The Usefulness of Artificial Intelligence Across the Research Cycle: A VISTA Framework for Responsible Scholarly Work

Daves Baltazar

Abstract

Artificial intelligence (AI) can accelerate research, but speed alone does not establish scholarly usefulness. A tool is useful only when its contribution to quality, efficiency, reproducibility, or accessibility exceeds the costs of verification and the risks introduced into the research record. This conceptual article examines AI use across problem formulation, literature discovery, research design, data preparation, analysis, interpretation, and scholarly communication. It distinguishes specialized analytical systems from general-purpose generative AI and argues that their value is task-dependent. AI is well suited to expanding search vocabulary, organizing candidate literature, drafting code, detecting patterns, proposing alternative explanations, and improving language. It is unreliable when treated as an authority for citations, methods, numerical results, or final interpretations. To convert this distinction into an operational process, the article develops the VISTA framework: Validate the task and permissible use; Investigate evidence through authoritative sources; Structure the method and analysis under human control; Test every consequential output; and Account for material AI use and human responsibility. Two tools accompany the framework: a research-cycle risk map and a minimum verification record. The analysis integrates literature on AI-assisted scientific discovery, language-model hallucination, research transparency, publication ethics, and fairness. It concludes that AI's most defensible role is augmentation rather than epistemic delegation. Researchers may delegate bounded operations, but they cannot delegate responsibility for the question, evidence, method, interpretation, or published claim. The framework offers researchers, reviewers, and institutions a proportionate basis for evaluating AI-supported work without assuming that all uses are either harmless or unacceptable.

Keywords: artificial intelligence; generative AI; research integrity; scholarly communication; reproducibility; human oversight

Introduction

Artificial intelligence is becoming part of the ordinary research environment. Machine-learning systems already support image analysis, molecular prediction, pattern recognition, and large-scale classification. Generative systems add a conversational interface that can summarize text, propose code, reorganize arguments, translate prose, and produce plausible answers within seconds. These capabilities create a practical attraction: tasks that once required repeated searching, drafting, or debugging can be completed more quickly. Wang et al. (2023) show that AI has contributed across several stages of scientific discovery, including hypothesis generation, experimental design, data interpretation, and the identification of scientific relationships. The relevant question is therefore no longer whether AI can be used in

research, but under what conditions its use improves the reliability and value of scholarly work.

That question cannot be answered by productivity claims alone. A fluent output may contain a false citation, an invented numerical detail, an inappropriate statistical procedure, or an interpretation unsupported by the data. Large language models generate likely linguistic sequences; they do not possess a researcher's accountable relationship to evidence. Ji et al. (2023) describe hallucination as the production of content that is unfaithful to source material or factual knowledge. In a controlled comparison, Gao et al. (2023) found that generated scientific abstracts could appear convincing while sometimes containing fabricated numbers. Such errors are especially dangerous in research because polished language can conceal the absence of a traceable evidentiary chain.

The debate is often reduced to two positions. One treats AI as an efficiency tool whose outputs can be corrected like ordinary drafts. The other treats any AI involvement as incompatible with authentic scholarship. Both positions are too broad. AI use is not a single intervention: using a model to suggest database search terms differs materially from asking it to supply references, choose an analysis, or interpret a study. The consequences also vary by discipline, data sensitivity, researcher expertise, and the reversibility of the task. UNESCO (2023) consequently emphasizes a human-centered approach, protection of data privacy, validation, and institutional capacity rather than unrestricted adoption.

This article develops a task-based account of usefulness. It proposes that AI is useful when it produces a net improvement in research quality, efficiency, reproducibility, or access after verification costs and foreseeable harms are considered. It then introduces the VISTA framework as a practical sequence for governing AI-supported research. The article is conceptual; it reports no experiment, survey, or institutional evaluation and makes no causal claim that VISTA has already improved research outcomes.

Purpose of the Article

The article has three objectives: (a) to identify where AI can contribute across the research cycle; (b) to distinguish augmentation from the delegation of scholarly judgment; and (c) to present a verification and accountability framework that can be applied by individual researchers, research teams, reviewers, and institutions. The guiding question is: How can researchers obtain the practical benefits of AI while preserving a transparent chain from research question to evidence, method, analysis, and claim?

Conceptual Approach

Scope and Source Base

The article uses integrative conceptual synthesis. It connects research on AI-assisted scientific discovery with work on language-model limitations, transparency, open research, scholarly publishing, ethics, and detection bias. The sources were selected to illuminate functions and safeguards rather than to estimate a pooled effect. The term AI includes specialized predictive or classificatory systems and general-purpose generative systems, but the discussion emphasizes generative AI because its

fluent outputs can enter literature review, coding, interpretation, and writing with unusually little friction.

The unit of analysis is an AI-assisted research task. This level is more precise than judging an entire application as acceptable or unacceptable. A single model may be relatively useful for brainstorming synonyms, moderately risky for screening article titles, and unacceptable for processing identifiable participant data on an unapproved service. Task-level analysis also makes safeguards proportionate: the greater the consequence and the weaker the independent evidence, the stronger the required validation.

The Principle of Conditional Usefulness

AI usefulness can be understood as a balance among four gains and three costs. The gains are quality, efficiency, reproducibility, and accessibility. Quality may improve when a system identifies anomalies, exposes overlooked alternatives, or enables analysis at a scale beyond unaided human review. Efficiency may improve when routine transformations, code scaffolding, or document organization are accelerated. Reproducibility may improve when AI-assisted code and workflow records are preserved and independently rerun. Accessibility may improve through translation, language support, alternative explanations, or interfaces that lower technical barriers.

The costs are verification burden, introduced risk, and opacity. Verification burden includes the time and expertise required to check an output. Introduced risk includes factual error, bias, privacy loss, intellectual-property exposure, and automation of an unsuitable method. Opacity concerns the inability to explain, reproduce, or audit how an output influenced a scholarly claim. An AI-generated draft that saves ten minutes but requires an hour of source checking has little efficiency value. A classification system that saves weeks and achieves documented performance on a representative validation sample may have substantial value. Usefulness is therefore conditional, not inherent in the presence of AI.

AI Usefulness Across the Research Cycle

Problem Framing and Literature Discovery

At the beginning of a project, generative AI can broaden a researcher's vocabulary. It can propose synonyms, adjacent constructs, exclusion terms, disciplinary labels, and candidate combinations for a database search. It can also help convert a broad interest into several answerable question forms. These are generative tasks: the output expands possibilities but does not determine which question is important or which literature is authoritative. Researchers remain responsible for novelty, feasibility, theoretical significance, and the relationship between the question and an identifiable body of evidence.

AI can also assist with literature triage by clustering titles, extracting candidate themes, or applying human-defined screening criteria. The gain can be considerable when the search set is large. The safeguard is retrieval from the original source. A model's citation, summary, or quotation is not evidence until the researcher has located the work, confirmed its bibliographic identity, read the relevant passage in context, and recorded the source. This rule directly addresses the tendency of language models to produce plausible but nonexistent references. AI may suggest where to look; it should not become the literature database.

Research Design and Protocol Development

During design, AI can act as a structured critic. A researcher may ask for alternative operational definitions, plausible confounders, threats to validity, missing questionnaire domains, counterexamples, or edge cases in a protocol. The output can improve design deliberation by making more candidate problems visible. AI can likewise draft a data dictionary, simulation plan, interview guide, consent-language outline, or analysis checklist from parameters supplied by the research team.

These contributions are useful only as proposals. A model cannot determine whether an instrument is valid for a population, whether a sample is adequate, whether a causal claim is identified, or whether participant-facing language satisfies local ethical requirements. Those decisions depend on disciplinary knowledge, context, and accountable review. Protocol elements should therefore be checked against primary methodological sources, pilot-tested where appropriate, and approved before data collection. When an AI suggestion materially changes recruitment, measurement, intervention, or analysis, that change belongs in the research record.

Data Preparation, Coding, and Analysis

AI can reduce labor in transcription, de-identification support, file classification, missing-value diagnostics, qualitative code suggestions, image segmentation, and code generation. Specialized systems may detect relationships in high-dimensional data that are difficult to identify manually. Conversational tools can explain error messages or translate an analytical intention into a provisional script. These uses can expand participation in computational work, particularly for researchers who are still developing programming expertise.

The central safeguard is separation between raw evidence and generated transformation. Raw data should be preserved unchanged, with access controls and version history. Generated code must be read, tested on known cases, and rerun in a controlled environment. Automated classifications require a validation set relevant to the target population, performance measures suited to the research question, and examination of consequential subgroup errors. Qualitative suggestions must be compared with source excerpts and the interpretive framework. Confidential, identifiable, proprietary, or peer-review material should not be entered into an external system unless the service and institutional arrangements expressly protect that use.

Interpretation and Scholarly Communication

AI can help researchers formulate alternative explanations, conduct a structured challenge to an interpretation, identify claims that require citations, improve organization, and edit language for clarity. Used in this way, the model functions as a fallible reader rather than an author or adjudicator. It may expose an ambiguity that the researcher then resolves by returning to the data, code, or literature. It may also improve access for multilingual scholars, although language assistance should not erase culturally specific meaning or standardize every argument into a dominant academic voice.

AI should not supply final numerical statements, quotations, citations, or conclusions without direct verification. Every number in a manuscript should be traceable to an analysis output or authoritative source; every quotation to the original text; every citation to a work the author has examined; and every conclusion to the

study's actual evidence and design. Material AI assistance should be described according to the policy of the relevant institution or journal. A tool cannot accept responsibility, respond to peer review as an accountable scholar, or approve the final record, and therefore should not be listed as an author (Nature, 2023).

Table 1

Task-based map of AI usefulness and research risk

Research stage	Potentially useful AI tasks	Principal risk	Required safeguard
Framing and search	Query expansion; concept mapping; candidate questions	Invented sources; narrowed framing; hidden omissions	Search scholarly databases; retrieve and read original sources
Design	Protocol critique; draft instruments; simulations; checklists	Invalid method; inappropriate construct; ethical mismatch	Method expert review; primary guidance; pilot or preregistration
Data preparation	Transcription; classification; cleaning suggestions; coding support	Privacy loss; altered evidence; unequal error	Approved data environment; frozen raw data; validation sample
Analysis	Code scaffolding; diagnostics; pattern detection; visualization drafts	Incorrect code; model mismatch; irreproducible result	Version control; test cases; independent rerun; sensitivity checks
Interpretation and writing	Counterarguments; organization; language editing; translation	Unsupported claims; fabricated details; concealed dependence	Return to data and sources; author approval; material-use disclosure

Note. Risk depends on context, data sensitivity, consequence, and the availability of independent evidence. The same tool can occupy different risk levels for different tasks.

The VISTA Framework

VISTA converts the principle of conditional usefulness into five linked decisions. It is designed as a minimum process rather than a new compliance layer. Each stage produces a record that makes the next stage possible. Failure at a stop condition means that the researcher should change the task, use an approved alternative, obtain additional review, or proceed without AI.

Validate the Task and Permissible Use

Validation begins before a prompt is written. The researcher specifies the exact task, expected benefit, possible consequence, data classification, and applicable institutional, funder, publisher, or participant obligations. The task should be narrow enough to review. 'Help with my research' is not a valid specification; 'propose synonyms for these three database concepts' is. Validation also asks whether the researcher has the expertise and source access needed to detect a poor output. A task should not be delegated merely because it is difficult to evaluate.

Investigate Evidence Through Authoritative Sources

The second step separates discovery from evidence. AI-generated leads are converted into verifiable objects: database records, primary articles, official datasets, documented software, or methodological standards. Claims are checked in their original context, not against another generated summary. The researcher records failed searches and excluded sources when these affect comprehensiveness. This stage is especially important in literature reviews, where a polished synthesis can hide missing studies, incorrect attribution, or dependence on sources that cannot be retrieved.

Structure the Method and Analysis Under Human Control

The third step places the AI operation inside a predefined scholarly workflow. Inputs, transformations, outputs, decision rights, and validation criteria are specified. For code, this means an environment, version history, test data, and expected results. For screening, it means explicit eligibility criteria and a human adjudication route. For qualitative work, it means a stated interpretive approach and access to the underlying excerpts. Structure prevents an attractive output from silently redefining the method after results are visible.

Test Every Consequential Output

Testing is proportional to consequence. Citations are matched to retrieved sources; quotations are compared word for word; code is tested on cases with known answers; statistical outputs are independently regenerated; classifications are evaluated against labeled data; summaries are compared with source passages; and interpretations are challenged through alternatives and sensitivity analysis. The researcher should be able to explain why the output is acceptable without appealing to the model's fluency or popularity. If a result cannot be independently checked, it should not support a consequential claim.

Account for Use and Human Responsibility

The final step records material use in enough detail for appropriate review. A minimum record includes the tool and available version, date, task, input restrictions, relevant prompts or workflow, output retained, checks performed, errors found, changes made, and responsible researcher. Disclosure in the published article should follow the venue's policy and should be meaningful rather than ceremonial. Regardless of disclosure format, the named authors retain responsibility for the integrity, originality, confidentiality, citations, analysis, and conclusions of the work.

Table 2

Operational decisions in the VISTA framework

Step	Decision question	Minimum record	Stop condition
Validate	Is the task permitted, bounded, beneficial, and reviewable?	Task, expected benefit, data class, applicable policy	Sensitive or prohibited use; no competent reviewer
Investigate	Can every evidentiary claim be traced to an authoritative source?	Retrieved sources, database path, inclusion decisions	Source cannot be found or checked
Structure	Is AI contained within a method defined by the researcher?	Inputs, transformations, criteria, versions, decision rights	Output silently determines the method
Test	Has each consequential output survived an independent check?	Test cases, comparisons, reruns, error and revision log	Result cannot be reproduced or explained
Account	Can reviewers understand material AI influence and human responsibility?	Tool, date, task, workflow, checks, accountable author	Use cannot be recorded or responsibly approved

Discussion

Productivity Without Epistemic Delegation

The principal implication of VISTA is a distinction between operational delegation and epistemic delegation. Researchers routinely delegate operations to instruments and software: a microscope extends vision, a statistical package executes calculations, and a database retrieves indexed records. Trust is justified through calibration,

documentation, validation, and expert interpretation. AI can be treated similarly when its role is bounded and tested. Epistemic delegation occurs when the researcher accepts the system's account of what is true, relevant, or justified without recovering the evidentiary chain. That move is incompatible with accountable scholarship.

This distinction avoids the false choice between total prohibition and unrestricted use. A researcher may use generated code when the code is understood, tested, documented, and rerun. The same researcher should reject generated references that cannot be located. A reviewer may accept language editing while requiring disclosure of substantial analytical assistance. The burden follows the claim: the more directly an AI output determines evidence, analysis, or interpretation, the more extensive the validation and record must be.

Equity, Bias, and Access

AI can widen access to research practices by supporting translation, programming, explanation, and drafting. These benefits are important for scholars working across languages or without extensive technical support. At the same time, model behavior reflects training data, design choices, and uneven representation. Bender et al. (2021) warn that scale and linguistic fluency can obscure embedded social and environmental costs. Automated systems may reproduce stereotypes, perform unevenly across groups, or normalize a narrow style of scholarly expression.

Fairness concerns also apply to enforcement. Liang et al. (2023) found that several GPT detectors disproportionately classified writing by non-native English authors as AI-generated. Detector scores should therefore not be treated as proof of misconduct. A defensible integrity process examines evidence, documented workflow, source use, version history, and the researcher's ability to explain the work. Institutions should not answer one opaque system with another opaque system.

Institutional Implications

Institutions need task-specific guidance rather than a single rule for all AI use. At minimum, policy should address confidential data, participant information, peer-review material, intellectual property, approved services, disclosure, authorship, validation expectations, and responsibility. Training should include practical verification exercises: locating a generated citation, testing generated code, comparing a summary with its source, evaluating subgroup errors, and documenting a workflow. These skills are more durable than rules tied to a single product version.

Publishers and reviewers can use VISTA to ask focused questions. Was the task permitted? Is every source retrievable? Did AI alter the method? How were outputs tested? Is material influence reported? These questions preserve proportionality. Minor language assistance should not carry the same evidentiary burden as automated classification of study data, while neither use removes author responsibility. Transparency should enable evaluation, not substitute for it.

A Minimum Researcher Record

A practical record can be maintained as a short table or version-controlled note. It should identify the research stage, the task assigned to AI, the reason for using it, the tool and available model designation, the date of use, the type of information submitted, and the output incorporated into the project. It should then state how the output was checked, who performed the check, what errors were found, and what

changed as a result. Prompts need not always be reproduced in a publication, but the internal record should preserve enough of the workflow to explain material influence and to repeat the operation when repetition is technically possible.

The record should focus on decisions rather than accumulating every conversational exchange. If AI suggested ten search phrases but none affected the final search, a brief note is sufficient. If generated code produced a table used in the article, the code, environment, tests, and corrected version should be retained with the analysis materials. If an AI-assisted classification determined which records entered a dataset, the decision rule, validation sample, disagreement process, and performance evidence should be documented. This proportional approach keeps transparency usable while preserving scrutiny where the research record depends on an automated output.

The same record supports handoff within a team. A coauthor or reviewer should not have to infer where generated material entered the work. Clear records allow another researcher to inspect sources, rerun code, revisit exclusions, or challenge an interpretation. They also make correction easier when a model or service changes. Documentation therefore serves more than disclosure: it protects continuity, supports reproducibility, and gives human accountability an observable form.

Research Agenda

VISTA is a conceptual framework and requires evaluation. Future studies could compare error detection, completion time, and reproducibility under different verification protocols; examine how AI support affects researchers with different levels of expertise; and test whether task-level disclosure improves reviewer judgment. Discipline-specific work is needed because acceptable validation differs across laboratory science, qualitative inquiry, computational social science, humanities research, and clinical or participant-facing settings.

Research should also measure the verification burden rather than reporting time saved at the first draft. Net usefulness depends on downstream correction, documentation, and review. Longitudinal studies could examine whether repeated reliance changes methodological understanding, whether access benefits persist after training, and which institutional controls reduce privacy or integrity incidents without excluding beneficial assistance. These questions can turn general enthusiasm or anxiety into testable propositions.

Conclusion

Artificial intelligence is useful to research when it extends scholarly capability without breaking the chain between question, evidence, method, analysis, and claim. It can broaden searches, organize information, propose designs, scaffold code, identify patterns, test alternative explanations, and improve communication. These gains are real but conditional. Fluency is not evidence, automation is not validity, and disclosure is not a substitute for verification.

The VISTA framework provides a compact discipline: Validate the task; Investigate evidence; Structure the method; Test consequential outputs; and Account for AI use. Bounded operations may be delegated, but scholarly judgment and accountability may not. With that boundary, AI can function as a useful instrument of inquiry rather than an unexamined source of claims.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623. https://doi.org/10.1145/3442188.3445922

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694. https://doi.org/10.1007/s11023-020-09548-1

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. npj Digital Medicine, 6, Article 75. https://doi.org/10.1038/s41746-023-00819-6

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248, 1-38. https://doi.org/10.1145/3571730

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. Journal of the Association for Information Science and Technology, 74(5), 570-581. https://doi.org/10.1002/asi.24750

Nature. (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature, 613, 612. https://doi.org/10.1038/d41586-023-00191-1

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., et al. (2015). Promoting an open research culture. Science, 348(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNESCO. (2023). Guidance for generative AI in education and research. https://unesdoc.unesco.org/ark:/48223/pf0000386693

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. Nature, 614, 224-226. https://doi.org/10.1038/d41586-023-00288-7

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., et al. (2023). Scientific discovery in the age of artificial intelligence. Nature, 620, 47-60. https://doi.org/10.1038/s41586-023-06221-2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J. G., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18