness, and are further measured through domain-adapted empirical benchmarks and the documented uptake in communities such as Kazakhstan, Colombia, and under-resourced educational institutions [5, 28, 58, 110, 113].

The development and deployment of **scalable and lightweight architectures** is central to democratizing automation. Evidence from case studies in multilingual survey analysis [5, 27, 108, 110] and mobile/low-resource settings—such as WhatsApp-based survey automation for refugee populations—confirm sustained gains in cost, scalability, inclusion, and completion rates [28]. Furthermore, the use of distributed agents, as in decentralized transport or water contamination monitoring, achieves measurable impact by lowering computational and technical barriers [17, 56, 71].

Looking forward, the realization of **living review systems**—dynamic, AI-enabled platforms providing ongoing, open updates to research syntheses—represents a transformative vision for open science. These systems, enhanced by privacy-preserving computation and adaptive fraud detection, are increasingly evaluated according to transparent update frequency, user engagement, and audit trail provision [6, 17, 27, 28, 42, 47, 56, 58, 59, 70, 71, 110, 113]. The consistent integration of robust fraud detection strategies is critical; layered, adaptive detection protocols are urgently needed to counter fast-evolving threats to research integrity in online surveys, as standalone indicators (e.g., traditional attention checks or IP filters) are no longer sufficient [40, 82].

Use Cases from Underrepresented Domains:
[leftmargin=*,nosep,label=]

- **Amharic Language NLP:** Fine-tuning compact transformers for Amharic news headline generation results in BLEU, ROUGE-L, and Meteor scores substantially surpassing previous methods, and sets a replicable, open-source benchmark for other low-resource languages [27].
- **Automated Survey Data Collection:** WhatsApp-based survey automation enables researchers to reach hard-to-reach or mobile populations, such as refugees, resulting in lower costs and higher completion rates relative to traditional methods. This inclusive approach is scalable to diverse geographies and is supported by open-source frameworks [28].
- **Multilingual Survey QA Evaluation:** Machine learning pipelines for survey item assessment effectively flag problematic questions across English and Spanish, providing efficient triage for large data sets and facilitating fair, scalable quality assurance in multilingual contexts [108].

Concrete Future Directions and Measurable Objectives:
[leftmargin=*,nosep,label=]

(1) *Expand and standardize open datasets and evaluation platforms* to provide consistent, domain- and language-inclusive benchmarks for academic automation tools.
(2) *Advance lightweight, domain-adaptable LLM architectures* with effective instruction-tuning and evaluation in underrepresented and low-resource research areas.
(3) *Integrate explainable AI and user-centric feedback mechanisms* to ensure transparency and participatory auditing in system output and decision processes.
(4) *Deploy multilayered, adaptive fraud and quality control strategies* in online research workflows—moving beyond legacy detection towards community-driven standards verified by empirical impact metrics.
(5) *Develop living, dynamic synthesis systems* with machine-readable accountability trails, aligned with open science expectations and responsive to evolving research needs and user groups.

In summary, although 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, there is clear evidence—across a growing diversity of languages, modalities, and disciplines—that emerging research is charting actionable paths forward. Sustained progress will depend on advancing interpretability, inclusiveness, reproducibility, and multidisciplinary collaboration. Automated systems must remain aligned to their core objective: augmenting, not replacing, the expert human reasoning central to scholarly inquiry.

12 Synthesis, Best Practices, and Conclusion

Survey Objectives and Impact:
At the outset, our survey set out to achieve the following concrete objectives: (1) systematically review and synthesize state-of-the-art advancements in the domain, (2) identify and highlight both prevailing challenges and emerging solutions, and (3) provide a unique taxonomy that offers a clearer conceptual framework for future research and application. The intent was to inform and guide both practitioners and scholars, emphasizing inclusivity across diverse and underrepresented domains. These measurable objectives are reiterated here to establish clarity and alignment for the reader as we transition to synthesizing best practices and drawing conclusive insights.

12.1 Synthesis of Key Findings

Throughout this survey, our analysis has surfaced several critical trends and exemplary practices that have shaped recent progress: - The field has witnessed a marked increase in interdisciplinary approaches, fostering successful collaborations that have enabled advances not just in established areas but also, notably, in less-represented sectors. - Notable case studies include applications in healthcare, environmental sustainability, and education. In healthcare, novel algorithms have driven improved early diagnostics in resource-constrained settings. In environmental sustainability, innovative predictive modeling has been leveraged for efficient resource allocation in agricultural planning. The educational domain has benefitted from adaptive learning models that cater to diverse learner populations. - Cross-cutting challenges such as data scarcity, bias mitigation, and scalability have been addressed through the adoption of transfer learning and transparent model architectures, which are now recognized as best practices.

Our survey's taxonomy organizes these developments into clear categories, serving as a conceptual map to navigate both technical advances and practical deployments in a coherent manner.

12.2 Best Practices

Our synthesis reveals the following best practices for researchers and practitioners: - Adopt modular, extensible frameworks that can accommodate novel data sources and evolving requirements. - Prioritize model transparency and interpretability, especially in high-stakes or regulated environments. - Engage with target-domain stakeholders early on to ensure alignment with real-world needs and to foster inclusivity. - Continuously benchmark against a variety of datasets, including those from underrepresented contexts, to validate generalizability and robustness.

12.3 Case Examples from Underrepresented Domains

In alignment with our inclusivity objective, we highlight several fine-grained use cases from less-explored fields that illustrate both opportunities and the adaptability of surveyed techniques: - In public health surveillance, AI-driven predictive systems have enabled earlier detection of rare disease outbreaks in low-resource regions, overcoming historically limited reporting infrastructure. - Conservation science has benefited from machine learning applications for wildlife monitoring via automated sensor networks, aiding biodiversity efforts in remote habitats. - In local government administration, data-driven policy modeling has improved service delivery for marginalized communities through targeted resource allocation strategies.

12.4 Future Directions and Challenges

While substantial progress has been realized, noteworthy challenges remain. Key open problems include: - Enhancing the fairness and accountability of deployed models across disparate populations. - Bridging gaps in domain knowledge transfer to facilitate more effective cross-context adaptation. - Scaling best practices to address the evolving complexity and ethical requirements of emergent scenarios.

12.5 Unique Taxonomic and Conceptual Contributions

One distinctive contribution of this survey is the introduction of a novel taxonomy that bridges methodological and application-oriented perspectives. This framework clarifies the landscape, providing a structured lens through which emerging researchers can identify gaps and opportunities.

12.6 Conclusion

In summary, this survey has consolidated best practices, identified positive trends, and surfaced both opportunities and obstacles across a wide spectrum of domains. By making explicit our measurable objectives and the impact of these synthesized findings, we aim to offer a valuable resource for ongoing research and practical deployment. Our hope is that the conceptual and taxonomic frameworks outlined here will serve to catalyze further inclusivity, robustness, and innovation within the community.

12.7 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 can erode 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.

12.8 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. In alignment with the overall objectives of this survey—namely, to systematically map the landscape of automated survey, review, and detection technologies—this subsection demonstrates how benchmarks underpin transparent, reproducible, and fair evaluation.

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, gaps persist. 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 typically benchmarked on recall, precision, workload reduction, and scalability. However, considerable heterogeneity continues in evaluation standards, underscoring the ongoing need for consensus methodologies and greater transparency in reporting [17, 42, 50, 56, 71, 76, 89, 93, 107]. This motivates taxonomic frameworks that distinguish between benchmark types (e.g., open-domain QA, structured review evaluation, fraud and integrity monitoring) and encourages best practices for dataset curation and transparent reporting.

Table 17 provides a comparative view of prominent datasets and systems, emphasizing their respective evaluation domains and key limitations. This taxonomic synthesis aims to clarify the state of available benchmarks across the field, identify salient positive trends—such as the emergence of structured, multi-document summarization corpora—and highlight structural challenges that continue to motivate research into diverse, multilingual, and standardized evaluation resources.

13.2 Example Code Listings and Pipelines

This section aims to elucidate the operational objectives and comparative landscape of major computational pipelines deployed for

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

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

literature review automation, survey administration, agent-based monitoring, topic modeling, and AI-generated text detection. By explicitly foregrounding the strengths, limitations, and novel organizational schema developed through our synthesis, we address both practical implementation considerations and broader research needs previously identified in this survey. In doing so, we seek to guide readers less familiar with earlier sections toward an understanding of how pipeline designs and code artifacts serve targeted goals of reproducibility, scalability, and interdisciplinary adaptability.

A diverse spectrum of computational pipelines supports literature review automation and monitoring, often implemented as code artifacts and modular workflows. Inverse style transfer for authorship, for example, exploits large language models (LLMs) to generate (neutral, stylized) pairs, efficiently replicating author style with minimal parallel data. While traditional LLM prompting often fails for rare author styles, inverse transfer data augmentation methods [94] deliver substantial gains for such cases, reframing style transfer as a data augmentation problem and extending the domain applicability beyond frequent styles. This approach, notably open-sourced, marks progress toward practical deployment but remains bounded by the representativeness of training data and the need for continued tuning when targeting synthetic or low-resource styles [7].

Prompt tuning pipelines, such as those leveraging P-Tuning and continuous prompt embeddings, address instability and sensitivity associated with hand-crafted prompt design for LLMs. Methods like P-Tuning have demonstrated empirical improvements in both fully supervised and few-shot scenarios [70], reducing performance volatility. However, these methods require representative prompt datasets and careful tuning to generalize across diverse language tasks, potentially limiting portability without additional retraining [6].

Survey automation pipelines consistently adopt modular architectures supporting tasks such as extraction, ranking, iterative classification, and external system integration. For instance, automated WhatsApp survey administration frameworks combine the WhatsApp Business API, cloud-based data flows, and modular branching logic to enable scalable and interactive data collection, demonstrating improved completion rates and lower respondent costs for hard-to-reach populations [28]. Nonetheless, such commercial interfaces can pose significant initial setup and API-related constraints, while technical robustness and workflow adaptability remain ongoing concerns [91][15]. Similarly, open-source modules in platforms like LiteRev and SciReviewGen support clustering and automatic literature review, but full system deployment may still require substantial domain adaptation and operator oversight [59].

Agent-based monitoring frameworks span simulation environments using BDI agent reasoning to reinforcement learning-based decision policies, allowing for customizable and reproducible monitoring scenarios [93][17][71][56]. A key comparative strength of agent-based models is their ability to capture emergent or decentralized decision processes [71], which traditional equation-based or centralized pipelines cannot easily represent. Recent frameworks employing local Gaussian processes with deep reinforcement learning [17] have improved estimation accuracy and adaptability in spatially distributed domains. However, the increased configurability of multi-agent pipelines can render cross-system benchmarking and code portability difficult, especially when decision logics are coupled to domain-specific input streams or sensor architectures [42][56].

Topic modeling pipelines, primarily employing Latent Dirichlet Allocation (LDA) or supervised LDA (sLDA), enable scalable and explainable survey analysis through open-source, annotated workflows [52][108][15]. While these methods provide increased transparency and facilitate large-scale multilingual and mixed-method analysis [108], the effectiveness and interpretability of models are sensitive to preprocessing quality and the credibility of topic-label associations. Double-coding strategies can boost classification accuracy, but at additional annotation cost [52].

In AI-generated text (AIGT) and fraud detection, the field relies increasingly on ensemble detection strategies, neural watermarking, and specialized classifiers. Despite the proliferation of public code and APIs, empirical studies underscore that detector effectiveness varies dramatically with respect to LLM size, language, input domain, and adversarial manipulation [40]. Methods are particularly challenged by out-of-distribution samples and cross-lingual data; remedies include more diverse, in-domain datasets, uncertainty calibration, and human-in-the-loop configurations, though these are not yet systematically integrated in most codebases.

Across these areas, the principal limitations are: (1) code and standardization fragmentation, which impedes robust benchmarking and reuse, particularly in multi-agent, multilingual, or hybrid commercial/academic deployments; (2) disparities in reproducibility due to opaque external APIs or limited documentation; and (3) the need for more explicit integration of fairness, explainability, and policy-aware modules, especially for pipelines with societal implications (e.g., survey item evaluation [107][108] or fraud detection [40]).

In sum, this section organizes the described pipelines according to their primary function and deployment context, highlighting the comparative strengths and trade-offs that arise from codebase openness, modularity, and adaptability to complex research and operational needs. By analyzing these pipelines in relation to the core objectives of standardized, scalable, and explainable computational survey research, our synthesis aims to both inform practical adoption and foreground avenues for methodological innovation.

13.3 Computational Roadmaps and Reproducibility Workflows

This segment of the survey aims to delineate the essential computational strategies and infrastructure underpinning reproducible and transparent automated review workflows. By focusing on the systematic organization, evaluation, and dissemination of all computational artifacts involved in automated survey, synthesis, writing, and monitoring pipelines, this section targets improved trustworthiness, scalability, and replicability—core objectives of this survey.

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].

By mapping these developments to the overall objectives of the survey, this section enables readers to better understand how computational reproducibility workflows serve as the foundation for reliable, scalable, and credible research outputs in the landscape of automated survey and review technologies.

13.4 Case Studies and Exemplary Systems

This section aims to illustrate the practical goals and operational realities of automated survey, review, monitoring, and detection systems by presenting representative case studies and exemplary systems. Specifically, it highlights how key agentic platforms, domain-specialized and multilingual models, and innovative architectures address persistent challenges in data quality, efficiency, engagement, and integrity. By synthesizing these concrete deployments, we elucidate best practices and persistent limitations, directly supporting the survey's overarching objective: to clarify the state of automated survey technologies and guide future research and application.

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 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]. By surveying these systems and their real-world impact, this section reinforces the survey's aim to bridge technology and application, drawing actionable insights to inform both research directions and practical deployments.

13.5 Metadata Proposals and Policy Guidance

This segment of the survey addresses the concrete aims and targeted outcomes associated with advancing metadata standardization and policy development for AI and agentic tool use in scholarly communication. Specifically, we focus on initiatives that support increased transparency, facilitate rigorous impact analysis, and ensure alignment with the broader survey objectives of promoting reproducibility, trust, and responsible integration of automation in academic research workflows.

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]. By systematically addressing metadata and policy guidance, this survey segment directly supports the overarching goals of transparent reporting, improved research integrity, and a more accountable scholarly ecosystem.

13.6 Research Summary Tables

This section aims to synthesize and critically evaluate key research gaps, comparative results, and strategic directions identified in the longitudinal, cross-cultural, and multi-domain studies reviewed thus far. Our objectives are threefold: (1) to map actionable recommendations alongside persistent challenges for automated survey and review systems; (2) to highlight the original organizational schema used to relate datasets, benchmarking, workflow reproducibility, and human–algorithm integration; and (3) to explicitly connect synthesized recommendations to overarching survey goals regarding transparency, fairness, and interdisciplinary adaptability.

Synthesis across diverse studies reveals persistent gaps and generates 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]. Notably, Zheldibayeva [113] and Jen and Salam [58] emphasize the positive potential of AI assistance coupled with peer or human oversight, but also highlight the limitations inherent in currently available datasets—especially for under-resourced languages and modalities—which impedes robust generalization or fair evaluation. Comparative analyses of system pipelines, as observed in studies such as Erin et al. [42], demonstrate trade-offs between system complexity, performance, and replicability, thereby illustrating the practical importance of harmonized benchmarking and reproducibility frameworks. Furthermore, adopting agent-based and multi-agent modeling perspectives, as surveyed by Delcea and Chirita [71] and Malas et al. [56], provides flexible frameworks for modeling heterogeneity and decentralization in automated review workflows, but these pipelines also require careful oversight to ensure transparency and comparability.

An emergent consensus supports the development of robust, dynamically updatable review systems—such as those pioneered by Fei et al. [28] for longitudinal engagement—capable of evolving alongside changing disciplinary landscapes and data streams. Despite the technical promise of such innovations, limitations like technical setup, attrition, and ethical assurance remain prevalent, especially in sensitive or mobile populations.

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. In this synthesis, novel attention is given to how openness, documented version control, adaptive mechanisms, and modular architectures can coalesce into a more resilient and equitable infrastructure for automated review. 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. In summary, this organizational schema both consolidates comparative strengths and exposes critical limitations of current pipelines, thereby offering a transparent framework for future methodological improvements—which directly tie into the survey's overarching objectives of advancing robustness, equity, and adaptability in automated review systems.

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.

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

[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

[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