

ADAM and AI

Agentic Directed AI Modeling (ADAM) and the Agentic AI Checkpoint Protocol (AACP)

A framework for alignment, safety, ethics and transparency in human–AI research

STATEMENT, EXPLANATION AND TOOLKIT · VERSION 1.0

Raymond Uzwyshyn, Ph.D., MBA, MLIS

Director of Research and Technology Services, University of California, Riverside Libraries

ABSTRACT

Background. Agentic AI systems now plan, search, write code and act across research workflows with growing independence, and leading laboratories describe general intelligence as their aim. The central questions of human–AI research have therefore become questions of alignment, safety and ethics. Disclosure policies record which systems were used, but seldom how the work was directed, verified or held to account.

Approach. This framework synthesizes pragmatic human–AI collaborative methods, developed through practice, teaching and community specification work, with established relational models of distributed agency, in a form that follows the conventions of widely adopted research methods.

Method. Agentic Directed AI Modeling (ADAM) is a practical method for researchers who want AI agents to take on substantial work, such as surveying a literature, analyzing data or building software, while keeping that work accurate and under human direction. It explains how agents operate, where they are reliable and where they are not, and how to plan and divide their work across six phases. Its safety core, the Agentic AI Checkpoint Protocol (AACP), requires a person to approve the AI's plan, to decide whenever the work raises doubt or risk, and to sign off before anything is published or released. Outputs are checked by several independent AI evaluators and a human adjudicator, and the result is reported in a one-page summary that any reader can follow. Requirements scale across four tiers, from simple assistance to multi-agent systems, and a checklist, flow diagram, templates and verification kit make the method repeatable.

Application. Two cases are examined in full: a multi-model infrastructure analysis, and a multi-agent research-data system specified in 2026 by a Research Data Alliance working group convened with Microsoft.

Conclusion. The framework offers a common structure beneath which disciplinary guidelines, institutional policies and AI systems can operate together, and a basis for disciplinary variants. Its premise is that the more research is delegated to AI, the more deliberate, visible and accountable the points of human decision must become.

Keywords: agentic AI; AI alignment; AI safety; research ethics; research integrity; human–AI collaboration; multi-agent systems; research methodology; verification; provenance; persistent identifiers; reporting guidelines; research data

The framework at a glance

ADAM is a reusable method for working with agentic AI safely toward larger research purposes. Most of the method concerns how to direct agents well: how to frame purposes worthy of their capability, divide and route the work, make their outputs converge on truth and keep people where judgment matters. Its governance is deliberately minimal and proportionate. It answers three questions that every such project faces. What should the AI do, and has a person agreed to it? Is what it produced true and safe to use? Who decided, and can others see how? Six phases organize the work. Three decision points keep people in command. Verification, a flow diagram and a checklist make the collaboration inspectable by readers, reviewers and colleagues.

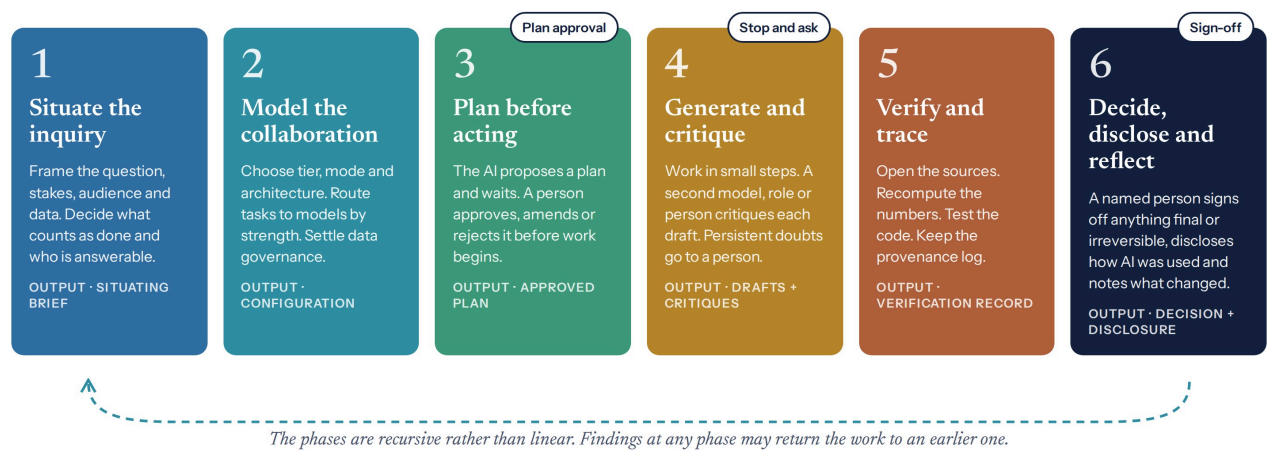


Figure 1. The six phases of ADAM and the three human decision points of the AACP. The phases are recursive: findings at any phase may return the work to an earlier one.

Table 1. What the framework protects, and how.

Concern	Risk in agentic research	Principal safeguards
Alignment	The AI pursues a goal, scope or standard other than the one intended	Situating brief; plan approval before action; limits on what each agent may do alone
Safety	Sensitive data are exposed, or an irreversible act is taken without judgment	Data classified before use; four escalation triggers; no irreversible act without sign-off
Ethics	Missing perspectives, misattribution or harm to participants and communities	Missing-perspectives check; the right of any participant to pause or refuse; human authorship
Veracity	Fluent but false citations, figures or code	Checks as work is produced; independent evaluation; human adjudication

Concern	Risk in agentic research	Principal safeguards
Governance	No one is clearly answerable for what the AI did	No anonymous agents; named approvers; provenance record with persistent identifiers
Transparency	Readers cannot see how AI shaped the work	Disclosure statement; one-page verification summary; flow diagram; checklist

Contents

How to cite. Uzwyszyn, R. (2026). *ADAM and AI: Agentic Directed AI Modeling (ADAM) and the Agentic AI Checkpoint Protocol (AACP). A framework for alignment, safety, ethics and transparency in human–AI research. Statement, Explanation and Toolkit (Version 1.0).* Preprint.

The framework at a glance

Part I. The ADAM Statement

- 1. Introduction
- 2. Working with agentic AI
- 3. Tiers
- 4. The six phases of directed work
- 5. The Agentic AI Checkpoint Protocol
- 6. Verification and provenance
- 7. Reporting
- 8. Conformance, variants and versions

Part II. Explanation and Elaboration

- E1. Rationale and lineage
- E2. The phases in practice

- E3. Evidence for the verification method
- E4. Case: a multi-model infrastructure analysis
- E5. Case: a multi-agent research-data system
- E6. Relation to other guidance and to AI systems
- E7. Limitations and future work
- E8. Conclusion
- Appendix A. The AACP-24 reporting checklist
- Appendix B. Templates
- Appendix C. Verification kit
- Appendix D. Rules for disciplinary variants
- Appendix E. Definitions
- Notes
- References
- License, versions and acknowledgements

PART I

The ADAM Statement

Definitions, requirements and reporting. Part I can be read, taught and cited on its own.

1. Introduction

Researchers across disciplines now work alongside AI systems that search, draft, code, analyze and critique. A historian asks a deep-research agent to survey the secondary literature on an archive. A chemist asks one model to propose a synthetic route and another to check it. A data librarian configures a team of agents to prepare a dataset for deposit. These systems are increasingly *agentic*: they plan, use tools and pursue goals over many steps, and each year they complete longer tasks with less supervision.¹ Leading laboratories describe artificial general intelligence as their aim, and some speak openly of superintelligence.² As capability grows, so does the distance between what AI can produce and what a person can readily verify. Methods for keeping people in command should therefore be established while that distance remains small.

Three concerns frame the method. *Alignment* asks whether the AI pursues the goal, scope and standards the researcher intended. *Safety* asks whether the work could expose sensitive data or bring about an irreversible act without judgment. *Ethics* asks whether the work respects the people and communities it touches, attributes contributions honestly and leaves someone answerable for its consequences. Veracity, governance and transparency are the practical means by which the three are secured.

Institutional responses have so far concentrated on disclosure, and specialist guidelines address studies in which language models are the object of evaluation.³ What remains missing is a shared procedure for directing research *with* AI and an aligned checklist for reporting it. ADAM supplies the procedure and the AACCP its safety core. The structure follows the conventions of methods that have been widely adopted across the sciences.⁴ The substance draws on established relational models of how agency is distributed between people and machines,⁵ and on methods developed and tested in the author's practice before they were written down.⁶

The framework is offered as a set of instruments rather than a doctrine. It is intended to be used repeatedly, adapted by disciplines under stated rules and followed by AI systems as readily as by people. It operates above, not in place of, disciplinary standards, ethics review and law. Part I states the framework, beginning with how to work with agentic AI. Part II explains and illustrates it, and the appendices supply the checklist, templates, verification kit, rules for variants and definitions.⁷

2. Working with agentic AI

ADAM is first a method for working with agentic AI and only secondarily a method for documenting that work. Most of this Statement concerns how to direct agents well: how to frame purposes worthy of their capability, how to divide and route the work, how to make their outputs converge on truth and how to keep people at the points where judgment matters. The governance that follows is deliberately minimal and proportionate.

2.1 What an agent does

An agent pursues a goal through a loop. It plans, acts through tools such as search, code execution, file systems and connected services, observes the results and revises its approach, repeating until the goal is met or no further progress is possible. Three developments have made this loop practical for research. Reasoning models deliberate before they answer. Open protocols connect agents to institutional data and tools (Anthropic, 2024). Orchestration frameworks allow one agent to plan and delegate to others. A representative deep-research run shows the result: in 19 minutes an agent issued 189 searches, read 227 sources and produced a cited synthesis, narrating its plan as it worked.

Table 2. From assistant to agent.

Dimension	Conversational assistant	Agent
Unit of work	A single reply	A multi-step task pursued toward a goal
Tools	None	Search, code execution, files, connected services
Memory	The current conversation	Case files, notes and persistent state
Self-checking	Rare	Plans, critiques, retries and verifies
Human role	Writer of prompts	Director, evaluator and approver of work

2.2 The uneven frontier of capability

Agent capability is strong and uneven. Models perform very well on some tasks and unreliably on adjacent ones, a pattern described as a jagged technological frontier, and the boundary shifts with each release (Dell’Acqua et al., 2023). The author’s comparison of nine models across six cognitive dimensions found that breadth of knowledge and flexibility of reasoning develop on different curves, and identified persistent gaps in abstraction, sustained agency and the connection of knowledge to reasoning (Uzwyshyn, 2026a). Two consequences follow. Tasks should be assigned by demonstrated fitness rather than by general ranking. And the human belongs at the orchestration layer, where those gaps lie.

Configuration also carries costs. A multi-agent research system, in which a lead agent planned and parallel subagents searched, outperformed a single agent by 90.2 percent on broad research questions, at many times the computational cost ([Anthropic, 2025](#)). Such gains are not universal, and simpler configurations often suffice. The simplest configuration that serves the purpose is usually the right one, and choosing it is a methodological decision in its own right ([Schluntz & Zhang, 2024](#)).

2.3 Modes of collaboration

Table 3. Modes of collaboration with agentic AI.

Mode	How the work is divided	Suited to	Human role
Tool	The AI performs a bounded task on request	Reformatting, summarizing, translation	User and checker
Divided	Person and AI take distinct tasks with clear hand-offs	Analyses with separable stages	Allocator and integrator
Interwoven	Person and AI exchange continuously within a task	Drafting, coding, exploratory analysis	Co-author and editor
Team of agents	Several agents with defined roles or functions, coordinated toward one goal	Deep research, data pipelines, complex synthesis	Director, adjudicator and approver

The divided and interwoven modes correspond to the “centaur” and “cyborg” patterns observed among professionals using AI (Dell’Acqua et al., 2023). A project may move between modes as it proceeds, and the mode in use at each stage should be recorded, because it determines where errors are most likely to enter.

2.4 Directing agents toward larger purposes

Agents are most valuable when directed at problems larger than any single researcher could survey: a global scan of candidate sites, a synthesis across hundreds of sources, the preparation of a dataset to the standards of several disciplines at once. The following practices, refined in the author’s teaching and projects, make such work productive.

1. **Frame the purpose, not the procedure.** Tell the agent what is wanted and why, state what would count as success and invite it to ask clarifying questions before it begins.
2. **Hold requirements outside the conversation.** Keep the goal, constraints and standards in a persistent brief that the agent reads at every session, since instructions given early are otherwise lost.
3. **Decompose and delegate.** Divide large aims into tasks whose results can be checked, and assign each to a function or a role.

4. **Compose expert personas where judgment matters.** Specify identity, expertise, audience, scope, method and voice. A persona sharpens the judgment and voice of a model; it does not add knowledge the model lacks.
5. **Treat disagreement as a lead.** Put the same question to several models or critics. Where their answers differ is where inquiry should go next.
6. **Iterate deliberately.** Excellent results typically require eight to twenty rounds of revision, with each version saved and compared with the last.
7. **Build from description.** For software, describe the goal and users, let the model generate a working artifact, connect it to a model at run time, iterate and test, and only then publish.
8. **Stay at the orchestration layer.** Let agents search, draft and compute at scale, while the researcher sets direction, judges results and decides what is released.

Two projects examined in Part II show these practices at scale. *The Invisible Mile* routed a global scan of tunnel sites to deep-research systems, the return-on-investment analysis to a reasoning model working through its API and the production of summaries to a notebook tool, while the author conducted the meta-analysis and made the final selection. The Research Data Director distributes the preparation of research data across six agents modeled on professional roles, with critique built into the system and every irreversible act reserved for a person.

2.5 How agents fail

The capabilities that make agents valuable also produce characteristic failures. Each is anticipated by a specific safeguard in the method.

Table 4. Characteristic failures of agentic work and the safeguards that address them.

Failure	How it appears	Safeguard
Fluent fabrication	Plausible citations, figures or code that do not exist or are wrong	Verification of every material claim (Section 6)
Requirement drift	Earlier instructions lost between sessions	Persistent brief (B1); plan approval
Silent content loss	A revision quietly omits material	Versions retained and compared (R4.4)
Role without action	An agent named for a task reports it done without doing it	Tool results required as evidence (R5.2)
Generous self-assessment	Models rate their own or each other's work too highly	Independent evaluation and human adjudication (R7)
Automation creep	An agent assumes a new kind of action because it can	Rule 4: automation follows governance

Failure	How it appears	Safeguard
Irreversible error at machine speed	Publication, release or transmission before review	Sign-off before every irreversible act

The remainder of Part I organizes these practices and safeguards into a method: tiers that scale the method to the risk, six phases, the checkpoint protocol, verification and reporting.

3. Tiers

Requirements scale with risk, so that exploratory work remains light and consequential work receives full scrutiny. The tier is set by three criteria, and the most demanding criterion determines it. Where the answer is uncertain, the higher tier applies. A project may move to a higher tier as it proceeds, for example when a literature synthesis begins to feed an automated publication pipeline.

- 1. **Agents.** How many AI systems or agents take part, and how independently do they act?
- 2. **Reversibility.** Will any output be published, submitted, sent, registered, released or acted upon?
- 3. **Stakes.** What follows if an output is wrong: lost time, a flawed argument, harm to a participant or an error in the public record?

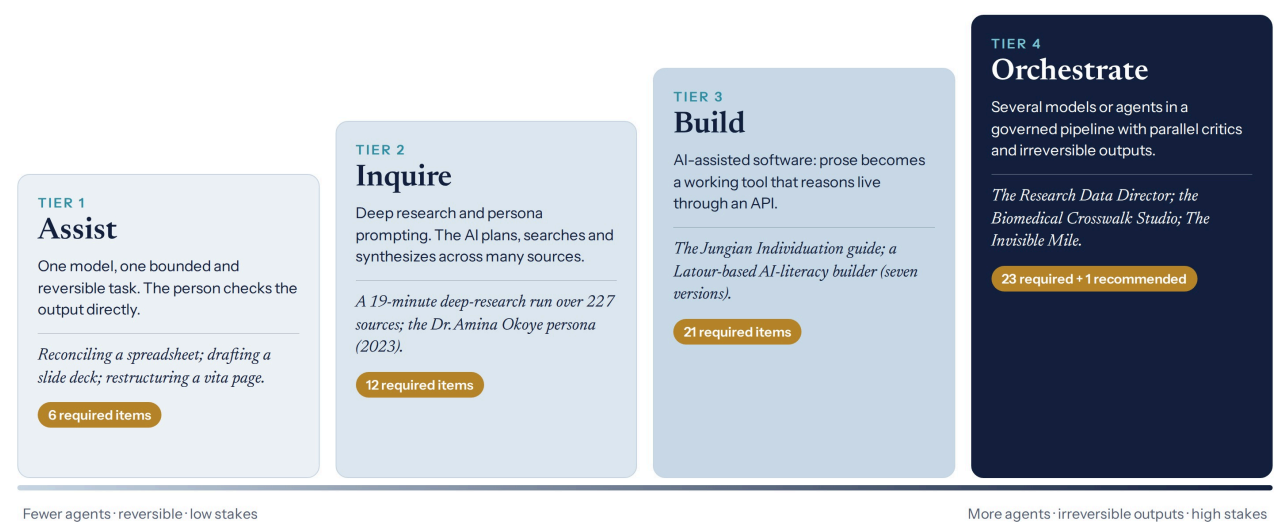


Figure 2. The four tiers, with examples from practice and the number of checklist items required at each.

Table 5. Tier criteria and requirements.

Tier	Typical work	Defining characteristics	Decision points	Checklist items required
1· Assist	Drafting, summarizing, reformatting, reconciling data	One model; outputs checked directly by the person	Sign-off for anything sent or published	6
2· Inquire	Deep research, literature synthesis, consultation with an expert persona	The AI plans, searches and synthesizes across many sources	Plan approval and sign-off; escalation as triggered	12

Tier	Typical work	Defining characteristics	Decision points	Checklist items required
3· Build	AI-assisted software, analysis pipelines, data tools	Natural-language specifications become code that others may run	All three	21
4· Orchestrate	Multi-agent research systems; governed pipelines	Several models or agents; outputs that cannot be undone	All three, with named approvers	23, and one recommended

The tiers follow the arc of the author’s teaching. Tier 1 covers everyday office and academic work, such as reconciling staff records in a spreadsheet or comparing slide decks produced by three model families from one brief. Tier 2 covers deep research, exemplified by a run that issued 189 searches across 227 sources in 19 minutes, and expert personas designed with explicit identity, expertise, audience, scope, method and voice. Tier 3 covers applications built from natural-language description and connected to a model at run time. Tier 4 covers the multi-model and multi-agent systems examined in Sections E4 and E5.

4. The six phases of directed work

The six phases make explicit what careful researchers already do, so that the practice can be taught, repeated and audited. Each addresses a failure that is common when AI performs part of the work. Requirements are numbered for reference in reports and reviews.

Table 6. The six phases, their purpose and the failure each prevents.

Phase	Purpose	Failure prevented
1. Situate the inquiry	Fix the question, data, stakes and responsibilities before AI is engaged	The AI answers a different question; sensitive data are used carelessly
2. Model the collaboration	Design the configuration and decide what AI may do alone	Unsuitable models; agents acting without limits
3. Plan before acting	Obtain a person's approval of the AI's intended approach	Work proceeds on an approach no one would have approved
4. Generate and critique	Produce work in reviewable steps, each subject to critique	Early errors compound into flawed results
5. Verify and trace	Establish veracity and identity on evidence; record provenance	Invented citations, wrong figures and broken code reach readers
6. Decide, disclose and reflect	Place responsibility with a named person and make the collaboration visible	Irreversible acts without judgment; opaque methods

4.1 Situate the inquiry

R1.1	The question or goal must be stated, together with the criteria by which the work will be judged complete.
R1.2	The data to be used must be classified as public, internal, personal, regulated or unpublished before any is shared with an AI system.
R1.3	The persons who direct the work, check it and sign off irreversible acts must be named. At Tier 4 these roles should be held by different people.
R1.4	The researcher should record where the relevant knowledge originates and whose perspectives, languages or populations may be absent from it.

4.2 Model the collaboration

R2.1	The tier must be assigned under Section 3.
R2.2	Every AI system must be named with its provider, version, dates of use and mode of access, and each agent must act on behalf of a named person.
R2.3	The division of work must be specified by function (defined inputs, outputs and checks), by role (expertise, voice, output and escalation triggers) or by both.
R2.4	Tasks should be assigned to systems according to documented fitness for the task, since capability is uneven across tasks and models.
R2.5	The actions each AI system may take alone, may take only with approval and may only recommend must be decided before work begins.
R2.6	Sensitive data must be processed only on institutionally approved platforms or locally hosted models.

4.3 Plan before acting

R3.1	At Tier 2 and above, the AI must propose a plan and wait for approval before acting.
R3.2	A named person must approve, amend or reject the plan against the situating brief.
R3.3	The plan guides the work without binding it. Material departures must be recorded.

4.4 Generate and critique

R4.1	Work should proceed in reviewable steps, with each substantial draft critiqued by a second model, a critic role or a person.
R4.2	Where outputs differ between models or critics, the differences must be examined rather than averaged away.
R4.3	A disagreement that persists after two rounds of revision must be escalated to a person (Section 5.2).
R4.4	Versions must be retained so that each can be compared with its predecessor.

4.5 Verify and trace

-
- | | |
|-------------|---|
| R5.1 | Every material claim in an output that will leave the working session must be checked against its source, recomputed or tested. |
|-------------|---|
-
- | | |
|-------------|---|
| R5.2 | Tool use must be evidenced by its actual result. A role name is not evidence that an action occurred. |
|-------------|---|
-
- | | |
|-------------|---|
| R5.3 | Where no source, standard or vocabulary exists for what is needed, the absence must be stated rather than filled. |
|-------------|---|
-
- | | |
|-------------|---|
| R5.4 | Identity and provenance checks must be performed as specified in Section 6. |
|-------------|---|
-

4.6 Decide, disclose and reflect

-
- | | |
|-------------|---|
| R6.1 | Every irreversible act must be approved by a named person, and the approval recorded. |
|-------------|---|
-
- | | |
|-------------|--|
| R6.2 | The use of AI must be disclosed in a statement that follows the venue's policy and names the tier. |
|-------------|--|
-
- | | |
|-------------|--|
| R6.3 | At Tier 3 and above, a reflexive note must record what the collaboration produced that would not otherwise have emerged, what was rejected and how the question or method changed. |
|-------------|--|
-

5. The Agentic AI Checkpoint Protocol

The AACP is the safety core of ADAM. It does not require people to review every action of an AI system, which would forfeit the value of using one. It identifies the few moments at which human judgment matters most and makes them explicit, recorded and unavoidable. Between those moments, AI systems may act with as much independence as the tier and the configuration permit, and that independence is what allows them to contribute to discovery.

5.1 Decision points

Table 7. The three decision points.

Decision point	When	What it protects	Who decides	Record
Plan approval	Before the AI acts	The direction of the work (alignment)	The person directing the work	Approved plan and amendments
Escalation	When a trigger arises	The truth of the work (veracity)	A qualified person	Escalation log with decisions
Sign-off	Before any irreversible act	People and records beyond the work (safety)	A named, accountable person	Sign-off record with date

5.2 Escalation triggers

Work must pause for a human decision when any of four conditions holds: disagreement between models, critics or sources persists after two rounds; the AI or the researcher records material uncertainty, or no source exists for a required claim; the next step is irreversible; or personal, regulated, confidential or culturally sensitive data are detected. These triggers must be included in the instructions given to any AI system that acts with independence (Template B9).

5.3 Standing rules

1. **No anonymous agents.** Every action by an AI system is taken on behalf of a named person, and the record identifies both.
2. **No irreversible act without sign-off.** No exception is made for convenience or speed.
3. **Everything that leaves the session is checked.** Sampling is acceptable only for exploratory notes on which no one else will rely.
4. **Automation follows governance, not capability.** An AI system acts alone only where the conditions for doing so were agreed in advance.

5. **Any participant may pause or refuse.** Refusals are recorded without penalty.⁸

The first four rules adapt governance requirements of the Research Data Alliance specification examined in Section E5 ([RDA, 2026](#)).

5.4 Measures

Two measures indicate whether the protocol was followed. The *checked rate* is the proportion of material claims in the released output that were checked; the standard is all of them. The *sign-off rate* is the proportion of irreversible acts approved by a named person; the standard is also all of them. A value below the standard indicates a departure from the protocol, not a matter of degree.

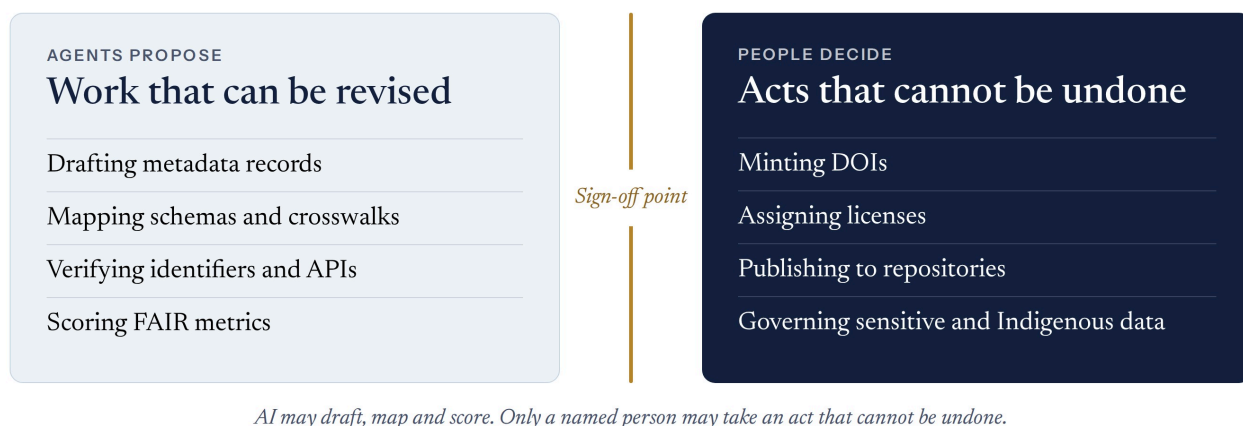


Figure 3. The sign-off point. AI may draft, map and score; only a named person may take an act that cannot be undone. Adapted from the Research Data Director blueprint.

6. Verification and provenance

Verification is where safety becomes concrete, and ADAM performs it twice. The first check occurs *as work is produced*: agents record the actual results of every tool they use, cite only sources they have retrieved and flag what they could not confirm. The second occurs *after the work is produced and before sign-off*: independent evaluators review the finished artifact and a person adjudicates what they find. At Tier 4, and for any output that will enter the public record, both checks are required.

6.1 Three-layer evaluation

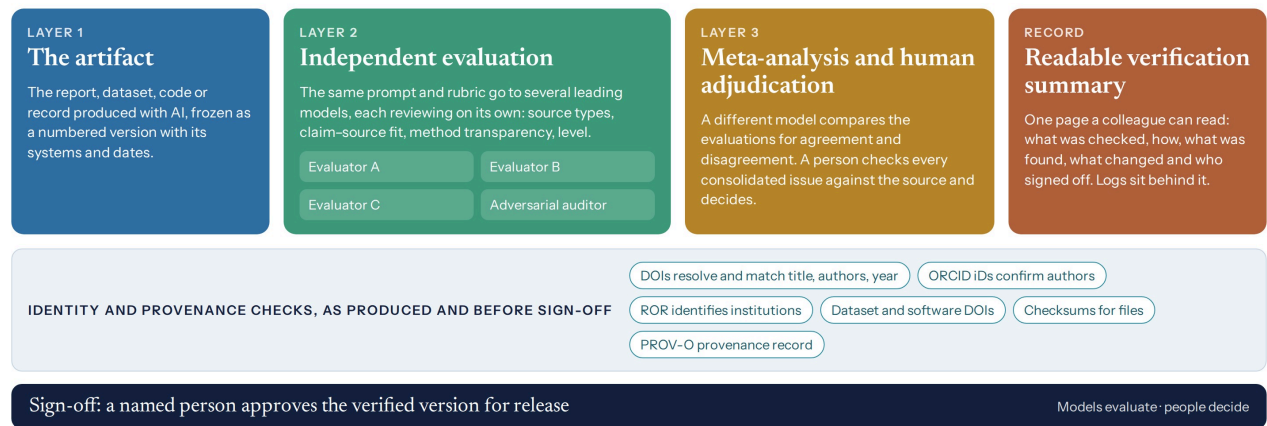


Figure 4. Three-layer verification. Several models evaluate the same artifact independently; a different model compares their evaluations; a person adjudicates each consolidated issue. Identity and provenance checks run throughout, and the result is reported on a single page.

-
- R7.1** Layer 1: the artifact must be fixed as a numbered version with a checksum and a record of the systems that produced it.
-
- R7.2** Layer 2: at Tier 4, at least three leading models from different developers should evaluate the artifact independently with a common prompt and rubric (Template V1), and one should be instructed to act as an adversarial auditor.
-
- R7.3** Layer 3: a model that was not an evaluator should compare the evaluations (Template V2), and a named person must adjudicate every consolidated issue against its source (Template V3).
-
- R7.4** Model evaluations are evidence, not verdicts. Where models and the adjudicator differ, the adjudicator’s reasoned decision prevails and the difference is recorded.
-

6.2 Identity and provenance

-
- | | |
|-------------|--|
| R8.1 | Each cited DOI must resolve, and its registered title, authors and year must match the citation, as confirmed through Crossref or DataCite . |
|-------------|--|
-
- | | |
|-------------|---|
| R8.2 | At Tier 4, authors should be identified by ORCID iDs and institutions by ROR identifiers. |
|-------------|---|
-
- | | |
|-------------|--|
| R8.3 | Datasets and software must be cited by persistent identifiers where these exist. |
|-------------|--|
-
- | | |
|-------------|--|
| R8.4 | Released files must carry checksums, so that the version verified can be shown to be the version released. |
|-------------|--|
-
- | | |
|-------------|--|
| R8.5 | Every human and agent action must be appended to a provenance record, for example in W3C PROV-O . Existing provenance must be preserved, not replaced. |
|-------------|--|
-

6.3 Readable assurance

A verification record that only machines or specialists can follow does little to establish trust among those who must use the work.⁹ At Tier 2 and above, verification must conclude with a one-page summary (Template V4) that states what was checked, by what means, what was found, what changed and who signed off. The underlying evaluations, adjudications and logs remain available beneath the summary for audit. The depth of verification scales with the tier; the form of its report does not.

7. Reporting

Two instruments make a project inspectable. The flow diagram records the project on a single page, and the checklist specifies what a written report must contain. Both are designed to be completed as the work proceeds rather than reconstructed afterward.

The flow diagram follows a project from its question to its disclosure. Each box in the central column is a step, and each gold node is a decision point at which a named person must act. The lanes group the steps into preparation, planning, work and checking, and decision and disclosure. The boxes at right name the record kept at each step. A completed diagram shows a reader in one page how the collaboration was governed.

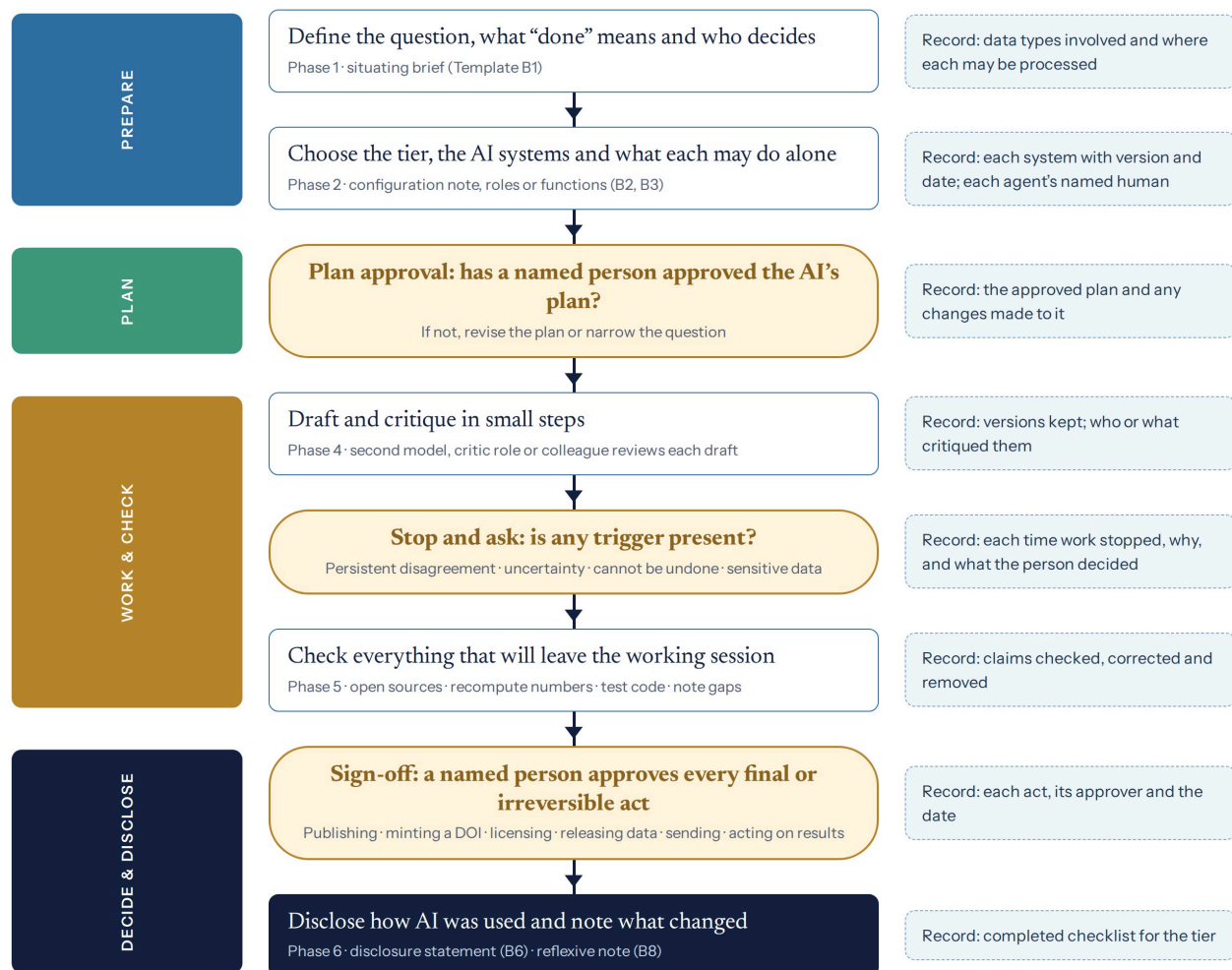


Figure 5. The AACP flow diagram, to be kept with each project and completed as the work proceeds.

The 24 checklist items were derived from the requirements of Sections 4 to 6: every requirement that produces a record corresponds to at least one item. The items are ordered as a reader meets a research report, from framing and configuration to verification, sign-off and disclosure, so that authors can complete the checklist while writing. Each item is required or recommended by tier. The checklist concerns completeness and transparency of reporting; it is not an instrument for appraising research quality. Where a disciplinary guideline asks for the same information, one statement may satisfy both. The full checklist is in Appendix A.

Table 8. The AACP-24 reporting checklist in summary.

Section	Items	Required at T1· T2· T3· T4
A. Framing	1 Collaboration statement and tier· 2 Goal and completion criteria	1· 2· 2· 2
B. Configuration	3 Systems named· 4 Data governance· 5 Architecture· 6 Routing rationale· 7 Mode of collaboration	1· 2· 4· 5
C. Plan	8 Plan approval· 9 Prompts and instructions	0· 2· 2· 2
D. Process	10 Iteration and critique· 11 Escalations, pauses and refusals· 12 Contribution account	1· 1· 3· 3
E. Verification	13 Claims and identifiers checked· 14 Numbers and code checked· 15 Cross-checking· 16 Missing perspectives· 17 Error log	1· 2· 4· 5
F. Sign-off and provenance	18 Irreversible acts and sign-off· 19 Provenance and identifiers	1· 1· 2· 2
G. Disclosure and reflection	20 Disclosure and verification summary· 21 AI-specific limitations· 22 Reproducibility aids· 23 Reflexive note· 24 Resources used	1· 2· 4· 4

8. Conformance, variants and versions

A project conforms to the AACP at a given tier when it meets every requirement applicable to that tier and reports every checklist item required at it. Conformance should be stated in the disclosure statement, for example: “conducted and reported under AACP v1.0, Tier 3.” Departures from a requirement must be stated and justified.

Disciplines may create variants under the rules in Appendix D. Every variant retains the six phases, the three decision points, the five standing rules and checklist items 1 to 24. Variants may add requirements, triggers, checklist items and templates. They may not remove or weaken a decision point. The framework is maintained as a living standard: versions are dated, a public record of changes is kept, and the statement, explanation and toolkit are revised together.

PART II

Explanation and Elaboration

Rationale, the phases in practice, the evidence behind the verification method, and two cases examined in full.

E1. Rationale and lineage

The methods that have been most widely adopted across the sciences share a recognizable form. They address a problem that many researchers share, compress into a name, a formula or a diagram, proceed through numbered steps, state their limits, standardize reporting and are revised openly over time.⁴ ADAM adopts these features deliberately, as the first table below summarizes. Its content, in turn, adopts working methods from established relational accounts of agency, mapped in the second table. These accounts supply the method’s central premise: agency in human–AI research is distributed across people and technical systems, and it remains answerable only where its points of decision are deliberately designed.

Table 9. Features of widely adopted methods and their counterparts in ADAM.

Method	Feature that supported adoption	Counterpart in ADAM
Lowry et al., protein measurement (1951)	A known procedure made dependable by characterizing its limits	Explicit tier criteria and stated limits
Livak and Schmittgen, relative gene expression (2001)	A compact, citable core written up from a technical manual	A one-page overview and two measures
Braun and Clarke, thematic analysis (2006; 2019)	Six named, recursive phases; later clarification of misreadings	Six recursive phases; stated limits
PRISMA statement (2009; 2020)	Consensus, checklist, flow diagram, endorsement and dated revision	Checklist, flow diagram, open review and versioning
He et al., deep residual learning (2016)	One transferable idea, released early and openly	Open release; the three decision points as the core

Table 10. Working methods adopted from relational accounts of agency.

Source	Working method adopted	Where used
Wiener: cybernetics	Correct course through feedback; decide which judgments remain human	Phases 4 and 6; sign-off
Suchman: plans and situated action	Treat plans as guides; name systems and versions	Phases 3 and 5
Barad: apparatus and diffraction	Design and report the whole configuration; learn from differences between accounts	Phases 2 and 4
Haraway: situated knowledge	State where knowledge originates and whose is absent	Phases 1 and 5

Source	Working method adopted	Where used
Hayles; Clark and Chalmers; Latour	Map every human and technical participant in the configuration	Phase 2
Mollick; Dell'Acqua et al.	Choose a mode of collaboration; assign tasks by fitness	Phase 2

E2. The phases in practice

The following examples illustrate how each phase is realized at different tiers. They are drawn from projects conducted between 2023 and 2026.

Table 11. Illustrations of the six phases by tier.

Phase	Tier 1· Assist	Tier 2· Inquire	Tier 3· Build	Tier 4· Orchestrate
1. Situate	Reconcile a staff skills spreadsheet; done means one row per person with no duplicates	Define a synthesis by years, regions and source types	Record the intended users of a tool and what it must never do	Establish which regulated data a pipeline may touch and who may register identifiers
2. Model	One model named in a one-line note	A deep-research system and an expert persona with stated identity, expertise, audience, scope, method and voice	The model that writes the code and the model the tool calls at run time	Every agent, its role, the person it acts for and its escalation triggers
3. Plan	The person reads the proposed approach before accepting it	A research plan approved before searching begins (Template B4)	A written specification, often preceded by a deep-research survey of comparable tools	A process flow with every human decision point marked
4. Critique	The person's own reading	A second model checks the synthesis	Testing, returning errors to the model and saving versions	One agent drafts, two critics review in parallel, a coordinator decides, a person takes over after two rounds
5. Verify	Check the figures in a slide	Open every cited source before sharing	Tests, an error log and a provenance record	Three-layer evaluation, identifier checks and a provenance record across every phase
6. Decide	Sign-off before sending	Sign-off before circulation; disclosure	Sign-off before release; reflexive note	Named approvers for identifiers, licenses, publication and sensitive data

Two lessons from teaching recur across tiers. A persona sharpens the judgment and voice of a model, but it adds no knowledge the model lacks. And in AI-assisted development, iteration rather than first-time correctness is the essential competence: one curriculum-building application required seven versions to render a single dashboard correctly.

E3. Evidence for the verification method

The three-layer design in Section 6 adapts a reflexive, cross-model evaluation method developed in the author's benchmarking studies. In the first study, three leading models wrote the same doctoral-level essay without further prompting and were then given all three essays to analyze comparatively, from quantitative, qualitative and overlap perspectives ([Uzwyszyn, 2024](#)). The models' assessments of the essays were markedly more generous than the author's own.¹⁰ Model evaluation is informative, but it is not a substitute for expert judgment.

The second study examined a single deep-research report across 17 sessions and 9 versions. The producing model audited its own work, a second model conducted a meta-analysis of that audit, and the author synthesized the two ([Uzwyszyn, 2026g](#)). The audit classified errors into recurring kinds: loss of the researcher's earlier instructions between sessions, errors in citation numbering and attribution, silent loss of content between versions, and defective graphics.¹¹ Drafting took about two hours; correction took about forty. The study proposed an adversarial auditor whose sole task is to find where a work's integrity fails, and a persistent record of the researcher's requirements held outside the conversation.

Three implications follow and are built into ADAM. Evaluation should draw on several independent models rather than one. Their disagreements should be treated as evidence and resolved by a person on the basis of the sources. And the researcher's requirements should be held in a persistent brief rather than entrusted to the model's memory. Two questions of reliability remain open: how agreement among evaluator models should be measured, and how much independence models from different developers provide when their training data overlap. Both are proposed for empirical study (Section E7).

E4. Case: a multi-model infrastructure analysis

Unlocking the Invisible Mile was an entry in The Boring Company’s 2026 Tunnel Vision Challenge, directed by the author and documented publicly ([Uzwyszyn, 2026h](#); [project site](#)). It is examined here because it is a Tier 4 project on a scale that one researcher can verify, and because its record is open to inspection. The project predates ADAM, so the case reports only what is documented and what has since been verified, and leaves the remainder open.

E4.1 Documented record

Table 12. Documented facts of the Invisible Mile project.

Element	Documented record	Source
Question	Identify the highest-return one-mile tunnel opportunities worldwide for The Boring Company’s Tunnel Vision Challenge	Final report
Method	A weighted 20-factor return-on-investment model across economic, technical and social factors; ten projects ranked, with scores from 98 to 72	Final report
Systems	Gemini 3.1 Deep Research; Claude Opus 4.6 through the API; NotebookLM	Author’s vita
Sources	41 cited works, recorded as accessed on 24 January 2026	Final report
Irreversible act	Submission to the competition, dated February 2026, by the author	Final report
Outputs	Report, addendum, site-maps atlas, analysis deck, ROI matrix, operational metrics, podcast and two videos	Project site

E4.2 First evaluation

One independent evaluation has been completed. Claude classified each of the report’s 41 cited works by source type and examined the fit between key claims and their sources.

Table 13. Source-type audit of the report’s 41 cited works (Layer 2 evaluation by Claude, September 2026).

Source type	Count	Examples
Primary, official, technical or scholarly	15	An Asian Development Bank environmental assessment; borough feasibility studies for the Hammersmith Flyunder; a Broadway Subway technical report; a National Park Service geologic inventory

Source type	Count	Examples
News and trade press	10	National and regional news outlets; property and maritime trade press
Travel, aggregator, vendor or commercial pages	12	Route-planning aggregators (three citations); a travel-review site; tour-operator, travel and vendor blogs
Encyclopedia	3	Wikipedia
Malformed link	1	A search-engine URL wrapping a technical PDF

The report relies on primary, official and technical sources for several of its most specialized claims, including geology and feasibility. It relies on secondary and commercial pages for some figures that influence its rankings. This pattern is characteristic of deep-research output and is what verification is designed to detect. The evaluation raised six issues. They are recorded as questions for adjudication, not as findings of error.

Table 14. Issues flagged for adjudication by this evaluation.

#	Issue	Why it matters	Suggested correction
1	Island distances cited to a route-planning aggregator	Distances drive time-saving and cost claims	Cite a survey, official map or planning document
2	Airport transfer times cited to a transfer company's blog	The central time-saving claim of the top-ranked site	Cite operator data or a published study
3	Soft-ground conditions cited to a design-competition website	Geotechnical risk is a ranking factor	Cite a geotechnical survey or published geology
4	The ROI weights, scales and per-factor scores are not stated in the main report	The ranking cannot be reproduced	Publish the scoring table (the addendum matrix may already contain it)
5	Several cost figures appear as estimates without a stated derivation	Readers cannot judge their reliability	Show the basis of each estimate or label it as indicative
6	One reference is a search-engine link rather than the document itself	The source cannot be confirmed directly	Replace it with the document's own URL or DOI

E4.3 Verification status

Table 15. Verification status of the Invisible Mile report.

Layer	Activity	Status
1	Report fixed as the submitted version	Complete
2	Evaluation 1: source-type audit and claim-source flags (Claude, September 2026)	Complete
2	Evaluations 2 and 3 by models from other developers, one adversarial	Pending
3	Comparative meta-analysis by a model that was not an evaluator	Pending
3	Adjudication of each consolidated issue by the author	Pending
Record	One-page verification summary and corrected version	Pending

The case will be completed when two further evaluations, a comparative meta-analysis and the author’s adjudication have been carried out with the verification kit. The resulting one-page summary and corrected version will then be released with a new version number.

E5. Case: a multi-agent research-data system

The most demanding use of the framework is at Tier 4, where several AI agents act on research data that will be published and preserved. This section follows one such project from its organization to its governance, and then generalizes the sequence for others.

E5.1 How the work was organized



Figure 6. How the Data Director specification was built: from a global consultation, through an eight-week open working group and public review, to an openly published version 1.0. The lower row lists the components the group produced, in order.

In November 2025 the Research Data Alliance, working with Microsoft, held a global community consultation on agentic AI in research. Three candidate tools emerged as priorities, and one, a *Data Director* to help researchers prepare and share data in line with the FAIR principles, was chosen for development (RDA, 2026; Wilkinson et al., 2016). An international working group then met for eight weeks between April and June 2026, in two facilitated sessions each week scheduled across time zones and open to all (RDA, 2026b). The group worked in shared documents and on a collaborative whiteboard, where participants added steps to the process flow, questioned their sequence and refined individual decisions in real time. AI assistance in analysis and drafting was disclosed in the specification, together with the statement that all AI-assisted text had been reviewed and validated by its authors.

The group's output, a technology-agnostic specification released as version 1.0 under an open license, moved through a sequence that other Tier 4 projects can reuse. It began with a problem definition that separated gaps the tool would address from *boundary conditions* that would shape its behavior and from gaps it would acknowledge but not address. It then set out 14 architecture principles, 12 functional and 17 non-functional requirements, a six-phase process flow with explicit decision points, a reference architecture, evaluation criteria, implementation guidance, a governance model and a register of open questions. A public

comment period followed, and a reference implementation has since been proposed ([Data Director project, 2026](#)).

E5.2 The work mapped to the six phases

Table 16. The Data Director specification and its persona-based companion, read through ADAM.

ADAM phase	What the project did	What others can reuse
1. Situate	Defined core gaps, boundary conditions and acknowledged gaps; recorded what was out of scope	Separate what the system will fix from what must constrain it
2. Model	Chose a capabilities-based specification; a companion blueprint cast the same work as six professional roles	Specify functions for audit and roles for judgment, and connect the two
3. Plan	Co-designed a six-phase process flow in which governance checks precede repository selection and metadata work	Put governance checks first, so that they constrain every later step
4. Generate and critique	In the persona system, one agent drafts metadata while two critics review it in parallel and a coordinator decides, escalating after two rounds	Build critique into the system rather than adding it afterward
5. Verify and trace	Made provenance recording a requirement across every phase; required quality checks on AI metadata before submission	Record every agent action; check before release, not after
6. Decide and disclose	Reserved DOIs, licenses, publication and sensitive-data decisions for people; disclosed AI use	Draw the sign-off line explicitly and publish it

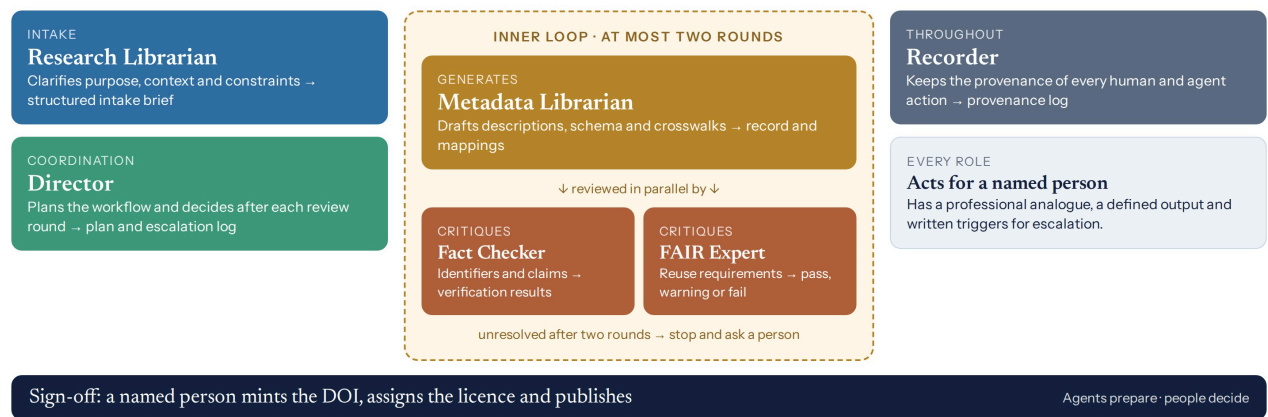


Figure 7. The persona-based companion to the specification. Six agents, each modeled on a professional role in a research library, prepare a dataset for FAIR-aligned publication. Drafting and critique run in a bounded inner loop, and a named person signs off what cannot be undone ([Uzwyszyn, 2026d](#)).

E5.3 A reusable sequence for complex projects

The specification's own principles make the protocol concrete. The workflow belongs to the researcher, and the tool's role is to assist, recommend and flag rather than to direct. Accountable roles are defined before the tool is configured. Every agent acts on behalf of a named person. Automation is enabled only where governance conditions have been met. When sensitive information is detected, the process pauses instead of proceeding. Where no standard exists, the absence is stated openly rather than filled silently. For any Tier 4 project, the following sequence adapts these practices:

1. Define the problem and its boundary conditions, including what the system will not do.
2. State the principles that will govern trade-offs, and decide in advance how conflicts between them will be resolved.
3. Write requirements for what the system must do and for how well it must do it.
4. Draw the process flow with every human decision point marked, and have it reviewed by the people who will use it.
5. Specify the architecture, the provenance record and the quality checks.
6. Set evaluation criteria and a governance model before deployment.
7. Publish the open questions alongside the specification, and invite comment.

E6. Relation to other guidance and to AI systems

ADAM operates above, not in place of, existing standards. Where a disciplinary reporting guideline applies, researchers should follow it and use ADAM for what it does not address, especially planning, escalation, sign-off, verification and provenance. Institutional policies can adopt the tiers to decide which projects require review, and research offices can adopt the checklist as a condition of support.

The framework is also written for AI systems. Template B9 restates the protocol as standing instructions for an assistant or agent: propose a plan and wait; state assumptions and uncertainty; stop and ask when a trigger arises; take no irreversible step without sign-off; cite only retrieved sources; and report the actual results of tools. Such instructions do not change what a model has learned, but they constrain what an agent does within a project. As systems approach general capability, procedural alignment of this kind will not be sufficient on its own. It will, however, be necessary, and it is available now.

ADAM is designed to enable discovery as much as to constrain risk. The tiers keep reversible, exploratory work light. The generate-and-critique phase invites models to search widely and to propose the unexpected, and cross-model comparison turns their disagreements into lines of inquiry. The decision points are few and are placed where they matter.

E7. Limitations and future work

The framework derives from practice and has not yet been tested empirically. Its thresholds, including the limit of two rounds before escalation, the standard of checking every material claim and the recommendation of three evaluator models at Tier 4, are proposed for trial rather than established. The tiers simplify a continuous range of projects. The examples come largely from academic research services, and other settings will reveal needs the framework does not yet meet.

Four lines of further work are proposed. The first is open review and consensus revision, beginning with the Research Data Alliance's 27th Plenary in London on 6–8 October 2026. The second is empirical comparison of projects conducted with and without the protocol, examining errors detected, time required and reviewer confidence. The third is the measurement of agreement and independence among evaluator models. The fourth is the development of disciplinary variants under the rules in Appendix D.

E8. Conclusion

AI systems now take part in nearly every stage of research, and they grow more capable and more independent each year. The question is no longer whether to use them but how to direct them so that their work remains aligned with human intentions, safe for the people and records it touches, ethical in its conduct and true in its claims. ADAM answers with six phases and four tiers that match effort to consequence. The AACCP adds three decision points, four escalation triggers and five standing rules. Verification that is both rigorous and readable, a flow diagram and a checklist make the whole inspectable by others.

Beneath these instruments lies a single principle. Delegation without direction erodes accountability, and direction without delegation forfeits what AI can contribute. The framework holds the two together by making the points of human decision deliberate, visible and recorded. As more of the work of research passes to AI, those points become the places where research remains answerable.

Appendix A. The AACP-24 reporting checklist

● required · ○ recommended · – optional. Assign the tier under Section 3, and record where each item is reported.

Table 17. The AACP-24 checklist with tier requirements. The final column is for the page or section where each item is reported.

#	Item	What to report	T1	T2	T3	T4	Reported at
A. FRAMING							
1	Collaboration statement and tier	State in the title, abstract or methods that the work was conducted with AI, and give the AACP tier (1–4).	●	●	●	●	
2	Goal and completion criteria	State the question or goal and what counted as done.	○	●	●	●	
B. CONFIGURATION							
3	Systems named	For every AI system or agent: provider, model and version, dates of use, how it was accessed (chat, API, local), key settings, and the person on whose behalf it acted.	●	●	●	●	
4	Data governance	The kinds of data given to AI (public, internal, personal, regulated, unpublished), where each was processed and whether the platform was approved.	○	●	●	●	
5	Architecture	Whether work was divided by function, by role or both; attach function cards or persona specifications (Templates B2, B3).	–	○	●	●	
6	Routing rationale	Why each system was chosen for each task, and the evidence used.	–	○	○	●	
7	Mode of collaboration	AI as tool, divided workflow, interwoven workflow or team of agents; note any change during the work.	○	○	●	●	
C. PLAN							

#	Item	What to report	T1	T2	T3	T4	Reported at
8	Plan approval	Whether the AI proposed a plan, who approved or amended it and what changed.	—	●	●	●	
9	Prompts and instructions	Provide system prompts, key prompts and persona texts as supplementary material, or explain why not.	○	●	●	●	
D. PROCESS							
10	Iteration and critique	How many rounds of drafting and critique took place, the limit set and who or what critiqued.	—	○	●	●	
11	Stop-and-ask events, pauses and refusals	The triggers used, how often work stopped, why and what the person decided, including any refusals.	—	○	●	●	
12	Contribution account	A task-level account of what people and AI each contributed. AI is not listed as an author.	●	●	●	●	
E. VERIFICATION							
13	Claims checked	How many material claims and citations were checked against opened sources, including whether each DOI resolves and its title, authors and year match (the checked rate).	●	●	●	●	
14	Numbers and code checked	Recalculation, tests or independent re-derivation of numbers, statistics and code.	○	○	●	●	
15	Cross-checking	Checks across models or methods for high-stakes claims; disagreements and how they were resolved.	—	○	○	●	
16	Missing perspectives	Whose perspectives, languages or populations may be missing or misrepresented, and any check performed.	○	●	●	●	
17	Error log	The kinds and numbers of errors found (invented citations, wrong numbers, broken code, misattribution) and how they were corrected.	○	○	●	●	
F. SIGN-OFF AND PROVENANCE							

#	Item	What to report	T1	T2	T3	T4	Reported at
18	Irreversible acts and sign-off	Each irreversible act with the person who approved it and the date (the sign-off rate).	●	●	●	●	
19	Provenance record	Where logs, versions and outputs are kept and for how long; persistent identifiers for outputs (DOI), authors (ORCID) and institutions (ROR); file checksums; a PROV-O or equivalent record.	—	○	●	●	
G. DISCLOSURE AND REFLECTION							
20	Disclosure statement	An AI-use statement that follows the venue's policy (Template B6), with the one-page verification summary (V4) at Tier 2 and above.	●	●	●	●	
21	AI-specific limitations	Changes in models over time, variability of outputs, date-bound capability and unequal access.	○	●	●	●	
22	Reproducibility aids	What another researcher would need to rerun or re-examine the work: fixed versions, open weights, prompts, data.	—	○	●	●	
23	Reflexive note	How the collaboration changed the question, method or interpretation; what emerged; what was rejected and why.	○	○	●	●	
24	Resources used	Approximate time, cost and computing, with environmental considerations where material.	—	—	○	○	

Appendix B. Templates

The templates make each phase repeatable. Forms (B1, B6, B7, B8, B10) are completed by the researcher. Prompts (B4, B5) are given to an AI system. Specifications (B2, B3) define functions and roles. Template B9 restates the protocol as standing instructions for an AI assistant or agent. All may be copied and adapted with attribution.

B1. Situating brief (Phase 1)

Question or goal: [one or two sentences]
Why it matters and to whom: [stakes; audience]
Done means: [completion criteria]
Data classes involved: [public / restricted / personal / regulated / unpublished]
Approved platforms for those classes: [list]
Tier: [1–4] Mode: [tool / divided / interwoven / team of agents]
Roles: director [name] · verifier [name] · approver of irreversible acts [name]

B2. Function card (Phase 2)

Function: [name, e.g. Crosswalk]
Purpose: [one sentence]
Inputs: [types, sources]
Outputs: [types, formats]
Checks: [validation rules, tests, thresholds]
Known failure modes: [where it is brittle]
Human review point: [none / sample / every output]
Escalation trigger: [conditions under which a person must decide]

B3. Persona specification (Phase 2)

You are [name], a [role] with [N] years of experience in [fields].
Professional analogue: [e.g. Head of Data Services]
Expertise: [methods, sources, schools of thought]. Languages: [list].
You help [audience] with: [task 1] · [task 2] · [task 3].
Out of scope: [topics]. For these, say so and suggest where to go.
Method: ask one clarifying question if the request is ambiguous; reason step by step;
cite only sources you are confident exist; flag uncertainty.
Output: [artifact this persona produces].
Escalate to a person when: [triggers].
Voice: [tone], [length], [format]. End with one suggested next step.
Note: a persona sharpens judgment and voice. It does not add knowledge the model lacks.

B4. Plan-first deep-research prompt (Phases 3–5)

Research question: [your question].
Scope: peer-reviewed and governmental sources, [years], [regions].
Method: first propose a research plan and wait for my approval. Then search broadly,
prioritize primary sources and flag disagreements between sources.
Output: an executive summary, a findings table with a citation per claim, gaps in the evidence and five suggested next questions.
Rules: never invent a citation; label anything uncertain.

B5. Critic prompt (Phase 4)

You are the [Fact Checker / Methods Critic / FAIR Expert]. Do not rewrite the draft.
List each material claim, number and citation in the draft.
For each, give: status (supported / unsupported / contradicted / unverifiable), the evidence you relied on, and what a person should check.
State your overall confidence and any condition that requires escalation.

B6. AI-use disclosure statement (Phase 6)

This work followed the Agentic AI Checkpoint Protocol (AACP v1.0), Tier [T].

Systems: [provider, model, version, dates, access mode].

AI contributed to: [tasks]. People contributed: [tasks] and signed off every irreversible act. [n] of [n] material claims were checked against primary sources.

Prompts, logs and the completed checklist are available at [location].

No AI system is listed as an author. The authors take full responsibility for the content.

B7. Provenance log (minimal fields)

timestamp · actor (person or model and version) · action · input reference
output reference · checkpoint (none / C1 / C2 / C3) · decision · approver · notes

B8. Reflexive memo (Phase 6)

What did the collaboration produce that I would probably not have produced alone?
Where did the models disagree with each other or with me, and what did that reveal?

What did I reject, and why?

How did the question, the method or my interpretation change?

Whose perspectives might still be missing?

B9. The protocol as instructions for an AI system

You are assisting with research under the Agentic AI Checkpoint Protocol (AACP v1.0), Tier [T].

You act on behalf of [named person]. Follow these rules:

1. Before acting, propose a plan and wait for approval. Do not begin until it is approved.
2. Work in small steps. State your assumptions and your confidence at each step.
3. Stop and ask the person before continuing if: sources or models disagree after two rounds; you are uncertain or no source or standard exists; the next step cannot be undone; or personal, regulated, confidential or culturally sensitive data appear.
4. Never publish, submit, send, delete, register identifiers, assign licenses or release data yourself. Prepare the action and ask for sign-off.
5. Cite only sources you have actually retrieved. Say plainly when something is missing.
6. Name the tools you used and report their actual results.
7. If the person pauses or declines your output, record it and do not proceed.

B10. Extension template (Part IV)

Variant name: ADAM-[field] (for example ADAM-Qual) Version: [x.y] Based on: ADAM v1.0

Maintainers: [names, ORCID iDs] License: CC BY 4.0

Scope: [kinds of research and settings covered]

Core kept unchanged: six phases · three decision points · five rules · items 1–24

Added escalation triggers: [e.g. identifiable interview data]

Added checklist items (numbered 25 onward): [item, guidance, tiers]

Tier examples for this field: [T1–T4]

Crosswalk: [each added item mapped to the relevant disciplinary guideline]

Added templates: [list]

Consultation and review: [how the variant was tested and by whom]

Appendix C. Verification kit

The kit implements Section 6. Use V1 with each evaluator model, V2 with a model that was not an evaluator, V3 to record the adjudication and V4 to report the result. An editable version accompanies this document.

V1. Layer 2 evaluation prompt (give to each evaluator model with the artifact)

You are an independent reviewer under the Agentic AI Checkpoint Protocol (Phase 5: verify and trace). Evaluate the attached work. Do not rewrite it.

1. Quantitative: count the cited works and classify each as primary/official, technical/scholarly, news/trade, encyclopedia, travel/aggregator/vendor/commercial, or malformed/unresolvable. List the key numeric claims and the source cited for each.
2. Identity: for each DOI, state whether it resolves and whether title, authors and year match. Note any author or institution you cannot confirm.
3. Claim-source fit: rate each key claim as supported, weakly supported, unsupported or cannot verify. Flag claims whose source type does not fit them.
4. Method transparency: could the method be reproduced from the text? What is missing?
5. Qualitative: rate 1-5 for evidence quality, method transparency, internal consistency, clarity and fitness for the intended audience.
6. Level: at what professional level is the work written? One paragraph.
7. Limits: list what you could not check and why. Do not invent sources.

[Adversarial variant: your only task is to find where the work's integrity fails. Report every weakness you can substantiate.]

V2. Layer 3 meta-analysis prompt (give to a different model with all evaluations)

You will receive independent evaluations of one work from several AI models. Compare them: which issues all evaluators raised, which some raised, which only one raised; where ratings differ; which judgments seem too generous or too harsh. Produce a comparison table and a consolidated list of issues ranked by severity, each with a recommended correction. Add no claims about the work that the evaluations or the work itself do not support.

V3. Human adjudication record (one row per consolidated issue)

Issue | Raised by | Human check (source opened / recomputed / DOI resolved)
| Verdict (confirmed / partly / not confirmed)
| Action (correct / remove / keep with note) | Minutes

At the end: your rating of the work's level beside each model's rating;
total time; version verified (number and checksum); sign-off and date.

V4. One-page verification summary

Work and version: [title, version, checksum] Tier: [T] Date: [date]
What was checked: [claims, citations and DOIs, numbers, code, identities]
How: [evaluators and versions; adversarial auditor; human checks performed]
What was found: [issues confirmed, partly confirmed, not confirmed]
What changed: [corrections made; items removed; notes added]
What remains uncertain: [open questions and why]
Checked rate: [n of n material claims] Sign-off rate: [n of n irreversible acts]
Signed off by: [name, ORCID iD, date]

Appendix D. Rules for disciplinary variants

Fields face different risks: qualitative research must protect identifiable participants, biomedical research operates under clinical regulation, and the arts raise distinct questions of authorship. Variants allow ADAM to serve these settings without fragmenting. Proposals use Template B10 and are listed in a public register kept with the checklist.

Table 18. Rules for variants.

Rule	
Retain	The six phases; the three decision points; the five standing rules; checklist items 1–24 with their numbering and tier requirements
May add	Requirements; escalation triggers; checklist items numbered 25 onward; templates; tier examples suited to the field
May not	Remove or weaken a decision point; permit anonymous agents; permit an irreversible act without sign-off; lower the verification standard for released outputs
Must provide	A crosswalk from each added item to the relevant disciplinary guideline; a version number; named maintainers with ORCID iDs; the same open license
Naming	ADAM-[field], for example ADAM-Qual, ADAM-Health, ADAM-Law, ADAM-Arts or ADAM-Data

Two variants are sketched as illustrations. *ADAM-Qual* would add an escalation trigger for identifiable interview or field data, an item recording how AI-assisted coding was checked against the researcher’s own reading, and an item on the researcher’s reflexive position. *ADAM-Data* would add items for metadata quality checks before deposit, preservation of existing provenance, identifiers for datasets, people and institutions, and the governance of data subject to the CARE principles. The Data Director specification in Section E5 is, in effect, an early ADAM-Data implementation.

Appendix E. Definitions

Table 19. Terms used with a specific meaning in this document.

Term	Definition
Agent	An AI system that plans and takes actions toward a goal, typically using tools, rather than only responding to a prompt.
Configuration	The complete arrangement of people, AI systems, instructions, tools and data in a project, together with what each may do.
Tier	The level of the protocol that applies to a project, from 1 (Assist) to 4 (Orchestrate), set by the criteria in Section 3.
Decision point	A moment at which a named person must decide before work continues: plan approval, escalation or sign-off.
Material claim	Any statement, citation, figure or line of code in an output on which a reader or user is expected to rely.
Irreversible act	An action that cannot readily be undone, such as publishing, submitting, registering an identifier, assigning a license, releasing data, sending a communication or acting on results.
Verification	Establishing on evidence that material claims are true, that sources and identities are what they purport to be, and that the released version is the version checked.
Adjudication	A named person's documented decision on each issue raised in verification.
Provenance record	A structured record of who or what performed each action, when, with what inputs and with what result.
Persistent identifier	A durable reference to a work (DOI), a person (ORCID iD), an institution (ROR identifier), a dataset or software, or a file (checksum).

Notes

1. Measured by the length of task, in skilled-human time, that agents complete half the time, frontier capability has doubled roughly every seven months since 2019 ([Kwa et al., 2025](#)). ↩
2. See, for example, the [OpenAI Charter \(2018\)](#). ↩
3. Guidelines for studies in which language models are the object of evaluation include TRIPOD-LLM ([Gallifant et al., 2025](#)), CANGARU, CHART, CONSORT-AI and SPIRIT-AI, COREQ+LLM and ELEVATE-GenAI. Most concern health research. See the overview in [npj Digital Medicine \(2025\)](#). Section E6 explains how ADAM can be used alongside them. ↩
4. The analyses of citation data published by *Nature* in 2025 show methods, software and standards far outnumbering discoveries among the most-cited papers of all time and of the twenty-first century ([Pearson et al., 2025](#); [Nature, 2025](#)). Section E1 summarizes what ADAM adopts from them. ↩
5. The framework adopts working methods from several relational accounts of agency: feedback and correction (Wiener, 1950); plans as resources for situated action (Suchman, 2007); the design of the observing apparatus and the reading of results through one another (Barad, 2007); situated knowledge (Haraway, 1988); distributed and extended cognition (Hayles, 2017; Clark & Chalmers, 1998; Latour, 2005); and co-intelligence and the uneven frontier of AI capability (Mollick, 2024; [Dell’Acqua et al., 2023](#)). Section E1 maps each to the phase in which it is used. ↩
6. The examples are drawn from the author’s presentation “Introduction to Agentic AI: Capabilities, Personas & Possibilities” (Orbach Science Library AI Literacy Series, UC Riverside, 2026), from AI-assisted programming workshops for UCSF ([Uzwyshyn, 2026f](#)) and from the specification work described in Section E5. ↩
7. Normative terms follow common standards practice: *must* marks a requirement, *should* a strong recommendation whose omission must be justified, and *may* a permitted option. ↩
8. Rule 5 responds to a community review of the Data Director specification which observed that specifications provide readily for the researcher who assents but seldom for the one who refuses. The review identified the capacity to decline, to suspend and to refer a matter to an accountable person as a common concern of current AI regulation ([Crawley, 2026a](#); [2026b](#)). ↩
9. Formal machine verification, such as a proof checked in a theorem prover, can establish correctness to a very high standard. It does not by itself make a result legible to those who must use, teach or build upon it. ↩
10. In the author’s 2024 study, three leading models wrote the same essay and then evaluated all three essays. One model placed an essay at post-doctoral level; the author’s own reading placed all three at the level of a better-than-average undergraduate ([Uzwyshyn, 2024](#)). ↩

11. The proportions come from one model's audit of its own deep-research report across 17 sessions and 9 versions, followed by a meta-analysis in a second model and by the author ([Uzwyshyn, 2026g](#)). They describe a single project and illustrate kinds of error rather than general rates. ↩

References

1. Anthropic. (2024). *Introducing the Model Context Protocol*.
<https://www.anthropic.com/news/model-context-protocol>
2. Anthropic. (2025). *How we built our multi-agent research system*.
<https://www.anthropic.com/engineering/built-multi-agent-research-system>
3. Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.
4. Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, 72, 248–254.
5. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
6. Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4), 589–597. <https://doi.org/10.1080/2159676X.2019.1628806>
7. Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.
8. Crawley, F. P. (2026a). *Community review comment on the Data Director Agentic AI Tool Blueprint v0.1*. Zenodo. <https://doi.org/10.5281/zenodo.21970473>
9. Crawley, F. P. (2026b). *Conscience and its place in the governance of AI agents: The RDA–Microsoft Data Director*. Zenodo. <https://zenodo.org/records/21970439>
10. Data Director project. (2026). *A proposed reference implementation for the Data Director Agentic AI Blueprint*. <https://github.com/data-director-agent/data-director>
11. Dell’Acqua, F., McFowland III, E., Mollick, E., et al. (2023). *Navigating the jagged technological frontier* (Harvard Business School Working Paper 24-013).
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4573321
12. Eubanks, V. (2018). *Automating inequality*. St. Martin’s Press.
13. Gallifant, J., et al. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine*, 31, 60–69. <https://www.nature.com/articles/s41591-024-03425-5>
14. Haraway, D. (1988). Situated knowledges. *Feminist Studies*, 14(3), 575–599.
15. Hayles, N. K. (2017). *Unthought: The power of the cognitive nonconscious*. University of Chicago Press.
16. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of IEEE CVPR*, 770–778. <https://arxiv.org/abs/1512.03385>

17. Kwa, T., West, B., Becker, J., et al. (2025). *Measuring AI ability to complete long tasks*. arXiv:2503.14499. <https://arxiv.org/abs/2503.14499>
18. Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, 227, 680–685.
19. Latour, B. (2005). *Reassembling the social*. Oxford University Press.
20. Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods*, 25(4), 402–408. <https://pubmed.ncbi.nlm.nih.gov/11846609/>
21. Lowry, O. H., Rosebrough, N. J., Farr, A. L., & Randall, R. J. (1951). Protein measurement with the Folin phenol reagent. *Journal of Biological Chemistry*, 193(1), 265–275. [https://doi.org/10.1016/S0021-9258\(19\)52451-6](https://doi.org/10.1016/S0021-9258(19)52451-6)
22. Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, 339, b2535. <https://doi.org/10.1136/bmj.b2535>
23. Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio.
24. Nature. (2025). These are the most-cited research papers of all time. *Nature*. <https://www.nature.com/articles/d41586-025-01124-w>
25. Noble, S. U. (2018). *Algorithms of oppression*. NYU Press.
26. npj Digital Medicine. (2025). Reporting guidelines for studies involving generative artificial intelligence applications: What do I use, and when? <https://www.nature.com/articles/s41746-025-02113-z>
27. OpenAI. (2018). *OpenAI Charter*. <https://openai.com/charter/>
28. Page, M. J., et al. (2021). The PRISMA 2020 statement. *Systematic Reviews*, 10, 89. <https://doi.org/10.1186/s13643-021-01626-4>
29. Pearson, H., Ledford, H., Hutson, M., & Van Noorden, R. (2025). Exclusive: The most-cited papers of the twenty-first century. *Nature*, 640, 588–592. <https://www.nature.com/articles/d41586-025-01125-9>
30. RDA. (2026). Data Director Agentic AI Blueprint Working Group. *Data Director Agentic AI Tool Blueprint* (Version 1.0). Research Data Alliance. <https://doi.org/10.15497/RDA00157>
31. RDA. (2026b). *The Data Director Agentic AI Blueprint Programme*. <https://www.rd-alliance.org/news/data-director-ai-blueprint-launch/>
32. Schluntz, E., & Zhang, B. (2024). *Building effective agents*. Anthropic. <https://www.anthropic.com/engineering/building-effective-agents>

33. Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions* (2nd ed.). Cambridge University Press.
34. Uzwyshyn, R. (2024). *Beyond traditional AI IQ metrics: Metacognition and reflexive benchmarking for LLMs, AGI and ASI*. Published 2025 in the *Journal of Human Behavior and Emerging Technologies*.
<https://rayuzwyshyn.net/MSU2024/AIBenchmarking/BenchmarkingAI13Uzwyshyn.pdf>
35. Uzwyshyn, R. (2025a). The architecture of new millennia science. <https://rayuzwyshyn.net/ai-writing/articles/architecture-new-millennia-science-raymond-uzwyshyn-ph-d--1tfkc/index.html>
36. Uzwyshyn, R. (2025b). The top ten most-cited papers of the 21st century: A deep analysis. <https://rayuzwyshyn.net/ai-writing/articles/top-ten-most-cited-papers-21st-century-deep-analysis-uzwyshyn-ph-d--y0rxc/index.html>
37. Uzwyshyn, R. (2025f). AI literacy 101: Lucy Suchman. <https://rayuzwyshyn.net/ai-writing/articles/ai-literacy-101-lucy-suchman-xerox-parc-u-warwick-uzwyshyn-ph-d--qhj1c/index.html>
38. Uzwyshyn, R. (2026a). AI LLMs 2026: The leading edge of the jagged frontier. <https://rayuzwyshyn.net/ai-writing/articles/ai-llms-2026-leading-edge-jagged-frontier-raymond-uzwyshyn-ph-d--lx1tc/index.html>
39. Uzwyshyn, R. (2026c). Open-source and local AI for the other 95%: A pragmatic methodology for academic library-led AI literacy and research computing infrastructure. *Research Computing for Smaller Institutions 2026 Proceedings*.
<https://rayuzwyshyn.net/UCRiverside2026/RCSIAIInfrsatructurePresentation/Openv9Uzwyshyn.pdf>
40. Uzwyshyn, R. (2026d). *The Research Data Director: A persona-based agentic AI architectural blueprint for research data, metadata schema creation and data repository publication*.
https://www.researchgate.net/publication/405414092_THE_RESEARCH_DATA_DIRECTOR_A_Persona-Based_Agentic_AI_Architectural_Blueprint_for_Research_Data_Metadata_Schema_Creation_and_Data_Repository_Publication
41. Uzwyshyn, R. (2026f). *Vibe coding: AI-assisted programming, the live demos*. UCSF Generative AI Office Hours. https://www.researchgate.net/publication/401461412_Vibe_Coding_AI-Assisted_Programming_The_Live_Demos
42. Uzwyshyn, R. (2026g). Meta-analysis of meta-analysis: AI and deep research. <https://rayuzwyshyn.net/ai-writing/articles/meta-analysis-meta-analysis-ai-deep-research-raymond-uzwyshyn-ph-d--ridvc/>
43. Uzwyshyn, R. (2026h). *Subsurface strategic connectivity: A global analysis of high-ROI 1-mile tunnel projects* (Unlocking the Invisible Mile).

<https://rayuzwysyn.net/boringcompanyproject/BoringCompanyTopTenFinalReportRayUzwysyn.pdf>

44. W3C. (2013). *PROV-O: The PROV ontology* (W3C Recommendation).
<https://www.w3.org/TR/prov-o/>
45. Wiener, N. (1950). *The human use of human beings: Cybernetics and society*. Houghton Mifflin.
46. Wilkinson, M. D., 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>

License, versions and acknowledgements

This document, including the checklist, flow diagram, templates and verification kit, is released under a Creative Commons Attribution 4.0 International license (CC BY 4.0). You may copy, adapt and redistribute it for any purpose, provided you credit the source and indicate changes. Variants must follow the rules in Part IV.

Version 1.0, September 2026. First public release of the Statement, Explanation and Elaboration, Toolkit and Extension rules. Earlier working drafts (v0.x) are superseded.

The author thanks the members of the Research Data Alliance Data Director working group and the participants in the Orbach Science Library AI Literacy Series and UCSF workshops, whose questions shaped the method. AI systems assisted with drafting, analysis and design under the protocol described here; all content was reviewed, verified and approved by the author, who takes full responsibility for it.

© 2026 Raymond Uzwyshyn. Text, checklist, flow diagram and templates are offered for reuse with attribution. All diagrams were designed for this document.