Inspectable AI for Science: A Research Object Approach to Generative AI Governance

Ruta Binkyte
CISPA, Helmholtz Center for Information Security

Sharif Abuaddba
DATA61,CSIRO

Chamikara Mahawaga Arachchige
Data61, CSIRO

Ming Ding
Data61, CSIRO

Natasha Fernandes
Macquarie University

Mario Fritz
CISPA, Helmholtz Center for Information Security

***Abstract*—This paper introduces AI as a Research Object (AI-RO) paradigm for governing the use of generative AI in scientific research. Instead of debating whether AI is an author or merely a tool, we propose treating AI interactions as structured, inspectable components of the research process. Under this view, the legitimacy of an AI-assisted scientific paper depends on how model use is integrated into the workflow, documented, and made accountable. Drawing on Research Object theory and FAIR principles, we propose a framework for recording model configuration, prompts, and outputs through interaction logs and metadata packaging. We implement a lightweight writing pipeline in which a language model synthesizes human-authored literature notes under explicit constraints and produces a verifiable provenance record. We present this work as a position supported by an initial demonstrative workflow, arguing that governance of generative AI in science can be approached as a documentation and provenance practice. Based on this example, we outline and motivate a set of necessary future developments required to make such practices practical and widely adoptable.**

1. Introduction

The rapid proliferation of generative artificial intelligence (AI), particularly large language models (LLMs), has fundamentally transformed scientific writing and knowledge production. Across disciplines, researchers increasingly rely on AI systems for ideation, literature review, code generation, data analysis, manuscript drafting, and peer review [18]. This transformation is reflected in the growing number of AI-assisted submissions to academic venues and in emerging institutional guidelines governing AI use in research [10], [31], [34]

Emerging Reality of AI Use in Science. At the same time, this shift has introduced substantial uncertainty regarding authorship, originality, plagiarism, and intellectual accountability [1]. A recent study by [2] identified 100 non-existent citations in recently accepted papers at NeurIPS'25, one of the major machine learning conferences, illustrating how AI-mediated workflows can undermine established norms of scholarly verification. Such incidents highlight broader risks, including fully fabricated studies generated by AI, that challenge traditional expectations of academic responsibility [32]. In response, many journals and conferences now require authors to disclose AI assistance. Yet current disclosure practices remain largely self-attested and describe AI use in coarse categories, such as writing or grammar support, without mechanisms to verify the accuracy, scope, or procedural context of those claims [11], [13], [17], [30]. Statements such as "an LLM was used for language editing" reveal little about how AI shaped the document, whether confidentiality constraints were respected, or how claims of responsible use might be audited without exposing sensitive materials. Moreover, disclosure itself may be stigmatized, discouraging openness and potentially eroding trust rather than reinforcing it [25]. We argue that the extent of automation will grow in the future, indicated by developments in autonomous AI scientists and co-scientists [14], [21]. Therefore, maintaining control and scientific integrity calls for urgent advances in establishing standardized, verifiable, and practical tools for disclosing AI use in scientific research.

Theoretical Approaches to the Use of AI. Theoretical discussions of AI in scientific and creative work largely frame governance in terms of authorship: whether AI should be treated as an author, collaborator, or a tool. A prominent perspective conceptualizes AI as an instrument that extends human capabilities without possessing creative agency [24], aligning with legal positions that prohibit listing AI systems as authors of copyrighted materials [8], [15]. Other scholars argue that AI may function as a legitimate collaborator when it substantively shapes intellectual direction, raising questions about credit and responsibility [9]. A third line of work reframes authorship around creative control, emphasizing human steering of prompts, curation, and iterative refinement [20], while related research highlights psychological dimensions of ownership and engagement [12], [16]. Accountability-centered perspectives further stress that authorship entails responsibility for claims and ethical implications [6]. For example, the Committee on Publication Ethics (COPE) states that AI tools cannot meet authorship criteria

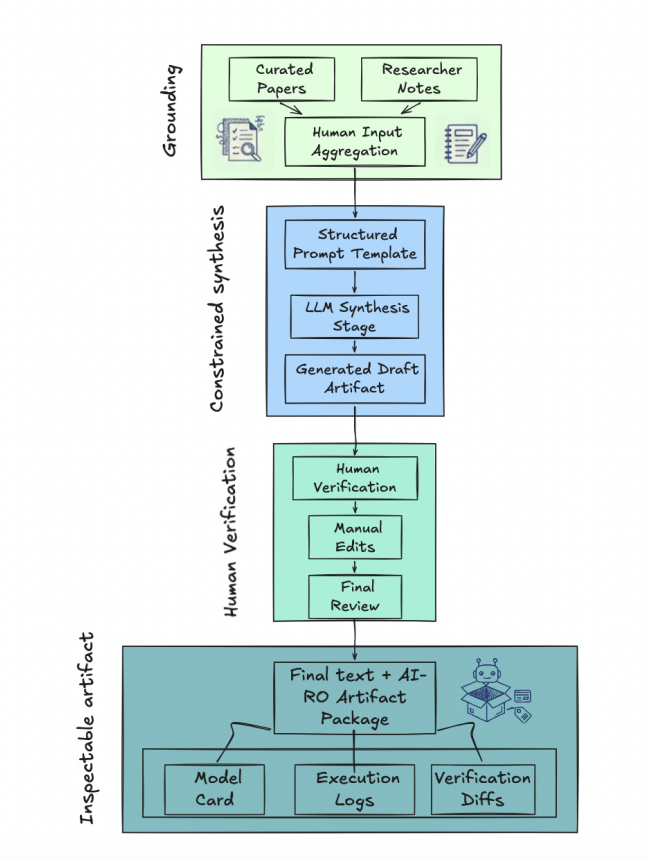


Figure 1: AI-RO Workflow for Literature Review Writing: From structured human input to verifiable outputs.

because they cannot assume responsibility for scholarly claims [7].

Limitations of Existing Approaches. Current theoretical and governance efforts typically rely on questions of authorship and contribution, disclosure statements, post-hoc AI detection efforts, or ad hoc prohibitions. Narrative disclosures are difficult to verify and often lack procedural detail. Detection methods are inherently error-prone and risk creating adversarial dynamics that discourage transparency. Blanket prohibitions ignore legitimate and beneficial uses such as language refinement or structured ideation, pushing AI use underground rather than governing it. Collectively, these approaches fail to provide structured, inspectable mechanisms for accountability and instead incentivize opacity over responsible transparency.

AI as a Research Object. We therefore propose a shift from authorship-centered debates toward infrastructural accountability. Drawing on the concept of Research Objects (ROs), structured digital artefacts that encapsulate datasets, software, workflows, and contextual metadata to support transparency and reproducibility [4], we argue that generative AI systems used in scientific writing should be governed as AI Research Objects (AI-ROs). Under this framing, AI is neither an author nor a collaborator, but a documented component of a research workflow whose provenance and configuration can be inspected alongside other scientific artifacts.

In this work, we propose reframing the use of AI in scientific research through the lens of the research object framework, emphasizing key distinctions from traditional research objects such as datasets or code. We further present an illustrative end-to-end transparency pipeline centered on one of the most prevalent and contested applications of AI: literature review writing. We emphasize that this example is intended to demonstrate the feasibility of the approach and to stimulate discussion on concrete implementations and best practices, rather than to serve as a definitive solution. Finally, we outline potential next steps toward AI governance in scientific writing and peer review, and discuss current limitations and areas requiring further development.

Scope. This work presents a conceptual and infrastructural framework for governing AI-assisted scientific workflows rather than a finalized technical standard or empirical evaluation. Our demonstration focuses on writing a literature review as a concrete case study to illustrate feasibility and design principles. However, this paper is a position and infrastructure proposal intended to stimulate discussion rather than report an evaluated system.

★ **Contributions.** 1) We introduce AI-RO, a framework for governing generative AI as inspectable research infrastructure; 2) We discuss operationalization of AI-RO; 3) We provide a feasibility demo; 4) We discuss the extension to governance interfaces; 5) Finally, we discuss limitations and provide actionable items for future development.

2. Conceptual Foundations: Research Objects

The notion of ROs was introduced as an abstraction for communicating, sharing, and reusing research results beyond traditional research papers [4]. Rather than treating scientific output as a static document, the RO paradigm frames research as a structured digital artifact that encapsulates complementary components, including hypotheses, experimental workflows, datasets, software tools, and derived results. By packaging these elements together, Research Objects make the procedural context of scientific claims explicit, supporting transparency, reuse, and reproducibility in digitally mediated research.

Building on this foundation, a growing ecosystem of standards and infrastructures has emerged to operationalize ROs in practice. RO-Crate, for example, provides a lightweight mechanism for describing research artifacts using schema.org annotations expressed in JSON-LD, lowering the barrier to adopting structured metadata in everyday workflows [27]. In domains with heightened privacy or regulatory constraints, Trusted Research Environments (TREs) establish secure contexts in which data, tools, and workflows can be accessed and audited under formal governance structures [28]. At a broader level, the FAIR principles articulate

widely accepted expectations for scientific artifacts, namely that they should be Findable, Accessible, Interoperable, and Reusable [35]. Together, these developments reflect a shift toward treating computational artifacts as first-class objects of scientific accountability.

The next section builds on this conceptual foundation to formalize generative AI systems as AI-ROs, outlining how their configuration, interaction traces, and outputs can be governed using established infrastructural principles for scientific transparency and reproducibility.

> **Takeaway 1.** Research Objects reframe scientific outputs as structured, inspectable artifacts that preserve procedural context, enabling transparency, reproducibility, and accountability in computational research.

3. AI as Research Object

Generative AI systems increasingly resemble established categories of scientific artefacts that are already governed as Research Objects. Like datasets, they embed prior information that conditions future results. Like software libraries, they implement reusable computational procedures. Like workflows, they transform inputs into outputs through configurable processes. In all these respects, AI systems function as components of research infrastructure rather than as independent epistemic agents. Their scientific relevance and/or risks arise from how they are integrated into scientific workflows.

In addition, similarly to traditional ROs, AI systems have identifiable versions, parameter settings (e.g., temperature), execution contexts, and interaction histories. This allows AI influence on outcomes to be documented and inspected. Therefore, treating AI usage as a Research Object allows researchers to leverage familiar tools and standards to support transparency, reproducibility, and auditability.

At the same time, AI Research Objects introduce properties that distinguish them from traditional artefacts and raise new methodological considerations. First, generative AI systems are inherently stochastic, meaning that identical inputs may yield different outputs even under controlled conditions. Second, many models are proprietary, limiting access to internal parameters, training data, or execution details that would traditionally support reproducibility. Third, AI-generated content often resembles human cognitive production, creating social and epistemic expectations that do not arise for conventional software or datasets. These characteristics complicate inspection and accountability, but they do not invalidate the Research Object framing. Instead, they motivate the use of additional documentation mechanisms, such as interaction logs and structured user prompts.

Table 1 summarizes the relationship between traditional Research Objects and AI Research Objects across key governance dimensions. Both ROs and AI-ROs are digital artefacts embedded in scientific workflows, subject to versioning, documentation, and human accountability. However, differences lie in the granularity of configuration, the limits of reproducibility, and the need to capture interaction-level provenance. For AI systems, parameters such as sampling settings and prompt histories become part of the methodological record, while evaluation shifts toward assessing output quality under human oversight.

Dimension	Traditional Research Objects	AI Research Objects
Ontological status	Digital scientific artefact	Digital scientific artefact
Typical examples	Datasets, software, workflows	AI use logs, AI agents
Versioning	Dataset/software versions	Model versions, checkpoints
Configuration	Parameters, scripts	Temperature, Top-p, Max tokens
Provenance	Data sources, pipelines	Prompt logs
Reproducibility	Full: Re-run workflow	Limited: Re-run model with same context
Documentation	README, metadata	Model cards, interaction logs
Evaluation	Performance metrics	Output quality, human oversight
Accountability	Human researchers	Human researchers
Automation	Low	Medium to High (in agentic settings)

TABLE 1: Systematic comparison between traditional Research Objects and AI Research Objects

> **Takeaway 2.** AI-ROs share the structure and workflow integration of traditional ROs while introducing stochastic and interaction-driven behavior.

4. Operationalizing AI Research Objects

Treating AI systems as Research Objects requires practical mechanisms for documentation, storage, inspection, and reuse. An operational AI Research Object should, at a minimum, capture the model identifier and version, system configuration and parameters, prompt and interaction logs, generated outputs, and human editing traces that distinguish AI-produced content from final authorial decisions. These artifacts form a provenance record that documents how AI systems influenced the research process. In alignment with FAIR principles, such packages can be indexed, stored, and shared using persistent identifiers and interoperable metadata schemas, enabling downstream inspection and reuse without redefining existing scientific infrastructure [35].

AI Agent as RO. This operational perspective becomes even more important when considering emerging agentic AI systems. Recent work demonstrates how AI agents can act as interactive interfaces to scientific content, encapsulating explanation, synthesis, or workflow execution within executable artifacts [22]. More advanced "AI scientist" systems aim to automate hypothesis generation, experimental design, and analysis pipelines. While such systems promise efficiency, they also risk obscuring the procedural basis of scientific claims if their internal

AI Research Object — Model Card

Run ID: ro-background
Research Topic: Paradigms of authorship in generative AI creation

Model Configuration

- Interface: OpenAI-compatible API
- Model class: LLaMA 3.1 (8B parameter family)
- Temperature: 0.2
- Top-p: 1.0
- Max tokens: 1200
- Input bundle SHA-256: `79b576eacbeee64171 dfe0bc8cd7a6e5e09da7f25259a97819cb ac5ce35d0860`

Artifacts Released

- Privatized interaction logs (taxonomy + synthesis)
- Generated taxonomy structure (JSON)
- Draft synthesis text (Markdown)
- Audit table (CSV)

Intended Use
Assistive structural synthesis and drafting using human-authored inputs. The system was not treated as an autonomous factual authority.

Human Oversight
All interpretations, claims, and references were verified by the authors.

Disclosure
A generative language model was used for structural synthesis and drafting under constrained prompts. Prompts and intermediate artifacts are released to support transparency.

Limitations
Outputs may contain model biases or structural artifacts. Human validation was required.

Reproducibility Note
Exact regeneration may vary due to backend nondeterminism. Parameters and artifact hashes support approximate replication.

Figure 2: Illustrative AI Research Object model card summarizing configuration, artifacts, oversight, and reproducibility information for the generative assistance workflow.

decision processes remain opaque. Framing agents as AI Research Objects provides a governance mechanism. Their architectures, configurations, and interaction traces become part of the documented workflow, preserving inspectability even when autonomy increases. As automation deepens, the importance of structured provenance will only grow: both for preserving the control and integrity of scientific research, and for using these artifacts to advance further development of scientific AI agents.

Takeaway 3. Treating AI and AI agents as Research Objects turns interaction histories into auditable artifacts, preserving transparency and accountability as scientific workflows become increasingly automated.

5. Example: Legitimate AI Contribution vs. Misuse in Scientific Writing of Related Work

To make the AI Research Object (AI-RO) framing concrete, we present a minimal paired example contrasting a legitimate use of generative AI in scientific writing with an opaque misuse scenario. Both cases involve preparing the related work section of a research paper. We focus on literature review writing because it is among the most common applications of AI in scientific authorship and is frequently associated with risks such as hallucinated citations and bias [2], [33]. The distinction lies not in whether AI is used, but in how its role is bounded, documented, and integrated into the research workflow.

In the legitimate case, the author uses an LLM as an assistive synthesis tool while retaining full intellectual responsibility. The process begins with a curated set of verified references and human-authored summaries derived from direct reading. The model is prompted to organize these materials and suggest comparative phrasing, which is treated as provisional text rather than final output. The author verifies all claims and citations against primary sources and revises the draft to reflect the paper's contribution.

Crucially, this interaction is preserved as an AI-RO artifact. Model identity, configuration context, and interaction traces are recorded, and human edits distinguish AI suggestions from final authorial decisions. The accompanying disclosure clarifies that the model supported structural synthesis and language refinement, while all scholarly judgments remained with the authors. Although stochastic generation limits exact regeneration, the artifact preserves a meaningful provenance trail that allows inspection of how AI influenced the workflow. Under this framing, the contribution is legitimate because the AI system operates as a transparent, bounded component of the research process.

The misuse case illustrates the opposite pattern: substitution of undocumented AI output for scholarly synthesis. Here, the author prompts the model to produce a complete related work section without supplying verified inputs or enforcing validation constraints. The resulting text may contain fabricated or misattributed citations, yet is incorporated with minimal review. Only a generic disclosure is provided, and no interaction records are retained. From an AI-RO perspective, the absence of provenance eliminates inspectability, making it impossible to reconstruct how claims were produced.

This paired contrast highlights a central claim of this paper: legitimacy in AI-assisted scientific writing is procedural rather than binary. Ethical use depends not on the presence of AI, but on whether its role is documented, bounded, and subject to the same expectations of transparency and accountability that govern other research artifacts.

Takeaway 4. The paired example shows that legitimacy in AI-assisted scientific writing depends on procedural transparency and documented human over-

sight, not on the mere presence or absence of AI tools.

6. Demonstration: Inspectable AI Workflows

To illustrate the practicality of the AI Research Object (AI-RO) framework, we present a workflow demonstration of an inspectable AI-assisted literature review pipeline (Figure 1). The purpose of this demonstration is to illustrate how documentation practices shape the inspectability and governance of AI-assisted scholarship, in contrast to unstructured, opaque use.

Workflow Design. The example workflow implements an AI-RO in which generative AI functions as a bounded synthesis tool embedded within a documented research pipeline. Human-authored inputs, including verified references, structured reading notes, and a contribution statement, are supplied to the model under explicit constraints. Outputs are treated as provisional drafts subject to human verification, and all model interactions are recorded as structured provenance artifacts. A detailed procedural specification of the workflow is provided in the Appendix section A.

Artifact Demonstration. Example workflow produces a concrete AI-RO artifact package that exposes the internal structure of AI-assisted synthesis. The central artifact is an AI Research Object model card (Figure 2), which summarizes model identity, configuration parameters, released artifacts, intended use, human oversight, and reproducibility considerations. While traditional model cards describe a model's capabilities, limitations, and recommended uses [23], this model card functions as a compact inspection interface, allowing readers to understand how the generative system was configured, what materials are available for audit, and how responsibility is maintained.

Supporting artifacts further expose workflow mechanics. Figure 3 illustrates the constrained prompt template used to bind model behavior to verified human-authored inputs. The artifacts and their role in auditing frameworks are summarized in Table 2. An anonymized companion research artifact is available at OSF.io via a view-only link: `Anonymous artifact link`. The artifact contains an executable workflow, structured inputs, generated outputs, and provenance documentation packaged as an RO-Crate. To preserve double-blind review and avoid disclosure of manuscript text, interaction transcripts and timestamps are redacted and replaced with integrity hashes, while retaining configuration parameters and workflow structure. A detailed description of the artifact packaging, RO-Crate structure, and anonymization procedure (including log redaction and provenance preservation) is provided in Appendix A.

Discussion. While we present illustrative examples of AI-related research artifacts, we do not prescribe these templates as fixed standards. The structure and content of such artifacts should be adapted to the requirements of specific venues and workflows. For instance, the strictness of prompt templates may vary with the intended level of model autonomy: some templates may allow external references while requiring the model to produce verification checklists for human review. Regardless of configuration, it is advisable to place hard constraints at the beginning or end of the template, as LLMs tend to exhibit reduced attention to instructions embedded in the middle of longer prompts [19]. Other directions may also include integrating retrieval-augmented generation (RAG) systems tailored for literature analysis and synthesis, enabling tighter coupling between curated sources and model outputs [3].

The artifacts discussed here focus on a single component of the research lifecycle, namely, literature review, as a concrete demonstration. Comparable AI-RO pipelines could be developed for other stages, including ideation, experimental design, data analysis, and automated experimentation. Together, these components could form a larger, integrated AI-RO pipeline that covers the entire manuscript development lifecycle. Consequently, in more advanced settings, releasing AI agents that assist across multiple stages of research as a part of a documentation package may provide a unified transparency layer spanning the entire workflow.

Takeaway 5. By structuring literature review synthesis as an AI Research Object, the workflow demonstrates how provenance artifacts enable transparency, human oversight, and governance of generative AI.

7. From AI Research Objects to Governance Interfaces

The AI-RO framework establishes how generative AI interactions can be documented as structured, inspectable components of a research workflow. A complementary question is how such artifacts might be surfaced within real scholarly processes, particularly peer review, without introducing excessive friction or compromising confidentiality.

To operationalize this idea, we propose a conceptual governance interface in which AI-RO documentation is packaged as a companion artifact submitted alongside the manuscript. Rather than exposing full conversational logs, this interface surfaces curated provenance indicators, such as model identity, configuration summaries, interaction categories, and evidence of human verification, that communicate how generative assistance was integrated into the workflow. Access to these artifacts can be tiered according to venue policy, preserving confidentiality while enabling meaningful inspection.

This interface reframes AI disclosure from a declarative statement into a structured accountability layer. Reviewers gain visibility into the procedural role of AI systems without being required to audit raw transcripts, and editors can evaluate compliance using standardized artifact summaries. Crucially, the model emphasizes incremental integration: AI-RO artifacts are designed to complement existing submission infrastructures rather than replace them, minimizing additional friction for authors and reviewers.

In this way, governance becomes an extension of documented workflow provenance rather than an external com-

Artifact	Captured Information	Governance Function	Primary User
Model card	Model identity, configuration parameters, intended use, released artifacts, limitations, and oversight description	Provides a high-level inspection interface summarizing how AI was used and governed; supports transparency and interpretability	Authors, Readers, reviewers
Prompt template	Constraints, task framing, allowed inputs, and generation rules	Documents how model behavior was bounded; prevents hidden or unconstrained synthesis	Authors, reviewers
Interaction/invocation logs	Timestamps, parameters, cryptographic hashes, prompt–response linkage	Preserves provenance and enables auditability without exposing sensitive raw content	Authors, Auditors, editors
Human-authored input bundle	Verified references, structured notes, contribution framing	Establishes human grounding of claims and prevents fabricated or unsupported synthesis	Authors, reviewers
Generated draft artifact	Model-produced synthesis text and intermediate reasoning structure	Exposes how AI influenced writing; enables comparison with final authorial decisions	Auditors, researchers
Human edit trace	Tracked revisions distinguishing AI suggestions from author decisions	Demonstrates human verification and responsibility over final claims	Reviewers, auditors
RO-Crate metadata manifest	Machine-readable description of files, relationships between artifacts, authorship roles, and provenance references	Enables interoperability and automated inspection	Repositories, automated analysis systems

TABLE 2: AI Research Object artifact stack and their governance roles. Each artifact externalizes a different aspect of AI-assisted workflow, enabling inspection, accountability, and procedural traceability.

PROMPT TEMPLATE

You are assisting with drafting a RELATED WORK section for a scientific paper on {{TITLE}}.

You will be given:

- a topic
- a contribution statement
- a taxonomy (clusters and assignments)
- structured human notes for each paper (summary, strengths, limitations, relation)

Task: Draft a related work section (target length: {{TARGET_WORDS}} words) that:

- Organizes discussion around the provided clusters
- Uses neutral, scholarly tone
- Highlights similarities and differences with respect to the contribution statement
- Mentions strengths/limitations carefully
- Avoids overclaiming (if evidence is weak, use cautious language)

Hard constraints (must follow):

- Use ONLY the provided papers and identifiers. Do NOT invent citations or claims.
- Every comparative claim must be supported by the human notes provided.
- Use citations in the form: ({{citation}}; {{pid}}).
- If a claim is not supported by the notes, mark it as [NEEDS HUMAN CHECK].

Output format: 1) "RELATED WORK (DRAFT)" header 2) Draft text in paragraphs 3) "CLAIM CHECKLIST" bullet list with supporting paper IDs

Figure 3: Prompt template used to guide the language model in drafting a structured literature review section under human-authored constraints.

pliance burden. Structured AI artifacts enable fine-grained transparency, support reproducibility, and align AI-assisted scholarship with familiar mechanisms of scientific accountability, while remaining practical for adoption within contemporary publishing pipelines.

Takeaway 6. Standardized interfaces for accessing and analyzing AI-RO artifacts provide reviewers with actionable transparency while preserving confidentiality and minimizing friction.

8. Challenges & Future Directions

While the AI-RO framework provides a structured path toward transparency and accountability in AI-assisted scholarship, its practical adoption raises intertwined social, technical, and institutional challenges. Addressing these barriers is essential if inspectable AI workflows are to move from conceptual proposal to routine scholarly practice.

Incentives for Transparency. Currently, disclosing AI assistance may negatively affect judgments of competence or credibility, creating incentives for underreporting [25]. In environments where transparency is penalized, responsible documentation becomes risky. Overcoming this stigma requires shifting evaluation norms away from whether AI was used toward how it was integrated, verified, and governed. In addition, it is necessary to incentivize transparency by rewarding the submission of AI-RO and by simplifying the AI-RO generation and disclosure pipelines.

Privacy and Security. Provenance artifacts may contain sensitive drafts, reviewer commentary, or proprietary data, while many widely used AI systems are closed-source, limiting insight into training data and internal behavior. These constraints complicate both auditability and reproducibility. Future work should explore privacy-preserving logging, selective disclosure models, and secure infrastructure to manage AI-RO artifacts without exposing confidential information.

Transparency-Safety Trade-off. The growing capabilities of AI models have raised concerns about potential misuse following their release [29]. At the same time, counterarguments emphasize that responsible disclosure can help mitigate misuse by enabling scrutiny and defensive development [26]. Lessons from established approaches to the

secure governance of research artifacts, therefore, provide a useful foundation for designing mechanisms that support controlled or limited disclosure.

Compliance Burden. Structured provenance introduces additional work for authors and new inspection responsibilities for editors and reviewers. Without automation and usable interfaces, AI-RO workflows risk becoming administratively heavy. Progress in automated metadata capture and lightweight audit tooling will be critical to ensuring that accountability mechanisms support, rather than obstruct, scholarly productivity.

Ethical and Data Governance. Broader concerns surrounding copyrighted training data, dataset provenance, and systemic bias remain unresolved across the AI ecosystem. While the AI-RO framework does not directly resolve these issues, it provides a structured context for documenting and highlighting them. Future governance models may link transparency incentives to the use of ethically sourced or auditable AI models.

Reproducibility. Reproducibility in generative systems is inherently limited. LLM outputs remain stochastic even under fixed seeds or low-temperature sampling [5], and proprietary models restrict access to internal representations. AI-RO practices, therefore, emphasize approximate reproducibility through preserved configuration, interaction context, and artifact hashes. Advancing this area will require methodological standards that recognize graded reproducibility appropriate to probabilistic systems.

> **Takeaway 7.** AI governance is not solely a technical problem, but a socio-technical transition that requires cultural norms, infrastructure design, and methodological innovation.

9. Conclusion, Limitations & Call for Action

Generative AI is rapidly becoming embedded in the everyday practice of scientific writing. Attempts to govern this shift through authorship debates, detection tools, or blanket prohibitions struggle to capture the infrastructural role that AI now plays in research workflows. This paper argues for a reframing: treating AI systems as *Research Objects* whose configuration, interaction history, and outputs can be documented, inspected, and evaluated alongside data and software artifacts.

By operationalizing AI usage as structured provenance, the AI-RO framework shifts governance from attribution disputes toward methodological transparency. Beyond a governance proposal, AI-RO serves as a metascientific intervention in how epistemic provenance is recorded and evaluated. The paired workflow example and documentation demonstration illustrate that legitimacy in AI-assisted scholarship depends less on whether AI is used than on whether its influence is inspectable. The companion model interface example further shows how such documentation can be integrated into peer-review infrastructure without compromising confidentiality.

Limitations. While we demonstrate the advantages and feasibility of the AI-RO approach through an example workflow, we acknowledge remaining limitations. The framework and demonstration intentionally focus on a single stage of the research lifecycle (i.e., drafting a RELATED WORK section) and do not constitute a comprehensive governance solution. While the proposed principles are designed to generalize beyond the literature review example, extending them to other stages of scientific workflows will require domain-specific adaptation, tooling, and validation. In addition, rigorous validation of the framework requires human study to evaluate the usefulness of the artifacts for scientific review and to assess the trustworthiness of the scientific work. Accordingly, the framework should be understood as an initial architectural proposal that delineates a design direction rather than a fully operational standard. Finally, we point toward a future research agenda. Below, we list the most urgent action points to support the use of transparent and ethical AI in scientific research.

> **Call for Action:**
>
> ✓ *Standardization of AI usage documentation*
>
> ✓ *Developing privacy-aware provenance infrastructure*
>
> ✓ *Supporting cultural norms that reward transparency usable audit interfaces*
>
> ✓ *Encouraging the development and adoption of open and responsibly trained models*
>
> ✓ *Developing protocols and tools guiding AI use in research*

References

[1] S. Altmäe, A. Sola-Leyva, and A. Salumets. Artificial intelligence in scientific writing: a friend or a foe? *Reproductive BioMedicine Online*, 47(1):3–9, 2023.

[2] S. Ansari. Compound deception in elite peer review: A failure mode taxonomy of 100 fabricated citations at neurips 2025, 2026.

[3] A. Asai, J. He, R. Shao, W. Shi, A. Singh, J. C. Chang, K. Lo, L. Soldaini, S. Feldman, M. D'Arcy, et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, pages 1–7, 2026.

[4] K. Belhajjame, C. Goble, and D. De Roure. Research object management: opportunities and challenges. In *Conference on Computer Supported Cooperative Work. CSCW*, 2012.

[5] R. E. Blackwell, J. Barry, and A. G. Cohn. Towards reproducible llm evaluation: Quantifying uncertainty in llm benchmark scores. *arXiv preprint arXiv:2410.03492*, 2024.

[6] E. Bonadio and H. Felisberto. Copyrightability of ai outputs: The us copyright office's perspective. *Available at SSRN 5252465*, 2025.

[7] Committee on Publication Ethics (COPE). Authorship and ai tools: Cope position, 2023. Accessed: 2026-02-01.

[8] G. Core. Louisiana, the potential safe-haven for creators of artificial intelligence: Louisiana's response to thaler v. perlmutter. *La. L. Rev.*, 85:651, 2024.

[9] J. Dempere, L. K. Ramasamy, J. Harris, et al. Ai as a research partner: Advocating for co-authorship in academic publications. *The Artificial Intelligence Business Review*, 1(2), 2025.

[10] N. Editorials. Tools such as chatgpt threaten transparent science; here are our ground rules for their use. *Nature*, 613(7945):612, 2023.

[11] Elsevier. Generative ai policies for journals, 2026. Accessed: 2026-02-10. Elsevier's publisher guidance on AI use, including the requirement to disclose AI assistance in manuscript preparation and restrictions on listing AI as an author.

[12] J. Fritz. Understanding authorship in artificial intelligence-assisted works. *Journal of Intellectual Property Law and Practice*, page jpae119, 2025.

[13] C. Ganjavi, M. B. Eppler, A. Pekcan, B. Biedermann, A. Abreu, G. S. Collins, I. S. Gill, and G. E. Cacciamani. Bibliometric analysis of publisher and journal instructions to authors on generative-ai in academic and scientific publishing. *arXiv preprint arXiv:2307.11918*, 2023.

[14] J. Gottweis, W. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, and V. Natarajan. Towards an ai co-scientist: A multi-agent system for scientific discovery. *arXiv preprint arXiv:2502.18864*, page 3, 2025.

[15] T. Hesman. The machine that invents. *St. Louis Post-Dispatch*, 2004.

[16] Y. Hwang, D. Shin, and J. H. Lee. Who owns ai-generated artwork? revisiting the work of generative ai based on human-ai co-creation. *Telematics and Informatics*, 98:102266, 2025.

[17] IEEE Publication Services and Products Board. *IEEE PSPB Operations Manual*. The Institute of Electrical and Electronics Engineers, 2018. Accessed on 2026-02-01.

[18] W. Liang, Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, et al. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024.

[19] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

[20] W. Liu and W. Huang. Authorship in human-ai collaborative creation: A creative control theory perspective. *Computer Law & Security Review*, 57:106139, 2025.

[21] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[22] J. Miao, J. R. Davis, Y. Zhang, J. K. Pritchard, and J. Zou. Paper2agent: Reimagining research papers as interactive and reliable ai agents. *arXiv preprint arXiv:2509.06917*, 2025.

[23] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[24] A. Sarkar. Enough with "human-ai collaboration". In *Extended abstracts of the 2023 CHI conference on human factors in computing systems*, pages 1–8, 2023.

[25] O. Schilke and M. Reimann. The transparency dilemma: How ai disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188:104405, 2025.

[26] T. Shevlane and A. Dafoe. The offense-defense balance of scientific knowledge: does publishing ai research reduce misuse? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 173–179, 2020.

[27] S. Soiland-Reyes, P. Sefton, L. J. Castro, F. Coppens, D. Garijo, S. Leo, M. Portier, and P. Groth. Creating lightweight fair digital objects with ro-crate. *Research Ideas and Outcomes*, 8:e93937, 2022.

[28] S. Soiland-Reyes, S. Wheater, T. Giles, J. Couldridge, P. Quinlan, and C. Goble. The five safes ro-crate: Fair digital objects for trusted research environments for health data research. In *Open Conference Proceedings*, volume 5, 2024.

[29] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

[30] Taylor & Francis. Ai policy, 2026. Accessed: 2026-02-10. Taylor & Francis outlines expectations for responsible use of generative AI tools, including disclosure of tool name, purpose, and version, and that AI cannot be listed as an author.

[31] H. H. Thorp. Chatgpt is fun, but not an author, 2023.

[32] R. Van Noorden and J. M. Perkel. Ai and science: what 1,600 researchers think. *Nature*, 621(7980):672–675, 2023.

[33] S. S. M. Vasu, I. Sheth, H.-P. Wang, R. Binkyte, and M. Fritz. Justice in judgment: Unveiling (hidden) bias in llm-assisted peer reviews. *arXiv preprint arXiv:2509.13400*, 2025.

[34] R. Watkins. Guidance for researchers and peer-reviewers on the ethical use of large language models (llms) in scientific research workflows. *AI and Ethics*, 4(4):969–974, 2024.

[35] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

Appendix

This appendix describes how the AI-RO corresponds to the literature review workflow example introduced in the paper and how it was prepared for anonymous inspection. The goal of the released artifact is to provide an inspectable execution trace of the proposed literature review workflow rather than to evaluate writing quality. The artifact, therefore, contains a representative run illustrating inputs, model invocation, verification steps, and provenance capture, and can be examined independently of the manuscript topic.

1. Workflow Overview

AI-assisted writing is operationalized as a bounded synthesis process embedded in a human research pipeline. The workflow proceeds in five stages:

1) preparation of structured input notes,
2) constrained model invocation,
3) intermediate artifact generation,
4) human verification,
5) provenance recording.

The language model functions as a synthesis and organization tool.

2. Structured Inputs

The model is never prompted with open-ended literature queries. Instead it receives a curated input bundle consisting of human-authored reading notes. Each record follows a fixed schema:

> identifier, citation, persistent identifier, summary, strengths, limitations, and relation to the contribution

These notes are derived from direct reading. The prompt constrains the model to reorganize and summarize only this material. The model is explicitly instructed not to invent sources or claims.

3. Model Invocation

The model is invoked in two stages. First, the model groups papers into thematic clusters and provides a brief rationale for each cluster.

Second, the model produces a structured narrative outline derived only from the provided notes.

Hard constraints embedded in the prompt require that: citations originate from the input bundle, comparative statements correspond to human notes, and uncertainty is explicitly flagged.

4. Generated Artifacts

Each run produces inspection artifacts rather than only a final text. The artifact includes:

- a taxonomy file describing conceptual groupings,
- a demonstration draft showing synthesis structure,
- an audit checklist linking claims to sources,
- a verification table for human review,
- a provenance record of model interaction.

These outputs externalize intermediate reasoning and make the writing process reviewable. The artifacts are available at OSF.io via a view-only link: `Anonymous artifact link`.

5. Provenance Logging

Every model invocation generates a structured log containing:

- model identity and configuration,
- generation parameters,
- cryptographic hashes of prompt, response, and input bundle.

Hashes allow verification that an artifact corresponds to a specific input without exposing the underlying text. The released artifact therefore enables auditability while preventing disclosure of manuscript material.

6. Log Redaction and Anonymization

For double-blind review, interaction transcripts are not publicly distributed. Instead:

- full prompts and responses are replaced by integrity hashes,
- timestamps are removed,
- local file paths and machine identifiers are removed,
- backend service endpoints are generalized.

This preserves verifiability while preventing attribution or reconstruction of the manuscript.

7. RO-Crate Structure

The artifact is packaged using the RO-Crate specification, which describes research objects using machine-readable metadata. The released archive contains:

- source code implementing the workflow,
- prompt templates,
- structured input notes,
- redacted provenance logs,
- generated demonstration outputs,
- RO-Crate metadata describing file roles and relationships.

The file `ro-crate-metadata.json` declares each component as code, data, or provenance material and links outputs to the process that produced them. The package can therefore be inspected both manually and programmatically.

8. Verification Procedure

The audit workflow is designed so a reviewer can:

1) inspect structured notes,
2) examine generated clusters and draft structure,
3) check claim-to-source mappings,
4) verify artifact hashes,
5) confirm that outputs derive from the recorded inputs.

Human verification is required before any generated text can be treated as scholarly writing.