

Ethical LLM-Assisted Research: A Framework for Responsible Delegation, Verification, and Epistemic Value

Kalin Stoyanov
Department of Automation
University of Chemical Technology and Metallurgy
email: kalin.stoyanov@uctm.edu

August 26, 2026

*“Evil begins when you begin to treat people as things.
Now, in the age of AI, evil begins
when you begin to treat things as people.”*
— adapted from Terry Pratchett

Abstract

Large language models (LLMs) are becoming routine instruments of scientific research, assisting with literature synthesis, hypothesis development, coding, and formal reasoning. Their use raises a central epistemic question: when parts of scientific reasoning are delegated to an artificial system, what conditions must remain under human control for the resulting knowledge claims to retain epistemic legitimacy and accountable authorship? This paper develops a normative and conceptual framework for analyzing such delegation. Scientific reasoning is treated as a distributed process in which the origin of a contribution may vary between human and machine, while responsibility for its acceptance into the scientific record remains human. The framework distinguishes content origin $O(g)$, completion of human verification $V(g)$, responsibility assignment $R(g)$, accountable human ownership $M(g)$, and epistemic outcome $E(g)$. These constructs separate the provenance of a claim from the process by which it is checked, the epistemic outcome of that checking, and the human responsibility attached to its disposition. The central proposition is that the ethical boundary of LLM-assisted research is determined primarily by adequate verification and accountable human ownership rather than by the degree of machine involvement itself. On this basis, the paper develops the notion of an *epistemic audit*: a structured record of delegation, verification, provenance, and responsibility intended to make AI-assisted reasoning transparent and reviewable. The resulting framework provides a

formal vocabulary for distinguishing responsible cognitive delegation from the transfer or neglect of epistemic responsibility in scientific research.

Keywords: Delegated cognition, Distributed reasoning, Epistemic accountability, Moral authorship, Large language models, Human-machine collaboration, Epistemic auditing, Philosophy of artificial intelligence

1 Introduction

Large language models (LLMs) and related AI systems are increasingly used in scientific work to support literature synthesis, argument development, coding, data analysis, and formal reasoning. Their growing role extends an established feature of scientific practice: cognitive work is routinely distributed across researchers, collaborators, instruments, software, and institutions. LLMs expand the scale and speed of such delegation because a single researcher can now assign cognitively substantial intermediate tasks to a computational system and iteratively incorporate its outputs into a research process.

Recent developments suggest that this role may extend beyond assistance with bounded research tasks. In mathematics, Terence Tao [2026] considers a setting in which AI systems are capable of performing research-level mathematical tasks and shifts attention from the question of technical capability to the goals and values that mathematical research is intended to serve. Klowden and Tao [2026] similarly interpret AI as part of the historical development of tools for creating, organizing, and communicating ideas, while arguing that its integration should remain fundamentally human-centered. A related position paper on AI for formal mathematics describes a prospective transition from systems designed primarily as problem solvers toward research agents capable of participating in open-ended mathematical investigation, with human-AI collaboration identified as a central challenge Jiang et al. [2026]. Together, these developments make the epistemic organization of human-AI research collaboration an increasingly substantive problem rather than a question confined to the use of AI as a writing or computational aid.

This development creates an epistemic problem that is distinct from the question of technical performance. A scientifically useful output may originate partly or extensively from an artificial system, while the resulting claim eventually enters the scientific record under human authorship. The relevant question is therefore how the origin of a contribution, its verification, and responsibility for its acceptance should be related within a hybrid human-machine research process.

The analysis developed here treats LLM-assisted research as a form of *delegated cognition*. This perspective is related to theories of distributed cognition and to the “extended mind” thesis Clark and Chalmers [1998], according to which cognitive activity can extend beyond the boundaries of an individual biological agent. The present framework adopts a more specific concern: how scientific claims generated within such extended systems acquire warrant and accountable authorship.

For this purpose, a distinction is required between the *production* of a candidate claim and its *acceptance* as part of a scientific contribution. Within the framework developed in this paper, an LLM output is treated as a candidate epistemic contribution rather than as independently warranted knowledge. The framework does not require a general metaphysical claim about whether artificial systems can possess epistemic agency. It assumes only that, under current scientific authorship practices and major publication guidance, an LLM is not treated as an author capable of independently assuming responsibility for a published claim International Committee of Medical Journal Editors [2026], World Association of Medical Editors [2023], Nature Methods [2026]. Human researchers therefore remain responsible for deciding whether machine-generated material is understood, verified, and incorporated into the scientific record.

This distinction becomes particularly important because the amount of machine involvement and the degree of human responsibility are separate dimensions. A claim may be substantially machine-generated and nevertheless undergo rigorous human verification. Conversely, a claim may originate entirely from a human researcher and remain inadequately checked. The epistemic and ethical status of a contribution therefore cannot be inferred from its origin alone. What matters is the structure through which candidate outputs are evaluated, accepted, and assigned to accountable human agents.

The central problem addressed in this paper can consequently be stated as follows:

Under what conditions can cognitively substantial work be delegated to an LLM while preserving the verification, accountability, and epistemic legitimacy required of scientific authorship?

To examine this problem, the paper develops a conceptual and formal framework organized around five related constructs. The *origin* function $O(g)$ characterizes the degree of human-machine contribution to a statement g . The *verification* function $V(g)$ records whether the verification procedure appropriate to the statement has been completed. The *responsibility mapping* $R(g)$ identifies the human agent or agents accountable for its evaluation. The *moral-ownership* function $M(g)$ records whether those responsible humans explicitly endorse and accept accountability for the epistemic disposition assigned to the claim. Finally, the *epistemic-value* function $E(g)$ characterizes the epistemic outcome of verification within the framework.

These constructs support a central proposition:

The legitimacy of delegated scientific reasoning depends primarily on accountable verification of the resulting claims, rather than on the proportion of cognitive work performed by a human or an artificial system.

The paper develops this proposition through four contributions. First, it models LLM-assisted research as a structured form of delegated cognition and separates the provenance of a contribution from the responsibility attached to

its acceptance. Second, it introduces a formal vocabulary linking origin, verification, responsibility, moral ownership, and epistemic value. Third, it interprets peer review as an additional layer of distributed verification while preserving the author’s primary responsibility for internal validation. Fourth, it develops the notion of an *epistemic audit*: a structured and reviewable record of delegation, provenance, verification, and responsibility within AI-assisted scientific work.

The purpose of the framework is normative and analytical. It does not provide an empirical measure of scientific quality, nor does it claim that the formal quantities introduced below already constitute validated metrics. Their role is to make explicit relationships that are often left implicit when cognitive work is delegated to AI systems. This distinction is particularly important for the later discussion of epistemic value, where quantitative extensions are treated as directions for further development rather than as established measurement instruments.

The remainder of the paper develops the argument progressively. Section 2 situates delegated reasoning within existing traditions of verification, distributed cognition, authorship, and auditing. Section 3 introduces the formal framework and examines its implications for human–machine collaboration and peer review. Section 4 considers possible extensions toward quantitative measures of epistemic value. Section 5 translates the framework into practical principles for responsible AI-assisted authorship, and Section 6 summarizes the resulting account of verification and human responsibility.

2 Background: Delegation and Verification in Scientific Practice

Scientific reasoning has long depended on forms of cognitive delegation. Researchers rely on collaborators, mathematical notation, experimental instruments, computational software, databases, and formal proof systems to perform parts of the work through which scientific claims are produced. Contemporary artificial intelligence extends this pattern by allowing increasingly complex cognitive tasks to be delegated to computational systems. Formal proof assistants such as *Lean* and *Isabelle* and generative systems such as *GPT* and *Gemini* differ substantially in architecture and epistemic function, but both illustrate a broader point: scientific reasoning can be mediated by artifacts without eliminating the need for human evaluation of the resulting claims.

The distinction between generation and verification is central to this process. Philosophical accounts of mathematical practice have repeatedly emphasized that conjecture generation and justification are different epistemic activities. Lakatos, for example, describes mathematical knowledge as developing through conjecture, criticism, counterexample, and reconstruction rather than through an uninterrupted sequence of initially correct propositions Lakatos [1976]. Related analyses of mathematical practice likewise emphasize the role of investigative and justificatory processes in stabilizing mathematical concepts and results

Schlimm [2011]. These perspectives support a general principle relevant to AI-assisted research: the epistemic standing of a claim depends on the process by which it is examined and justified, rather than solely on the mechanism by which it was initially generated.

This distinction becomes particularly important when reasoning is distributed across humans and artifacts. Hutchins’ account of *distributed cognition* shows that cognitive processes can be realized across coordinated systems containing multiple people and material or symbolic artifacts Hutchins [1995]. Scientific research provides a natural instance of such distribution. Experimental instruments transform physical states into measurements, software performs calculations that would be impractical to execute manually, and collaborators contribute specialized analyses to a common research objective. The final scientific result can therefore emerge from a system whose cognitive labor is distributed across heterogeneous components.

LLMs add a distinctive form of delegation to this established pattern. Unlike a calculator or a conventional deterministic program, an LLM can generate extended linguistic arguments, candidate explanations, code, summaries, and intermediate inferences whose internal derivation is not normally available to the researcher as an inspectable chain of reasoning. The relevant epistemic difficulty is therefore not simply that computation occurs outside the researcher’s mind. Scientific instruments and software have long done so. The more specific difficulty is that an LLM can produce apparently coherent candidate claims whose reliability cannot be inferred from their fluency or from direct inspection of the generative process. Verification must consequently be performed at the level of the resulting claims, evidence, references, calculations, and inferential relations.

Distributed cognition must also be distinguished from distributed responsibility. A scientific result may depend on many cognitive and technical components, while responsibility for accepting and publishing that result remains assigned to identifiable human researchers. This distinction is already implicit in established norms of authorship. Discussions of plagiarism, for example, show that scholarly integrity depends in part on truthful attribution of intellectual contributions and on preserving trust in the provenance of published work Helgesson and Eriksson [2014]. The present paper extends this concern to AI-assisted research: provenance identifies how a contribution was produced, whereas responsibility identifies who accepts the obligation to evaluate and endorse it. The two dimensions are related, but they are not identical.

Contemporary publication guidance makes a similar distinction operational. The International Committee of Medical Journal Editors states that AI-assisted tools should not be listed as authors because authorship requires responsibility for accuracy, integrity, and originality, and it places responsibility for AI-assisted submitted material on human authors International Committee of Medical Journal Editors [2026]. The World Association of Medical Editors likewise treats human accountability and transparency about generative-AI use as central requirements World Association of Medical Editors [2023]. Nature Portfolio guidance similarly requires human-led accountability, verification, and disclo-

sure when AI materially contributes to scholarly work Nature Methods [2026]. These policies are discipline-specific examples rather than universal philosophical premises, but they illustrate the practical institutional context in which the present framework locates responsibility.

This distinction also clarifies why machine assistance does not by itself determine whether delegation is responsible. Extensive computational involvement may coexist with careful human verification, while entirely human reasoning may remain inadequately checked. The relevant normative question is therefore not how much of a result was generated by a machine, but whether the transition from generated material to accepted scientific claim is accompanied by adequate verification and identifiable human responsibility.

A second conceptual resource comes from the literature on auditing. Strathern Strathern [1997], Shore and Wright Shore and Wright [2015], and Power Power [1997] examine how audit practices make institutional activities visible, documentable, and subject to external evaluation. Much of this literature is explicitly critical of the behavioral and institutional consequences of expanding audit regimes. The analogy developed here therefore does not imply that scientific reasoning should be converted into a managerial audit system. Instead, it isolates a narrower and potentially useful property of auditing: the creation of a traceable relation between an activity, the evidence concerning that activity, and the agent responsible for its evaluation.

Applied to AI-assisted research, this property motivates the notion of an *epistemic audit*. An epistemic audit is understood here as a structured account of how cognitively delegated outputs become accepted scientific claims. At minimum, such an account identifies the provenance of relevant contributions, the human agent responsible for their acceptance, and the verification activities through which that acceptance is justified. Its purpose is not to reproduce every intermediate interaction with an AI system, but to make the epistemically significant transitions in the research process reviewable.

The resulting conceptual structure separates three issues that can otherwise become conflated. First, *provenance* concerns where a candidate contribution originates. Second, *verification* concerns whether sufficient grounds exist for accepting it. Third, *responsibility* concerns which human agent assumes accountability for that acceptance. These distinctions provide the basis for the formal framework developed in the following section. They also support the central position of this paper: cognitive labor may be distributed across human and artificial components, while the warrant for publication and the corresponding responsibility remain matters of accountable human judgment.

3 Delegated Reasoning and Distributed Cognition

3.1 Human–Machine Collaboration

Consider a researcher S who coordinates a scientific task with human collaborators $s_1, s_2, \dots, s_n$. Each collaborator may produce candidate contributions such as literature summaries, intermediate derivations, computational results, or proposed interpretations. These contributions enter a broader integrative process in which they are selected, evaluated, combined, and, where necessary, rejected or revised before becoming part of the final scientific work.

Let

$$\mathcal{G}_H = F_S(s_1, s_2, \dots, s_n) \quad (1)$$

denote the set of substantive claims incorporated into a scientific result produced through such a human collaborative workflow. The operator F_S does not represent mere aggregation. It denotes the researcher’s integrative activity, including interpretation, contextualization, consistency checking, and final acceptance of the claims that enter the work.

Now consider a workflow in which an LLM L performs some of the cognitively delegated tasks. The corresponding set of accepted claims may be represented as

$$\mathcal{G}_L = F_S(L), \quad (2)$$

or, more generally, for a mixed human–machine workflow,

$$\mathcal{G}_{HL} = F_S(s_1, \dots, s_n, L). \quad (3)$$

These expressions are not intended to imply that human collaborators and LLMs are interchangeable epistemic or moral agents. They represent alternative distributions of intermediate cognitive labor within a research process. Human collaborators may possess independent expertise, intentions, and responsibilities, whereas the present framework treats the LLM as a source of candidate contributions whose acceptance into the scientific record requires human judgment.

The relevant comparison therefore concerns the criterion applied to the resulting claims rather than the identity of the system that generated them. For any accepted claim

$$g \in \mathcal{G}_H \cup \mathcal{G}_L \cup \mathcal{G}_{HL},$$

the framework requires

$$V(g) = 1, \quad (4)$$

where $V(g)$ records that the claim has undergone the human verification appropriate to its type. A mathematical derivation, an empirical statement, a citation, and an interpretive claim may require different verification procedures, but machine origin does not lower the required standard of warrant.

This formulation separates two stages that are easily conflated in AI-assisted research:

generation of a candidate contribution

↓

human acceptance of a scientific claim

An LLM may participate extensively in the first stage. Within the framework adopted here, the transition to the second stage remains a human act.

Responsibility is therefore attached to accepted claims rather than to the mechanism that generated their precursors. In the simplest case of a single accountable researcher S , let

$$R(g) = \{S\}, \quad g \in \mathcal{G}, \quad (5)$$

where $\mathcal{G}$ denotes the set of claims accepted into the final work. In a multi-author setting, $R(g)$ may instead identify the set of human agents who accept responsibility for the relevant claim. The important point is that $R(g)$ maps a published or publishable claim to accountable human agency; it does not assign responsibility to the LLM itself.

The analogy with a principal investigator and delegated research activity is therefore limited and structural. It concerns the separation between delegated cognitive work and the responsibility for accepting its results. It does not imply that human collaborators can be treated as tools, nor does it determine questions of authorship credit. Human contributors who satisfy the relevant authorship criteria may appropriately be co-authors. The LLM occupies a different position in the present framework because it does not independently assume responsibility for claims entering the scientific record.

This distinction permits substantial machine assistance without making the amount of machine involvement the primary ethical criterion. A predominantly human-generated claim may remain inadequately verified, whereas a substantially LLM-assisted claim may be subjected to rigorous human examination. What determines admissibility within the framework is therefore the passage from candidate generation to accountable acceptance.

The following subsection formalizes this distinction by separating origin, verification, responsibility, moral ownership, and epistemic value at the level of individual claims.

3.2 Formal Framework of Verification and Responsibility

The distinction developed above can be formalized at the level of individual claims. Let $\mathcal{C}$ denote the set of *candidate claims* produced during a research process. A candidate claim may originate from a human researcher, an LLM, or a mixed human-machine interaction. Let $\mathcal{H}$ denote the set of human agents participating in the research process.

The framework distinguishes five properties of a candidate claim: provenance, responsibility, verification, accountability, and epistemic outcome. These properties are related, but they are not interchangeable.

Provenance. For each $g \in \mathcal{C}$, let

$$O(g) \in [0, 1] \quad (6)$$

denote an operationalized provenance coordinate. The limiting cases

$$O(g) = 0 \quad \text{and} \quad O(g) = 1$$

represent purely human and purely machine-generated origin, respectively, while intermediate values represent mixed production.

The numerical value of $O(g)$ is not assumed to be an intrinsic property of the claim. It depends on the provenance convention adopted for a particular application—for example, contribution weights, documented transformations, or another explicitly specified measure of human–machine involvement. Where no defensible scalar operationalization exists, provenance may instead be recorded categorically. The conceptual role of $O(g)$ is descriptive: it records how a candidate contribution was produced. It does not determine whether the contribution is epistemically acceptable.

Responsibility mapping. Responsibility is assigned independently of provenance. Define

$$R : \mathcal{C} \longrightarrow 2^{\mathcal{H}}, \quad (7)$$

where $R(g)$ is the set of human agents accountable for evaluating and, if appropriate, accepting g into the scientific work.

Before such responsibility is assigned,

$$R(g) = \emptyset.$$

For a single-author workflow coordinated by researcher S ,

$$R(g) = \{S\}.$$

In a collaborative project, $R(g)$ may contain more than one human agent.

The codomain of R contains human agents only. This is a normative assumption of the present framework: an LLM may contribute to the production of g , but it is not treated as an agent capable of assuming responsibility for the claim’s publication.

Verification as a process. Verification is represented separately from the epistemic outcome of verification. Define

$$V : \mathcal{C} \longrightarrow \{0, 1\}, \quad (8)$$

where

$$V(g) = \begin{cases} 1, & \text{if the verification appropriate to } g \text{ has been completed,} \\ 0, & \text{if the required verification remains incomplete.} \end{cases}$$

The phrase “verification appropriate to g ” is intentionally claim-dependent. A mathematical derivation may require proof checking; an empirical statement may require examination of data and methodology; a citation claim may require inspection of the cited source; and a computational result may require reproduction or independent validation.

Importantly,

$$V(g) = 1$$

does *not* mean that g has been shown to be true. It means that the required verification process has been performed. Verification may support a claim, reject it, or leave it unresolved. This distinction prevents the act of checking a proposition from being conflated with the outcome of that check.

Responsibility and moral ownership. The assignment of responsibility and the accountable ownership of an epistemic decision are distinct. Let

$$M : \mathcal{C} \longrightarrow \{\perp, 0, 1\}, \quad (9)$$

with

$$M(g) = \begin{cases} \perp, & \text{responsibility unassigned,} \\ 0, & \text{responsibility assigned; endorsement absent,} \\ 1, & \text{responsibility assigned; endorsement present.} \end{cases} \quad (10)$$

Here $\perp$ denotes that responsibility has not yet been assigned, rather than a moral failure. The value $M(g) = 0$ means that a responsible human agent has been identified but has not yet explicitly owned the decision concerning the claim. This is common during an ongoing research process and is not by itself misconduct. The value $M(g) = 1$ records explicit human endorsement and acceptance of accountability for the epistemic disposition assigned to the claim—for example, acceptance as warranted, classification as unresolved, or rejection after verification.

The function $M(g)$ is therefore not determined automatically by $V(g)$. Verification records that the appropriate checking process has been completed; moral ownership records that an identifiable human accepts accountability for what is done with the resulting epistemic assessment. The minimal consistency condition is

$$M(g) = 1 \implies R(g) \neq \emptyset. \quad (11)$$

For a claim presented publicly as warranted, accountable ownership should follow adequate verification. More generally, $M(g)$ is neither a truth value nor an evaluation of the social consequences of the content of g ; it records human accountability for the epistemic disposition of the claim.

Epistemic outcome. The epistemic result of verification is represented by

$$E : \mathcal{C} \longrightarrow \{-1, 0, +1\}, \quad (12)$$

where

$$E(g) = \begin{cases} +1, & \text{if verification supports acceptance as warranted,} \\ 0, & \text{if the epistemic status remains unresolved,} \\ -1, & \text{if verification supports rejection or defeat.} \end{cases} \quad (13)$$

The three values are deliberately interpreted as *epistemic states*, rather than as a cardinal measurement scale. In particular, the notation does not imply that two claims with $E(g) = +1$ possess equal scientific importance, or that epistemic values can already be meaningfully averaged.

The definitions impose the consistency conditions

$$V(g) = 0 \implies E(g) = 0, \quad (14)$$

and

$$E(g) \in \{-1, +1\} \implies V(g) = 1. \quad (15)$$

A completed verification process may therefore lead either to acceptance or rejection. Both outcomes are compatible with responsible scientific practice.

Acceptance into the scientific record. Let $\mathcal{G} \subseteq \mathcal{C}$ denote the set of claims accepted by the researchers as part of the scientific contribution. Within the present framework,

$$\mathcal{G} = \{g \in \mathcal{C} \mid E(g) = +1, M(g) = 1\}. \quad (16)$$

Since $E(g) = +1$ presupposes completed verification, every accepted claim also satisfies $V(g) = 1$. Equation (16) therefore expresses the central normative condition of the framework: a claim enters the accepted body of the work when it has both adequate epistemic warrant and accountable human ownership.

Notably, $O(g)$ does not appear in this acceptance condition. Machine involvement may affect which verification procedures are prudent, how provenance should be documented, or how much scrutiny a particular output requires, but provenance alone neither validates nor invalidates a claim.

This independence can be represented conceptually in Fig. 1.

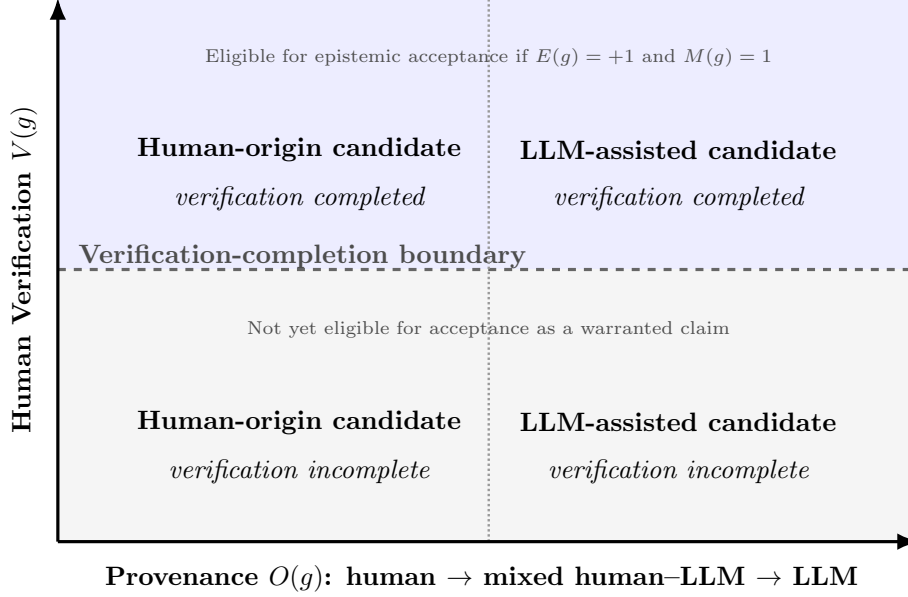


Figure 1: Provenance and verification as independent dimensions of delegated reasoning. The horizontal axis represents the provenance coordinate $O(g)$, while the vertical distinction records whether the required human verification has been completed. Crossing the horizontal boundary does not itself imply that a claim is accepted: accountable acceptance additionally requires $E(g) = +1$ and $M(g) = 1$. No normative threshold is imposed on $O(g)$.

Epistemic state of a candidate claim. The complete state of a candidate claim can consequently be represented as

$$\Sigma(g) = (O(g), R(g), V(g), M(g), E(g)). \quad (17)$$

The components have distinct roles: $O(g)$ records provenance; $R(g)$ assigns human responsibility; $V(g)$ records completion of verification; $M(g)$ records accountable human endorsement of the epistemic disposition; and $E(g)$ records the epistemic outcome of verification.

This representation intentionally avoids collapsing these dimensions into a single numerical score. Whether and under what assumptions such states can be aggregated into a quantitative measure of epistemic value is a separate question considered in Section 4.

3.3 The Moral Economy of Peer Review

Peer review provides an external layer of critical scrutiny between an author’s internally validated work and its acceptance by the scientific community. Its role is therefore related to verification, but it should not be conflated with the

verification responsibility of the author. Authors and reviewers occupy different positions in the epistemic process: authors construct and internally validate a scientific contribution, whereas reviewers independently examine whether the evidence and reasoning offered for that contribution are sufficiently convincing.

This distinction becomes particularly important in LLM-assisted research. Because an LLM can rapidly generate arguments, references, derivations, code, and plausible explanatory text, it can also increase the amount of material whose apparent coherence exceeds its actual warrant. If such material is submitted without adequate internal examination, the cost of detecting its defects is transferred from the researcher who produced the manuscript to the reviewers who evaluate it.

The framework developed in the preceding subsection makes this division of responsibility explicit. Let

$$\mathcal{G}_{\text{assert}}$$

denote the substantive claims that an author presents in a submitted manuscript *as warranted*. From the author’s own epistemic position, such claims should satisfy

$$E(g) = +1 \quad \text{and} \quad M(g) = 1, \quad \forall g \in \mathcal{G}_{\text{assert}}. \quad (18)$$

Equation (18) does not assert that every submitted claim is objectively true, nor that reviewers must agree with the author’s assessment. It expresses a different requirement: a claim presented as warranted should have adequate grounds, according to the verification procedures appropriate to that claim, and identifiable human ownership of that judgment. A submitted manuscript may also contain hypotheses, conjectures, or open questions with $E(g) = 0$, provided that their unresolved status is stated explicitly rather than presented as established knowledge.

Peer review then introduces an independent epistemic challenge to this author-side assessment. A reviewer may identify an error, expose an unsupported inference, request additional evidence, or show that an apparently verified claim remains unresolved. Such disagreement does not imply that the author’s prior verification was meaningless. Rather, it illustrates the value of independent scrutiny in a fallible epistemic system.

The relationship can therefore be represented schematically as

$$\begin{aligned} \text{candidate claim} &\longrightarrow \text{author verification} \longrightarrow \text{author acceptance} \\ &\longrightarrow \text{independent peer scrutiny.} \end{aligned}$$

The sequence is important. Peer review supplements internal verification; it does not replace it.

This structure also clarifies what may be called the *moral economy of peer review*. Scientific publishing depends on a reciprocal allocation of epistemic labor. Authors are expected to perform the primary work of constructing, checking, and defending their claims. Reviewers contribute a second, independent layer of criticism on behalf of the scientific community. The system becomes inefficient

and epistemically fragile when manuscripts are submitted primarily so that reviewers can discover whether their central arguments, references, calculations, or data are reliable.

This principle does not require authors to anticipate every objection or to establish certainty before submission. Scientific claims are inherently fallible, and one purpose of peer review is precisely to expose weaknesses that authors have failed to recognize. The normative distinction lies instead between *fallible but responsible verification* and the deliberate or careless transfer of basic verification work to referees.

LLM-assisted research makes this distinction especially salient because the marginal cost of generating additional text or candidate reasoning can be very low, while the cost of evaluating that material remains substantial. A system capable of producing ten plausible arguments does not thereby provide ten warranted arguments. Each claim incorporated into the manuscript must still pass through the author’s verification and responsibility structure described by $V(g)$, $R(g)$, $M(g)$, and $E(g)$.

Consequently, the ethical problem is not delegation itself. Delegation becomes problematic when it allows the generation of candidate material to outrun the human capacity or willingness to verify what is ultimately presented as scientific knowledge. Responsible use of an LLM therefore preserves an asymmetry between generation and acceptance: generation may be extensively delegated, whereas acceptance into the scientific argument remains subject to human verification and accountable judgment.

Peer review then performs its proper epistemic function. It tests the author’s warranted claims from an independent perspective, rather than serving as the first systematic verification of material that the author has delegated but not adequately examined. In this sense, responsible AI-assisted research does not reduce the significance of peer review. It preserves the conditions under which peer review can operate as a genuinely independent layer of scientific criticism.

3.4 The Epistemic Verification and Improvement Loop

The framework developed above can be interpreted as a recursive process of epistemic control. A useful analogy is provided by the Plan–Do–Check–Act (PDCA) cycle associated with quality improvement and with Deming’s broader account of iterative learning Deming [1986]. The analogy is structural rather than literal: scientific reasoning is not a quality-management process, and epistemic states are not quality-control measurements. PDCA is useful here because it makes explicit the recurrent relation between planning, production, verification, and corrective action.

Within AI-assisted research, the cycle can be interpreted as follows.

Plan: define the epistemic task and assign responsibility. The research objective, relevant constraints, and appropriate verification criteria are specified before candidate outputs are accepted. Human responsibility is also identified

at this stage. For a candidate claim g ,

$$R(g) \neq \emptyset$$

indicates that one or more human agents have been assigned responsibility for evaluating whether the claim may enter the scientific work.

This stage is particularly important when reasoning is delegated to an LLM. The task delegated to the system and the standard subsequently applied to its output are distinct decisions. Delegation specifies what the system is asked to produce; the verification plan specifies what a human researcher must establish before that output can be accepted.

Do: generate candidate epistemic material. Human reasoning, experimentation, computation, LLM assistance, or some combination of these processes produces candidate claims

$$g \in \mathcal{C}.$$

At this stage, generation does not imply acceptance. A fluent explanation, a plausible citation, a mathematical derivation, or an apparent empirical pattern remains a candidate contribution until the relevant verification process has been completed. Its provenance may be recorded through $O(g)$, but provenance does not determine its epistemic status.

Check: perform verification and determine epistemic outcome. The responsible human agent applies the verification procedure appropriate to the claim. Once that procedure has been completed,

$$V(g) = 1.$$

Completion of verification does not by itself determine moral ownership. The result of verification is then represented by

$$E(g) \in \{-1, 0, +1\}.$$

The three possible outcomes have different meanings:

$$E(g) = +1$$

indicates that verification supports acceptance of the claim;

$$E(g) = 0$$

indicates that its epistemic status remains unresolved; and

$$E(g) = -1$$

indicates that verification provides grounds for rejection or defeat.

The Check stage therefore does not exist merely to confirm candidate claims. Its epistemic function includes discovering that a candidate is unsupported, incorrect, or insufficiently determined.

Act: accept, revise, reject, or investigate further. The responsible human agent or agents now decide what epistemic disposition to assign to the candidate and explicitly accept accountability for that decision. When this accountable endorsement is present,

$$M(g) = 1.$$

The action taken then depends on the epistemic outcome. If

$$E(g) = +1 \quad \text{and} \quad M(g) = 1,$$

the claim satisfies the framework's condition for inclusion in the accepted set $\mathcal{G}$.

If

$$E(g) = -1,$$

the claim should not enter $\mathcal{G}$. It may instead be rejected or used to identify a defect in the reasoning, evidence, model, experimental design, or delegation procedure.

If

$$E(g) = 0,$$

additional evidence, analysis, or reformulation is required before an acceptance decision can be made.

Revision may generate a new candidate claim g' . The new claim re-enters the cycle as an epistemically distinct object:

$$g \longrightarrow \text{verification outcome} \longrightarrow \text{revision} \longrightarrow g'.$$

The significance of the loop therefore lies in *error-sensitive iteration*, rather than in the monotonic increase of a numerical epistemic score. Scientific progress may consist in confirming a claim, modifying it, identifying that further evidence is needed, or demonstrating that it should be abandoned. All four outcomes can represent successful epistemic control.

Figure 2 summarizes this interpretation.

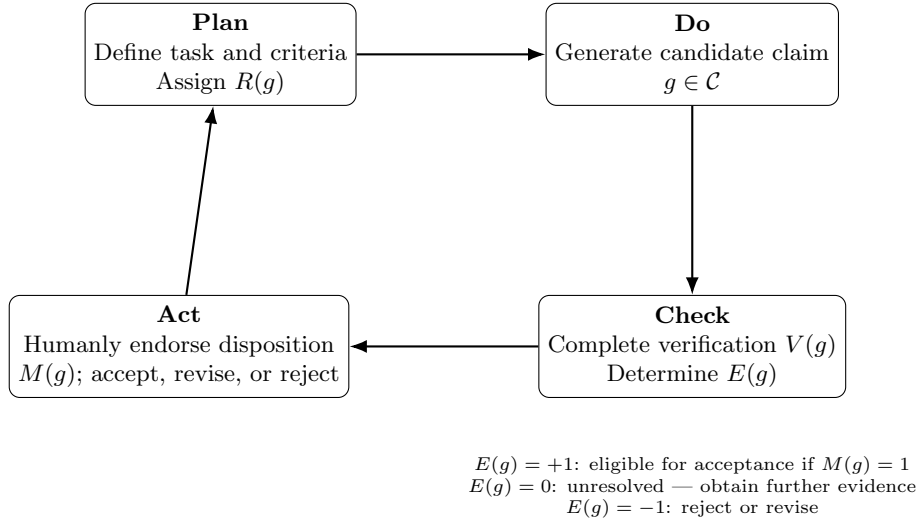


Figure 2: Epistemic verification and improvement loop. Planning assigns responsibility and specifies verification criteria; reasoning or experimentation produces candidate claims; verification determines their epistemic status; and the responsible human agent explicitly owns the resulting disposition before accepting, rejecting, investigating, or revising the claim. The loop improves epistemic control rather than presupposing a monotonic increase in a numerical epistemic-value score.

The PDCA analogy also clarifies the role of LLMs within the framework. An LLM may contribute substantially to the Do stage and may assist with parts of the Check stage, for example by identifying inconsistencies or generating tests. However, assistance with verification does not alter the responsibility mapping: the human agent identified by $R(g)$ remains accountable for deciding whether the verification is adequate and for accepting the resulting claim.

In this sense, responsible delegation is not a one-time act but a controlled cycle. Candidate outputs are repeatedly generated, examined, retained, modified, or discarded. The ethical function of responsibility and the epistemic function of verification thereby become operationally connected: responsibility determines who must perform or supervise the relevant checking, while verification determines which candidate claims may legitimately survive the cycle.

The resulting notion of improvement is consequently broader than an increase in the number of accepted claims. A research process improves when it becomes better able to distinguish warranted claims from unresolved or defeated ones, to identify failures in its own reasoning, and to revise its procedures in response to those failures. This interpretation provides the transition from the claim-level framework developed in this section to the more speculative question considered next: whether aspects of epistemic performance can be represented by defensible quantitative metrics.

4 Future Directions: Toward Operational Measures of Epistemic Value

The framework developed above assigns candidate claims the epistemic state

$$E(g) \in \{-1, 0, +1\}.$$

This representation is intentionally qualitative. It distinguishes claims that verification supports, leaves unresolved, or defeats, but it does not establish a cardinal scale of epistemic quality. A natural direction for future work is therefore to investigate whether richer and empirically defensible measures can be constructed without collapsing distinct epistemic properties into an arbitrary numerical score.

This question is relevant to a broader problem in research evaluation. Citation-based indicators such as citation counts, journal impact measures, and the h -index primarily record forms of visibility and uptake rather than the epistemic reliability of individual scientific claims. Their limitations as proxies for research quality are well documented Fire and Guestrin [2019], Aksnes et al. [2019], Hicks et al. [2015], Lisciandra [2025]. The Leiden Manifesto likewise emphasizes that quantitative indicators should support rather than replace substantive judgment Hicks et al. [2015]. These concerns suggest that any metric derived from the present framework should be treated as an aid to epistemic assessment, rather than as a substitute for scientific evaluation.

The measurement problem. The first requirement is to distinguish *classification* from *measurement*. In Section 3.2, $E(g)$ identifies an epistemic outcome:

$$-1, \quad 0, \quad +1.$$

These labels provide an ordering relevant to acceptance decisions, but the numerical symbols do not establish meaningful distances between the states. It therefore does not follow that expressions such as

$$\frac{1}{n} \sum_{i=1}^n E(g_i)$$

constitute valid measurements of the epistemic quality of a scientific contribution.

A quantitative extension would require additional assumptions and independent operationalization. In particular, it would need to specify what is being measured, how observations are mapped to numerical values, and under what conditions arithmetic operations on those values are meaningful.

The present paper therefore treats quantitative epistemic value as an open measurement problem rather than as an already defined metric.

Requirements for a future epistemic metric. Any defensible quantitative extension of the framework should satisfy at least five requirements.

1. **Construct validity.** The metric must specify the epistemic property it is intended to represent. Logical validity, empirical support, reproducibility, explanatory power, originality, and evidential strength are related properties, but they are not interchangeable.
2. **Claim-type sensitivity.** Verification evidence must be appropriate to the type of claim being evaluated. A mathematical theorem, an empirical observation, a causal inference, a literature claim, and an interpretive argument cannot be assessed by an identical verification procedure.
3. **Traceability.** Every numerical assessment should remain connected to identifiable evidence and verification acts. A metric that produces a score without allowing the evaluator to reconstruct its evidential basis would conflict with the epistemic-audit principle developed in this paper.
4. **Uncertainty representation.** Scientific evidence is rarely complete. A quantitative system should represent uncertainty explicitly rather than forcing unresolved claims into an apparently precise scale.
5. **Calibration and validation.** Proposed scores must be compared with independently observable outcomes, such as successful replication, subsequent correction, independent confirmation, or other discipline-appropriate evidence. Numerical aggregation should be introduced only after the resulting scale has been shown to behave as a meaningful measurement instrument.

These requirements transform the problem from one of assigning convenient numbers to claims into one of constructing and validating an epistemic measurement system.

Internal warrant and external validation. A further distinction is required between the epistemic status of a claim at the time of publication and evidence that accumulates afterward.

Within the framework of this paper, $E(g)$ concerns the result of the verification available when the claim is evaluated. Subsequent scientific developments may provide additional evidence. Examples include

- independent replication or reproduction,
- independent implementations,
- successful or failed predictions,
- later theoretical support or contradiction,
- corrections or retractions, and
- robust adoption in subsequent scientific or technical work.

Such observations can be interpreted as *external validation signals*. They may justify revising an earlier epistemic assessment, but they should not be identified automatically with epistemic value itself. A highly cited result, for example, may later be shown to be incorrect, while an important and well-supported result may initially receive little attention.

A future model can therefore distinguish an internal epistemic state

$$E_t(g)$$

from later evidence that may update that state over time:

$$E_t(g) \xrightarrow{\text{new verification evidence}} E_{t+1}(g).$$

This temporal interpretation is more consistent with scientific practice than a fixed impact score. Epistemic assessment becomes revisable as new evidence enters the scientific record.

Separating epistemic value, scientific impact, and moral consequence.

The original motivation for an epistemic metric also raises a second measurement issue. Three dimensions should remain conceptually distinct:

1. **Epistemic status:** how strongly a claim is warranted;
2. **Scientific or societal impact:** how extensively the claim is used, replicated, implemented, or influential;
3. **Moral consequence:** whether its application produces beneficial or harmful effects.

These dimensions may interact, but one cannot generally be inferred from another. A true scientific result can have harmful applications; a false claim can become highly influential; and a scientifically sound contribution can remain obscure.

For this reason, future quantitative work should prefer a multidimensional representation

$$\mathbf{Q}(g) = (Q_{\text{epi}}(g), Q_{\text{impact}}(g), Q_{\text{moral}}(g))$$

over an immediate reduction to a single scalar score.

The notation $\mathbf{Q}(g)$ represents a proposed measurement architecture, not an operational metric established by the present paper. Each component would require an independent definition, data model, calibration procedure, and uncertainty estimate before meaningful numerical values could be assigned.

Aggregation from claims to scientific contributions. A further open question concerns aggregation. A paper, dataset, model, or scientific theory normally contains multiple claims

$$G = \{g_1, \dots, g_n\}$$

with different significance and evidential status.

Any attempt to construct a contribution-level quantity

$$Q(G)$$

would therefore require a justified aggregation rule. At minimum, such a rule would need to determine

- which claims constitute the relevant epistemic units,
- whether all claims should contribute equally,
- how central and auxiliary claims should be distinguished,
- how dependencies among claims should be represented, and
- how uncertainty should propagate through the aggregation.

These questions are substantive rather than merely mathematical. Until they are answered, a weighted average of claim-level states should be regarded as an illustrative possibility rather than as a validated measure of scientific quality.

Epistemic audits and machine-assisted evaluation. The notion of an epistemic audit provides a possible empirical foundation for such future measurement. An auditable research record can associate claims with their provenance, responsible human agents, verification procedures, supporting evidence, and subsequent external validation.

LLMs and other AI systems may assist in maintaining this structure. Possible functions include

- identifying claims within manuscripts,
- tracing citations and checking bibliographic consistency,
- linking claims to supporting evidence,
- detecting logical or factual inconsistencies,
- recording replication and correction events, and
- identifying claims whose verification status remains unresolved.

Such systems should be understood as instruments for organizing epistemic evidence rather than as autonomous arbiters of scientific truth. In terms of the framework developed here, an AI system may assist with the collection and analysis of evidence relevant to $V(g)$ and $E(g)$, while responsibility for accepting the resulting assessment remains assigned through $R(g)$ to human agents.

Toward self-evaluating discovery systems. This distinction may become increasingly important as AI systems participate in longer scientific reasoning loops. Emerging architectures such as *Real Deep Research (RDR)* Zou et al. [2025] illustrate research workflows in which large models participate in literature analysis, hypothesis generation, and iterative investigation.

In such systems, the framework proposed here suggests a useful architectural separation between

generation, verification, epistemic assessment, human acceptance.

Candidate hypotheses with unresolved status should be preserved as

$$E(g) = 0$$

rather than prematurely promoted to accepted knowledge or discarded merely because they cannot yet be confirmed. Claims defeated by verification can be marked

$$E(g) = -1,$$

while claims supported by the required verification can become eligible for human acceptance.

This provides a possible foundation for what may be termed *epistemic safety*: the design of AI-assisted research systems so that uncertain, defeated, and accepted claims remain explicitly distinguishable. Such systems would aim to limit the propagation of unsupported claims while preserving potentially valuable unresolved hypotheses for further investigation.

The central future challenge is therefore broader than designing a new scalar research metric. It is to develop measurement, provenance, verification, and uncertainty mechanisms capable of representing how scientific claims acquire, lose, and revise epistemic warrant over time. The framework developed in this paper provides a conceptual structure for that programme, while its empirical operationalization remains a subject for future work.

5 Applied Reflection: Responsible Use of AI in Authorship

The framework developed in the preceding sections can be translated into a small number of practical principles for researchers who use LLMs during scientific reasoning, analysis, coding, or writing. These principles are not intended as a universal compliance protocol. Their purpose is to preserve the distinction between delegated generation and accountable scientific acceptance.

The practical question is therefore not whether an LLM participated in the research process, but whether the claims ultimately incorporated into the work can be traced to adequate verification and identifiable human responsibility. In

the notation of Section 3.2, a substantive claim included in the final scientific contribution should satisfy

$$E(g) = +1 \quad \text{and} \quad M(g) = 1. \quad (19)$$

Because $E(g) = +1$ presupposes completed verification, this condition also implies $V(g) = 1$.

The practical principles below specify how this condition can be supported in ordinary research practice.

5.1 Core Principles for Researchers

1. Human control of the research question.

The researcher determines the scientific problem, the intended contribution, the relevant assumptions, and the criteria by which proposed results will be evaluated. An LLM may assist in refining or exploring these elements, but their adoption remains a human research decision.

2. Explicit delegation.

Tasks delegated to an LLM should be sufficiently specified for the researcher to understand what cognitive function the system is performing. Examples include literature search support, generation of alternative hypotheses, code generation, mathematical manipulation, criticism of an argument, or linguistic revision.

Explicit delegation improves provenance because it helps distinguish the production of candidate material from its subsequent evaluation.

3. Claim-appropriate verification.

LLM-generated material should be verified according to the epistemic type of the claim rather than according to a single generic procedure.

For example:

- bibliographic claims require examination of the cited source;
- mathematical claims require derivation or proof checking;
- computational results require inspection or reproduction of the relevant computation;
- empirical claims require examination of the corresponding data and methodology; and
- interpretive claims require evaluation of their evidential and argumentative support.

The relevant condition is completion of the required verification, $V(g) = 1$, rather than confidence based on the fluency of the generated output.

4. Reference authenticity and entailment.

A citation should be checked both for existence and for relevance. The verification task is therefore stronger than confirming that a reference is real: the cited source must actually support the proposition for which it is invoked.

This requirement is especially important for LLM-assisted literature work, where a bibliographically plausible reference and an evidentially adequate reference are distinct possibilities.

5. Human responsibility for acceptance.

Every substantive claim incorporated into the final work should have an identifiable human responsibility assignment,

$$R(g) \neq \emptyset.$$

The responsible researcher need not have generated the claim personally, but must be able to explain why it was accepted, what evidence supports it, and what verification was performed.

6. Proportional traceability.

Documentation should be proportional to epistemic significance. Routine linguistic editing need not generate the same record as an LLM-assisted derivation, literature synthesis, data-analysis procedure, or central theoretical claim.

Traceability should therefore preserve the information necessary to reconstruct epistemically significant transitions:

$$\begin{aligned} \text{delegated task} &\rightarrow \text{candidate output} \\ &\rightarrow \text{verification} \\ &\rightarrow \text{acceptance or rejection.} \end{aligned}$$

7. Transparency of material AI assistance.

Where AI assistance materially affects the reasoning, analysis, code, or construction of the manuscript, its role should be disclosed in accordance with the relevant disciplinary, institutional, or publication requirements International Committee of Medical Journal Editors [2026], World Association of Medical Editors [2023], Nature Methods [2026].

Transparency concerns the role of the system in the research process; it does not transfer authorship responsibility to the system.

8. Authorial understanding and endorsement.

Human authors should understand and endorse the substantive claims for which they accept responsibility. Material that an author cannot adequately explain, defend, or verify should remain outside the accepted claim set $\mathcal{G}$ until those conditions are satisfied.

These principles can be condensed into a practical question that the responsible researcher should be able to answer for every substantive accepted claim:

What is the basis on which I accept this claim, and can I reconstruct the verification that justifies that acceptance?

This formulation places epistemic responsibility on the act of acceptance rather than on the origin of the candidate material.

5.2 Example: Responsible Integration

Consider a researcher using an LLM while developing a theoretical argument.

The researcher asks the system to identify possible weaknesses in an argument and to propose alternative formulations. Several candidate objections and revisions are generated. Some are discarded immediately; others appear substantive and are examined further.

For an LLM-generated claim g_1 , the researcher consults the relevant literature and finds adequate support. Responsibility has been assigned, verification is completed, and the claim is accepted:

$$R(g_1) \neq \emptyset, \quad V(g_1) = 1, \quad M(g_1) = 1, \quad E(g_1) = +1.$$

A second suggestion g_2 appears plausible but relies on a citation that does not support the generated interpretation. Verification therefore defeats the claim:

$$V(g_2) = 1, \quad M(g_2) = 1, \quad E(g_2) = -1.$$

The claim is excluded from the manuscript.

A third proposal g_3 is interesting but cannot be resolved with the evidence currently available:

$$E(g_3) = 0.$$

It is retained as an unresolved possibility or direction for future investigation rather than presented as established knowledge.

This example illustrates an important feature of responsible delegation: successful use of an LLM does not require accepting its outputs. Rejection and suspension of judgment are themselves legitimate outcomes of an effective verification process.

The epistemic contribution of the researcher lies partly in distinguishing among these outcomes.

5.3 Example: Failure of Accountable Delegation

Consider instead a workflow in which an LLM generates a technically fluent section containing several substantive claims and references. The researcher edits the language but does not examine whether the citations support the claims, does not reproduce the calculations, and does not independently inspect the central reasoning.

At this stage, the claims remain candidate outputs for which the required verification is incomplete:

$$V(g) = 0.$$

If responsibility has already been assigned but no accountable disposition has yet been endorsed, then

$$M(g) = 0,$$

and the epistemic state remains

$$E(g) = 0.$$

The mere existence of such unverified material during drafting is not itself a failure of research ethics. Candidate claims are routinely generated before they are checked.

The normative failure occurs if the researcher nevertheless promotes these claims into the submitted scientific argument as though they were warranted. The problematic transition is therefore

$$\text{unverified candidate} \not\rightarrow \text{accepted scientific claim}.$$

Submitting the material without completing the relevant verification transfers basic epistemic work to reviewers and obscures the distinction between generation and justification. The problem is therefore not that the content was generated by an LLM, but that the researcher treated unverified material as accepted knowledge.

An equivalent failure can occur in entirely human reasoning. The framework does not assign a special epistemic defect to machine-generated error; it identifies inadequate verification and unfulfilled responsibility as the relevant failure conditions.

5.4 Practical Epistemic Record

The epistemic-audit principle does not require a complete archive of every prompt, intermediate response, or stylistic edit. Such exhaustive logging could produce substantial documentation without corresponding epistemic value.

Instead, records should preserve information that is material to reconstructing how important claims were produced and accepted. Depending on the research task, a minimal epistemic record may contain four components.

1. Provenance record.

Identify significant uses of AI that materially contributed to a claim, derivation, analysis, code component, literature synthesis, or other substantive element of the work.

The purpose is to preserve relevant information about $O(g)$, not to assign moral significance to the amount of machine involvement.

2. Verification record.

Record the verification action appropriate to important claims: sources consulted, calculations reproduced, code tested, data inspected, or alternative explanations examined.

The record should make it possible to determine why $V(g) = 1$ was assigned.

3. Decision record for critical claims.

For claims central to the scientific contribution, retain enough information to explain why the evidence resulted in

$$E(g) = +1, \quad E(g) = 0, \quad \text{or} \quad E(g) = -1.$$

This is especially useful when an initially plausible AI-generated proposal is modified or rejected.

4. AI-assistance statement.

Where required or scientifically relevant, provide a concise disclosure of the systems used and the substantive functions for which they were employed. The disclosure should describe the role of the tool without implying that responsibility for the resulting claims has been transferred to it.

The amount of documentation should depend on the epistemic importance and risk of the delegated task. A grammar correction and a machine-generated proof of a central theorem do not warrant identical records.

This proportional approach also clarifies the purpose of an epistemic audit. Its objective is not surveillance of every human-machine interaction. It is to preserve sufficient evidence to reconstruct the accountable path by which substantive candidate material became accepted scientific knowledge.

The practical architecture can therefore be summarized as

$\text{delegate} \rightarrow \text{trace} \rightarrow \text{verify} \rightarrow \text{evaluate} \rightarrow \text{accept, revise, or reject}$

with human responsibility maintained throughout the transition from candidate generation to scientific acceptance.

In this sense, responsible AI-assisted authorship does not require minimizing machine participation. It requires ensuring that increasing capacity for delegated generation is matched by an equally explicit capacity for human verification, judgment, and accountable acceptance.

6 Conclusion

The increasing use of large language models in scientific research expands the amount and complexity of cognitive work that can be delegated to computational systems. This development changes the organization of research activity,

but it does not by itself determine the epistemic legitimacy of the resulting claims. The central argument of this paper is that provenance, verification, and responsibility must be treated as distinct dimensions of AI-assisted scientific reasoning.

The framework developed here formalizes this distinction at the level of individual candidate claims. The provenance coordinate $O(g)$ records how a claim was produced; the responsibility mapping $R(g)$ identifies the human agent or agents accountable for evaluating it; $V(g)$ records whether the required verification has been completed; $M(g)$ records whether the responsible human agent or agents explicitly endorse and accept accountability for the epistemic disposition of the claim; and $E(g)$ records the epistemic outcome of verification.

This separation resolves an ambiguity that becomes increasingly important in LLM-assisted research. Generation and justification are different epistemic activities. An LLM may contribute substantially to the generation of a candidate argument, derivation, explanation, program, or literature synthesis without thereby establishing that the resulting claim is warranted. Conversely, machine involvement does not itself diminish the epistemic standing of a claim that has subsequently undergone adequate human evaluation.

Within the framework, the acceptance condition is therefore

$$g \in \mathcal{G} \iff E(g) = +1 \text{ and } M(g) = 1, \quad (20)$$

where $\mathcal{G}$ denotes the set of claims accepted into the scientific contribution.

This condition has two components. The first is epistemic: verification must provide adequate grounds for accepting the claim. The second is responsibility-based: identifiable human agents must explicitly endorse and accept accountability for that acceptance. Provenance $O(g)$ does not itself appear in the acceptance condition.

This does not imply that provenance is irrelevant. Machine involvement may affect the verification procedures that are prudent, the documentation required for traceability, and the degree of scrutiny appropriate to a particular output. Provenance remains important for transparency and reconstruction of the research process. The narrower conclusion is that provenance alone is neither sufficient for acceptance nor sufficient for rejection.

The distinction between verification and epistemic outcome is equally important. Completing verification,

$$V(g) = 1,$$

does not guarantee acceptance. A responsible verification process may yield

$$E(g) = +1, \quad E(g) = 0, \quad \text{or} \quad E(g) = -1.$$

It may support a claim, leave it unresolved, or provide grounds for rejecting it. In each case the verification process can have functioned correctly. The epistemic value of scientific control therefore lies partly in the capacity to discard or suspend judgment on attractive but insufficiently supported candidate claims.

This point is particularly significant for generative AI. LLMs can reduce the cost of producing plausible candidate material without equivalently reducing the cost of establishing its warrant. Responsible AI-assisted research must therefore preserve an asymmetry between generation and acceptance:

generation may be delegated extensively;
scientific acceptance remains accountable.

The framework also clarifies the role of peer review. Reviewers provide an independent layer of epistemic scrutiny, but peer review should not serve as the first systematic verification of claims that authors themselves have not adequately examined. Authors remain responsible for submitting claims that they have reason to regard as warranted, while reviewers test those judgments from an independent perspective. The two forms of verification are complementary rather than interchangeable.

The notion of an *epistemic audit* follows from the same architecture. Such an audit need not preserve every interaction between a researcher and an AI system. Its purpose is narrower: to retain sufficient information to reconstruct the epistemically significant path

delegation $\rightarrow$ candidate output $\rightarrow$ verification
 $\rightarrow$ epistemic assessment $\rightarrow$ human acceptance or rejection.

This makes transparency operational without requiring exhaustive surveillance of the research process.

The framework also provides a basis for future quantitative work, while placing clear limits on what the present theory establishes. The values

$$E(g) \in \{-1, 0, +1\}$$

represent epistemic states rather than a validated cardinal measurement scale. They should therefore not be aggregated automatically into numerical measures of scientific quality. A future epistemic metric would require independently defined constructs, claim-sensitive operationalization, calibration, uncertainty representation, and empirical validation.

For the same reason, epistemic status, scientific impact, and moral consequence should remain conceptually distinct. A result can be well warranted yet have little influence, highly influential yet later defeated, or scientifically sound while producing harmful applications. Future systems for research assessment should preserve these dimensions rather than compressing them prematurely into a single score.

The broader implication is that LLM-assisted science can be understood as a new configuration of an older epistemic structure: cognitive work is distributed, candidate knowledge is generated through heterogeneous means, and scientific

legitimacy is established through accountable processes of verification and judgment. What changes with LLMs is the scale and character of delegated generation. What remains essential is the human responsibility for deciding which generated material deserves to enter the scientific record.

The resulting principle can be stated concisely:

The ethical boundary of AI-assisted scientific reasoning is determined less by where a candidate claim originates than by whether its acceptance is supported by adequate verification and identifiable human responsibility.

Under this view, responsible delegation does not require minimizing the use of artificial intelligence. It requires maintaining a disciplined transition from generated possibility to warranted knowledge. LLMs may expand the space of candidate reasoning; human researchers remain responsible for determining what within that space can legitimately be accepted, defended, and communicated as science.

Generative AI Assistance and Author Responsibility

Generative AI was used during the preparation and revision of this manuscript. OpenAI’s ChatGPT was employed as an assistive tool for language editing, structural reorganization, critical examination of arguments, and the generation of alternative formulations. AI-generated material was treated as candidate content rather than as independently warranted scholarly output.

The research question, conceptual framework, formal definitions, interpretation of the framework, and decisions concerning the inclusion, modification, or rejection of AI-assisted material remained under the control of the author. Substantive claims, references, and formulations retained in the manuscript were reviewed by the author, and cited sources were checked for their relevance to the claims they support. The author reviewed and approved the complete manuscript and accepts full responsibility for its accuracy, integrity, originality, and final content.

The use of generative AI in preparing this manuscript was conducted in accordance with the relevant principles of the World Association of Medical Editors (WAME) recommendations on chatbots and generative artificial intelligence in scholarly publications World Association of Medical Editors [2023]. In particular, the AI system is not identified as an author; its role is disclosed; responsibility for the manuscript remains exclusively with the human author; and AI-assisted material is subject to human verification and source checking.

This disclosure also illustrates the central distinction developed in this paper: AI may participate in the generation and revision of candidate scholarly material, whereas epistemic acceptance and responsibility for publication remain human.

References

- Dag W. Aksnes, Liv Langfeldt, and Paul Wouters. Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, 9(1):1–17, 2019.
- Andy Clark and David J. Chalmers. The extended mind. *Analysis*, 58(1):10–23, 1998. doi: 10.1093/analys/58.1.7.
- W. Edwards Deming. *Out of the Crisis*. MIT Press, Cambridge, MA, 1986. Revised edition, 2000.
- Michael Fire and Carlos Guestrin. Over-optimization of academic publishing metrics: Observing goodhart’s law in action. *GigaScience*, 8(6):1–20, 2019. doi: 10.1093/gigascience/giz053.
- Gert Helgesson and Stefan Eriksson. Plagiarism in research. *Medicine, Health Care and Philosophy*, 18(1):91–101, 2014. doi: 10.1007/s11019-014-9583-8.
- Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. The leiden manifesto for research metrics. *Nature*, 520(7548):429–431, 2015. doi: 10.1038/520429a.
- Edwin Hutchins. *Cognition in the Wild*. MIT Press, Cambridge, MA, 1995.
- International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals: Use of artificial intelligence in publishing. <https://www.icmje.org/recommendations/browse/artificial-intelligence/>, 2026. Updated January 2026. Accessed August 23, 2026.
- Eric Jiang, Xiao Liang, Yikai Zhang, Yingjia Wan, Mengting Li, Haikang Deng, Alexander K. Taylor, Justin Baker, Rushil Raghavan, Junyi Zhang, Ying Nian Wu, Andrea L. Bertozzi, Kai-Wei Chang, Raghu Meka, Matthew Sottile, Nanyun Peng, Amit Sahai, Terence Tao, and Wei Wang. From solvers to research: Large language model-driven formal mathematics at the research frontier, 2026. URL <https://arxiv.org/abs/2607.07779>.
- Tanya Klowden and Terence Tao. Mathematical methods and human thought in the age of ai, 2026. URL <https://arxiv.org/abs/2603.26524>.
- Imre Lakatos. *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press, Cambridge, 1976.
- Chiara Liscandra. Citation metrics: A philosophy of science perspective. *Episteme*, 2025. Forthcoming, Cambridge University Press.
- Nature Methods. Using ai responsibly in scientific publishing. *Nature Methods*, 23:271, 2026. doi: 10.1038/s41592-026-03020-1.

- Michael Power. *The Audit Society: Rituals of Verification*. Oxford University Press, Oxford, 1997.
- Dirk Schlimm. Mathematical concepts and investigative practice. In Bart van Kerkhove and Jean Paul Van Bendegem, editors, *Philosophy of Mathematics: Sociological Aspects and Mathematical Practice*, pages 165–189. Springer, Dordrecht, 2011.
- Cris Shore and Susan Wright. Governing by numbers: Audit culture, rankings and the new world order. *Social Anthropology / Anthropologie Sociale*, 23(1): 22–28, 2015. doi: 10.1111/1469-8676.12098.
- Marilyn Strathern. Improving ratings: Audit in the british university system. *European Review*, 5(3):305–321, 1997. doi: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.
- Terence Tao. Mathematics in the age of ai, 2026. URL <https://arxiv.org/abs/2608.16753>. Essay based on a public lecture delivered at the 2026 International Congress of Mathematicians.
- World Association of Medical Editors. Recommendations on chatbots and generative artificial intelligence in relation to scholarly publication. <https://wame.org/pdf/Chatbots-Generative-AI-and-Scholarly-Manuscripts.pdf>, 2023. Accessed August 23, 2026.
- Xueyan Zou, Jianglong Ye, Hao Zhang, Xiaoyu Xiang, Mingyu Ding, Zhaojing Yang, Yong Jae Lee, Zhuowen Tu, Sifei Liu, and Xiaolong Wang. Real deep research for ai, robotics and beyond. arXiv preprint arXiv:2510.20809, 2025.