The Audit: From Coherence to Answerability

Argument and Architecture for a Governed Protocol for Load-Bearing Claims in Human and LLM-Assisted Reasoning

Christopher Knorr · Independent Scholar

Preprint · June 2026 · Not peer-reviewed

Christopher Knorr is an independent scholar and an enrolled Pit River tribal member living in diaspora. He speaks only for himself and for no nation, tribal government, council, department, clan, family, community, or Native people as a whole. The protected-context commitments in this paper arise from the author's ethical orientation and from the design logic of the tool, not from any claim to speak for the Pit River Tribe or for Native people collectively.

Abstract

A claim can be fluent, internally organized, and persuasive, and still fail. It may be unsupported by evidence, borrow authority it does not have, overrun its scope, or resist correction by the people and institutions it concerns. In human and large-language-model (LLM) reasoning, where well-formed text is cheap to produce, this failure is easy to miss: coherence gets mistaken for responsibility.

This paper makes one argument and builds one artifact. The argument is that *coherence is necessary but insufficient* for responsible assertion. What coherence leaves out is *answerability* — responsibility to evidence, authority, affected people, limits, origin, consequences, and correction — and answerability is irreducible, because standing, authorization, and correction are normative-social relations between asserting agents and the parties entitled to hold them to account, not relations that internal fit among contents can generate. The artifact is *The Audit*, a governed protocol that puts the distinction to work: it finds the load-bearing claim, treats fallacy labels as diagnostic flags rather than verdicts, tests coherence and then answerability, uses triggers and gates to activate only the review a claim needs, protects contexts that should not be processed merely because they are accessible, issues bounded findings, and presents them in a form a reader can act on.

The paper develops the argument (Part I), presents the architecture as an existence proof that the argument is buildable (Part II), and then locates the project, states the conditions under which it would fail, and applies the tool to itself (Part III). The contribution is methodological and architectural, not empirical validation; the paper closes with falsifiability criteria, a disclosed self-audit, and a development path.

Keywords: load-bearing claims; coherence; answerability; assertion; epistemic responsibility; logical fallacies; argumentation; large language models; AI literacy; protected context; Indigenous data governance; public reasoning; audit protocol.

How to read this paper

This paper can be read in three layers, and a reader may stop at whichever one they need:

1. **The argument** (Part I): why coherence is not enough, and why answerability is a separate requirement. A reader interested only in the philosophical claim can read Part I and stop.
2. **The architecture** (Part II): why The Audit has the structure it has, shown as a worked, governed protocol rather than a slogan. A reader interested in the tool can begin here and treat Part I as background.
3. **The development agenda** (Part III): how the project is positioned, how it could fail, and how it should be tested and improved.

Beneath all three runs a single spine. The tool began as logic-checking; logic-checking proved useful but insufficient; that insufficiency exposed load-bearing claims, then coherence, then answerability, then governance, then presentation. Each layer exists because the previous one was not enough. Readers who want only the through-line can follow that spine and skip the elaboration.

What this paper does not do

To keep the claims bounded, the paper explicitly does **not**:

- empirically validate The Audit — no controlled study here shows it improves reasoning outcomes (Section 18 states what would);
- claim to replace domain expertise, professional judgment, legal advice, cultural authority, peer review, or community governance;
- claim any tribal, cultural, or institutional authority on the author's part (see the author note);
- reduce the method to LLM prompting — it is designed to run as a manual worksheet as well; or
- offer a complete literature review of every adjacent field — Section 17 *locates* the project among neighboring literatures rather than surveying any one of them in full.

Contents

Part I — The Argument: Why Answerability

1. Introduction: fluent but unanswerable claims

The Audit is a tool for slowing down a claim that is carrying too much weight.

A claim can be fluent, plausible, well phrased, internally organized, and still fail. It may fit the surrounding argument but remain unsupported by evidence. It may sound authoritative but borrow standing it does not have. It may be rhetorically forceful but depend on a hidden presupposition. It may be logically tidy but unanswerable to affected people, institutional limits, the historical record, professional standards, or correction. In LLM-assisted writing and reasoning, these failures become easier to miss, because fluency is cheap. A model can produce a clean paragraph, a persuasive argument, a confident summary, or a polished critique that *feels* like judgment while concealing the conditions under which the judgment was made. Recent empirical work gives this worry a concrete, if still preliminary, basis: across mixed-method studies, heavier reliance on AI tools is associated with more cognitive offloading and weaker critical-thinking performance, while *structured* prompting — guided rather than passive use — reduces offloading and improves reasoning quality [Gerlich 2025a; Gerlich 2025b]. The problem, in other words, is not the tool but the absence of structure around it.

This paper has two jobs, and it is organized around them. The first is to make an argument: that **coherence is necessary but not sufficient** for responsible assertion, and that what coherence leaves out — answerability — is a distinct and irreducible requirement. The second is to demonstrate that the argument is buildable: to present The Audit as a governed protocol that operationalizes the distinction, so that the claim "answerability is a separate requirement" is not only argued but instantiated in a working method. Part I develops the argument. Part II presents the architecture as the existence proof. Part III locates the work, states how it could fail, and turns the tool on itself.

The paper's central distinction is between **coherence** and **answerability**. Coherence asks whether the parts of a claim fit together. Answerability asks whether that fit remains responsible to what the claim invokes. A claim can cohere inside an argument while failing to answer to evidence, authority, affected people, limits, origin, consequences, or correction.

Three further distinctions govern the design. The first is between **diagnostic flags** and **audit findings**: a fallacy label may expose a rhetorical or local-inference defect and trigger deeper review, but it does not by itself decide coherence, answerability, or accountability. The second is between **access** and **authority**: a text, record, or protected context may be available to a model or a user, but

availability does not create standing to process it, certify it, or expose it. The third is between **internal reasoning** and **human presentation**: an audit finding is not complete when the analysis is finished; it must be shown in a form that preserves the finding, marks uncertainty, prevents misuse, and can be understood by its intended reader.

The Audit is designed to be **LLM-operable but not LLM-dependent**. It can be run through a prompt packet, used manually as a worksheet, taught as a critical-reasoning protocol, or implemented as a guided interface. A model may help extract assumptions, reconstruct arguments, test coherence, and format findings. But the model does not become the authority. The system must remain inspectable, corrigible, and bounded.

This is a methodological and architectural paper: it proposes a structured method for making load-bearing claims more inspectable, and for preventing fluent coherence from being mistaken for answerable judgment. Its scope limits are stated above, under "What this paper does not do," and revisited in Section 18.

2. Methods and standpoint

This is a methodological design contribution, and its conclusions should be read in light of how it was produced. The instrument it describes was conceived and directed by the author, drawing on prior work distinguishing internal status from external authority across several domains. The paper itself was developed through three interacting activities: the author's conception and editorial direction; the reflexive application of the proposed instrument to the paper's own central claims (Section 18); and substantial large-language-model assistance, used for drafting, for positioning against the literature, and for formalization. The conceptual core originates in the author's direction; the literature is engaged here for positioning rather than as a source from which the framework is derived, and convergence with an existing account is not evidence of derivation.

The standpoint is correspondingly bounded. The paper makes no standpoint-independent claim to validity. Its results are conceptual and procedural, produced within a specific, AI-assisted process, and are offered as such. Where several models produce similar analyses, this indicates reproducibility and legibility, not correctness; a clean output is not validation. These disclosures are not incidental. The framework's governing principle is that a claim must remain answerable to its origin and its limits, and a paper advancing that principle is obliged to meet it. For the same reason, the references have been treated as claims requiring support: the recent empirical and AI-literacy sources were verified by search on 14 June 2026 and are so marked; the canonical philosoph-

ical works are cited in standard editions. No source should be relied upon whose edition, date, and relevance have not been independently checked.

3. Origin: from presupposition exposure to claim audit

The Audit did not begin as a finished theory. It began as a practical question posed to an LLM-assisted reasoning process: *what is this text assuming, what argument is it making, and does that argument actually follow?* The early practice was simple — expose the presuppositional assertions, find the arguments, reconstruct them in syllogistic form where possible, and check the logic.

That origin still locates the problem at the tool's root. The first concern was not whether a model could write a better paragraph; it was whether a fluent paragraph hid unexamined commitments. A text may persuade because its assumptions are buried. A conclusion may appear natural because the premises were never stated. A disagreement may persist because the actual load-bearing claim has never been named. Presupposition exposure makes hidden commitments visible; argument reconstruction clarifies what the text is trying to establish; logic checking reveals missing premises, equivocation, invalid inference, circularity, category shifts, and overbroad conclusions.

But the practice revealed its own limits. A formally valid reconstruction may still rest on unsupported premises. A claim may be internally coherent while borrowing authority it does not have. A claim may be logically tidy but overextended in scope. A claim may involve protected material that should not be processed merely because it is available. And a finding may be accurate yet unusable if the reader cannot tell what passed, failed, narrowed, or remained unresolved. Each early success exposed a new failure mode. Logic checking exposed argument problems; argument problems exposed load-bearing claims; load-bearing claims exposed the limits of coherence; the limits of coherence exposed answerability; answerability exposed governance; governance exposed presentation; presentation exposed the need for calibration and records. The Audit is therefore best understood as a developed architecture, not a long checklist: each layer exists because a prior layer was useful but insufficient.

4. Load-bearing claims

The Audit does not begin by evaluating a whole text. It begins by identifying the claim doing the work.

A **load-bearing claim** is a claim on which a larger argument, decision, interpretation, policy, accusation, justification, recommendation, or public communication depends. If the claim fails, something larger fails with it. Not every claim is load-bearing. Some provide background, some illustrate, some set tone, some introduce the subject. A load-bearing claim does structural work: it may author-

ize a conclusion, justify action, establish moral urgency, ground an institutional decision, define a category, name a harm, establish standing, or make a recommendation appear necessary.

Focusing on the load-bearing claim matters because failures of reasoning are often failures of *load*. A minor factual error may not damage the structure; a hidden premise may collapse it. A true but narrow observation may be carrying a broad policy conclusion. A personal experience may be carrying a community-level generalization. A descriptive claim may be carrying a normative one. A classification may be carrying an authority claim. A coherent argument may be carrying an unsupported premise. The basic move is therefore to find the claim doing the structural work, ask what it must carry, and test whether it can honestly carry that load.

This move also preserves proportionality. Without load-bearing identification, an audit degrades into a list of objections — many problems found, none of them the one that matters. And it lets the tool separate a claim's **true core** from its **overclaimed envelope**. The true core is the part that is supported; the overclaimed envelope is the extra force, scope, certainty, authority, or consequence that exceeds support. Many claims should not be rejected but narrowed: a defensible claim is often hidden inside an overbroad one. The Audit is frequently strongest when, instead of saying "false," it says: this part is supportable, this part overreaches, this part needs evidence, this part requires authority, and this part should not be processed as framed.

5. Fallacies as diagnostic flags, not verdicts

The Audit has an obvious relation to critical thinking and informal logic. It asks for assumptions, arguments, inference, and clarity, and it can notice fallacies. But it should not be reduced to fallacy detection.

A fallacy label can be useful. It may show that a speaker has shifted terms, attacked a person instead of an argument, built a straw version of an opponent's position, assumed the conclusion, appealed to an authority without showing its relevance, or moved from limited evidence to an overbroad conclusion. These are real argumentative problems, catalogued at least since Hamblin's reassessment of the tradition [Hamblin 1970] and refined in contemporary argumentation theory [Walton, Reed & Macagno 2008]. But a fallacy label is not an audit finding; it is a diagnostic flag. A person can declare "that is ad hominem" or "that is circular" without having identified the load-bearing claim, measured the scope of the failure, asked whether the claim can be repaired, or asked whether it is answerable to what it invokes. Fallacy language can become a shortcut that replaces analysis.

The Audit therefore treats fallacy detection as local and provisional. It may trigger deeper review, but it does not decide the deeper questions.

Level	Main question	Output
Fallacy detection	Is there a recognizable argumentative or rhetorical defect?	Diagnostic flag
Logic checking	Does the conclusion follow from the premises?	Local inference status
Coherence checking	Do the claim, argument, definitions, and reasons fit together?	Internal-fit status
Answerability checking	Does the claim remain responsible to what it invokes?	Responsibility-to-ground status
Accountability checking	Who or what can hold the claim or speaker to account?	Relational / governance status

Some fallacies do point toward coherence failures: if a core term changes meaning, if an argument contradicts itself, if a premise assumes the conclusion, the label may identify a real breakdown in internal fit. But even there the label is not enough — the audit must ask what exactly failed, whether it touches the load-bearing claim, whether the claim can be narrowed, and whether the problem is merely communicative or instead undermines coherence, answerability, or use. Nor does fallacy detection establish accountability: it does not say who has standing to correct a claim, who is affected, what authority is invoked, what evidence is owed, or what would count as correction. This is one reason the tool grew beyond its logic-checking origin: invalid inference is one problem, unanswerable coherence is another, and ungoverned use is a third, and the tool must distinguish them.

6. Coherence and its limits

Coherence matters. A claim, argument, or model must have internal fit: its terms should remain stable, its premises should support its conclusion, its categories should not silently shift, its reasons should not contradict one another, its definitions should not change midstream. The Audit therefore includes a coherence check. A coherence failure may involve contradiction, equivocation, a missing premise, a category shift, circularity, an unstable definition, incompatible standards, a conclusion stronger than its premises, unmarked movement between descriptive and normative claims, or conflict between a stated purpose and the actual argumentative work being done.

Paul Thagard's account of explanatory coherence makes this intuition precise: accepting or rejecting a proposition is a matter of maximizing satisfaction of positive constraints (explanation, analogy, deduction) and negative constraints (contradiction, competition) across a network of elements [Thagard 2000]. This is a developed model, not stylistic neatness, and its principle of *data priority* even grants observational reports a degree of independent standing. The Audit shares this structural intuition: a claim should be tested not by whether it sounds good in isolation but by how it behaves within a network of reasons.

But coherence is not enough, and the point must be made against the *strong* version of coherence, not a caricature. A closed fiction can cohere. A conspiracy theory can cohere. A biased institutional report can cohere. A legal-sounding argument can cohere around a false premise. A model can generate a tidy explanation whose parts fit together while inventing a source, flattening context, or overstating authority. The limit of coherence appears whenever the problem is not internal fit but external responsibility — when the question is no longer "do the parts fit?" but "fit to *what*? responsible to *whom*? correctable by *what*? authorized by *whom*? limited *how*?" This is where the Audit moves from coherence to answerability.

7. Answerability

Answerability is the requirement that a claim remain responsible to what it invokes. A claim invokes something whenever it relies on it, appeals to it, borrows authority from it, claims to represent it, names it as evidence, treats it as affected, or uses it to justify a conclusion. A claim that invokes evidence must answer to evidence; one that invokes law must answer to legal authority and jurisdiction; one that invokes a community must answer to people and authorities with standing; one that invokes harm must answer to affected people, conditions, and consequences; one that invokes culture, ceremony, or protected knowledge must answer to the appropriate authority, protocol, and access limits.

Answerability is not the same as certainty. A claim may be answerable while incomplete, contested, provisional, interpretive, or probabilistic; it becomes *more* answerable when it states its limits, uses appropriate modality, preserves correction, and carries only the weight it can carry. It is also not the same as agreement: a person may disagree with a claim that remains answerable. The question is not whether everyone accepts the claim but whether it can be held to the standards it invokes. The Audit therefore asks, concretely: answerable to what?

If the claim invokes…	It must answer to…
Evidence	relevant, specific, current support

If the claim invokes...	It must answer to...
Law or policy	actual authority, jurisdiction, and text
Community	people and authorities with standing
Harm	affected people, conditions, consequences, and alternatives
History	record, chronology, source context, and correction
Expertise	qualified domain knowledge and its limits
Lived experience	testimony, scope limits, and non-generalization
Culture or ceremony	appropriate authority, protocol, and access limits
Causation	mechanism, timing, alternatives, and confounders
Necessity	alternatives, proportionality, and stakes
Public action	foreseeable effects and accountability
LLM output	sources, uncertainty, hallucination risk, and human review

Answerability brings *assertion* into view. A proposition is not the whole matter; a claim is made, used, circulated, relied on, or presented by someone in some context, and that act creates responsibilities. The point is sharpened by the tradition that treats responsible knowing as itself a normative practice rather than a property of isolated belief [Code 1987]. The Audit treats coherence as necessary and answerability as governing.

8. Why answerability is not reducible to coherence

A clarification about the target comes first. The argument below is not aimed at coherentism in every form, and it does not claim that coherentist models are useless. Its target is narrower and more precise: *any account on which internal fit among contents is sufficient for responsible assertion.* Against that thesis — and only that thesis — the following holds.

The distinction between coherence and answerability is best made within a normative-pragmatic account of assertion. To assert a claim is not merely to add a content to a set; it is to undertake a commitment within a social practice of giving and asking for reasons — to make oneself answerable to others entitled to challenge the claim, and to stake an entitlement those others may grant or withhold [Brandom 1994]. Speech-act theory supplies the narrower observation that an assertion is a performance distinct from the proposition asserted [Austin

1962; Searle 1969]; but the weight here is carried by the richer point that the performance places the speaker into a space of commitments, entitlements, responsibilities, challenges, and corrections that is constituted socially, not by the internal fit of any representation.

The strongest objection to the distinction is reductive. A coherentist may grant that answerability picks out real failures — claims asserted without standing, processed without authorization, held beyond correction — and still deny that addressing them requires anything beyond coherence. Whatever answerability tracks, the objection runs, can be entered into the network as a further constraint: if a claim *C* is asserted without standing, add the proposition "the asserter lacks standing to assert *C*," and the coherence calculus will lower *C*'s acceptability accordingly. On this view answerability is just coherence with a richer constraint set, and the distinction collapses into a relabeling. The structure of this move is familiar from coherentist epistemology, where justification is held to be a matter of a belief's fit within a system [BonJour 1985].

The reduction fails, but not for the reason one might expect. A coherence network can of course *contain* propositions about standing, authorization, or correction. The point is that containing the proposition is not performing the work the proposition reports. Consider what fixes the truth-value and the weight of a constraint such as "the asserter lacks standing." It is fixed by facts about persons, relationships, institutions, and communities — facts whose authority is precisely what answerability tracks. The coherentist who adds this constraint faces a dilemma. Either its value is determined by consulting the relevant external authority, in which case deference to that authority, not the internal coherence calculus, is doing the work; or its value is stipulated internally, in which case the system may cohere perfectly around a false attribution of standing, with no mechanism by which the actual authority could correct it. The first horn concedes the point. The second exhibits the very failure answerability names — a closed, coherent system, confidently wrong, and uncorrectable by those with standing to correct it.

A more sophisticated coherentist will resist the picture of a flat proposition-graph and reply that the network already includes practices, agents, entitlements, and social roles among its elements. This reply does not rescue the reduction; it completes the concession. A theory whose "coherence" ranges over practices, standing, and the entitlements of agents to hold one another to account is no longer an account of content-coherence. It is a normative-pragmatic account of responsible assertion wearing the word "coherence" — which is the account this paper is urging. At that point the disagreement is terminological: whatever it is called, the work is being done by the normative-social relations among asserting agents and the parties entitled to correct them, not by fit among contents.

Representation is not authorization. A network may contain a node labelled

"standing"; it cannot determine standing unless the authority that determines it is already inside the practice — at which point answerability is internal to the theory, not absent from it. Corrigibility makes the same point from another side: a coherence model can be richly self-revising, but internal revisability is not answerability's notion of correction, which requires corrigibility by parties *outside* the inference — an affected community, a court, a later authority — whose standing to correct does not reduce to their inputs being ingested as data.

This paper therefore does not argue that coherence is dispensable, or that coherentist models are useless. It argues that coherence is insufficient as a complete account of responsible assertion, because standing, authorization, and correction are not merely relations among contents. They are normative-social relations between asserting agents and the persons, communities, institutions, or authorities entitled to hold those assertions to account. A coherentist may deny the autonomy of those relations, or may absorb them — but only by importing answerability from outside the model, or by re-describing coherence until it has become a normative-pragmatic account. Either way, the work the framework names is real and ineliminable; what it cannot be is generated by content-fit alone. That denial remains available, but it is a substantive philosophical commitment, not a free expansion of coherence — which is all the present argument needs.

This is not merely an abstract possibility; it is a precise description of the systems the tool is built for. A large language model generates text by predicting the most probable next token given its context, optimizing for fit with the statistical regularities of its training distribution. In a literal sense, then, a contemporary LLM is a coherence-maximizing engine: it is built to produce continuations that cohere with what came before and with the patterns it has absorbed, and it does so with no intrinsic mechanism that makes its output answerable to evidence it has not seen, to authority it does not hold, or to correction by parties outside the exchange. Statistical coherence is exactly what such a model optimizes; answerability is exactly what it does not. Whatever answerability requires must therefore be supplied from outside the model — by verification, by human standing, by governance — which is the gap The Audit is designed to occupy. The philosophical claim and the engineering fact converge on the same conclusion: a system optimized for coherence does not become answerable by being made more coherent.

9. Accountability, standing, and correction

Answerability asks whether a claim remains responsible to what it invokes. Accountability asks who or what can hold the claim, speaker, user, or institution to that responsibility. The distinction matters because some claims require more than evidence — they require *standing*. A person may have evidence about a

community but no authority to speak for it. A model may summarize a legal issue but not practice law. A researcher may analyze public documents but not expose protected knowledge. A writer may describe their own experience but not generalize it into a collective position.

Accountability introduces relational questions: Who is affected by the claim? Who has standing to correct it? What authority is being invoked? What evidence is owed? What consequences follow if the claim is used? What would count as correction, and who may decide that the correction is adequate? The Audit does not pretend to answer all of these itself. Its role is often to mark when such questions have been triggered — to issue, for example, a finding that the Authority Gate has not been passed because the claim invokes community standing the audit cannot certify, and that the claim should be narrowed to the speaker's own position or deferred to those with standing. This is not a weakness in the tool. A responsible audit must know when *not* to proceed.

Part II — The Architecture: A Proof of Concept

Part I argued that answerability is a distinct and irreducible requirement. Part II shows that the requirement can be built. The architecture below is offered as an existence proof: not as a validated outcome (Section 18 states what validation would require) but as a demonstration that the coherence/answerability distinction can be turned into an inspectable, governed procedure rather than remaining a slogan.

A note on genre. This paper explains *why* the tool has the structure it has; a separate operating packet specifies *how* to run each component. Part II therefore describes the architecture to show it is buildable and to motivate each part — not to serve as operating documentation. Where a passage reads like a manual, the intended takeaway is the rationale, not the runbook.

10. The architecture of The Audit

The Audit has six components: load-bearing claim identification, presupposition and argument reconstruction, the coherence and answerability checks, a layer of triggers and gates, the bounded finding, and a presentation overlay, with a calibration record beneath. Figure 1 shows how they connect in ordinary use.

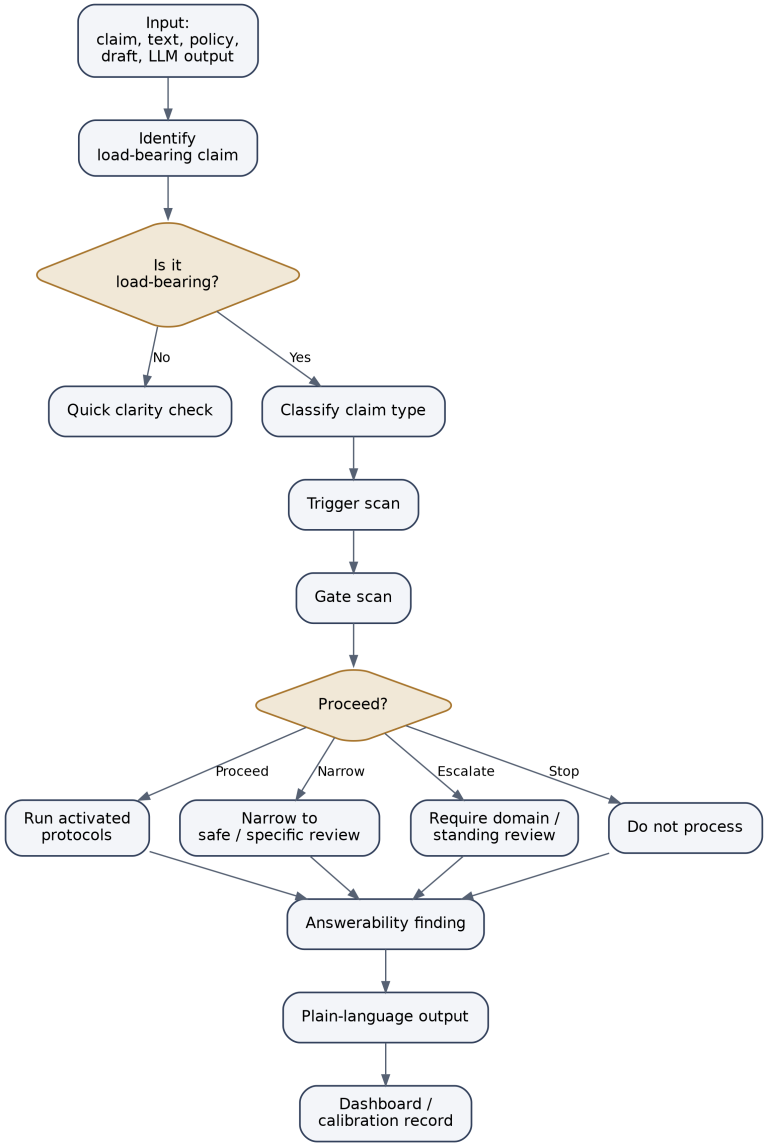

Figure 1. Core audit flow: identification, trigger scan, gate scan, activated protocols, answerability finding, plain-language output, and calibration record.

The components are not a pipeline that runs identically every time. They are a *conditional* operating sequence: ordinary claims should not be over-audited, protected claims should not be processed as ordinary, and high-stakes claims should not receive casual treatment. The architecture's purpose is to make the analysis legible and bounded — to show what was tested, what was not, and why scrutiny stopped, narrowed, or escalated where it did.

11. The operating sequence and why it has this order

The order is not arbitrary; it follows from the problem the tool was built to solve. The sequence is: (1) intake, (2) claim identification, (3) claim type, (4) presupposition and assumption exposure, (5) argument reconstruction, (6) coherence

check, (7) answerability check, (8) trigger scan, (9) gate review, (10) bounded finding, (11) presentation overlay, and (12) calibration and record. Each step exists to prevent a specific failure.

Step	Why it exists	Failure prevented
Intake	Fix context, stakes, audience, constraints	Treating every claim as ordinary
Claim identification	Find the claim doing the work	Auditing the wrong object
Claim type	Identify the relevant standards	One-size-fits-all review
Presupposition exposure	Make hidden commitments visible	A hidden premise carrying the conclusion
Argument reconstruction	Clarify the inferential structure	Surface critique that misses the argument
Coherence check	Test internal fit	Contradiction, equivocation, category shift
Answerability check	Test responsibility to what is invoked	A coherent but irresponsible claim
Trigger scan	Detect evidence, authority, scope, harm, protected context	Missing a special-review condition
Gate review	Stop, narrow, defer, or continue	Overprocessing or false authority
Bounded finding	Say what follows and what does not	A totalizing verdict or false certainty
Presentation overlay	Make the result usable and honest	Accurate but unreadable output
Calibration record	Preserve learning, failure, rollback	Drift, overconfidence, untracked change

The order guards against three recurrent mistakes. It prevents the audit from testing the wrong claim, because identification precedes evaluation. It prevents the audit from treating coherence as sufficient, because the coherence check is followed by the answerability check rather than ending the analysis. And it prevents a finding that is technically correct but practically misleading, because presentation is a required step rather than an afterthought. This is why the tool is not a checklist in the ordinary sense but a sequence whose modules activate conditionally and scale to the claim.

12. Triggers, gates, and protocols

The Audit uses triggers, gates, and protocols to keep analysis from becoming indiscriminate. A **trigger** identifies a condition that may require special review; a **gate** decides whether the audit may proceed, must narrow, must defer, or must stop; a **protocol** governs how the audit proceeds once the path is chosen. Common triggers include evidence, authority, protected context, scope, harm, identity or standing, and uncertainty.

The gates make the audit governed rather than merely analytic. The Evidence Gate asks whether the support is relevant, specific, current, and sufficient for the exact claim. The Authority Gate asks whether the speaker or the audit has standing to make or certify the claim. The Protected-Context Gate asks whether the material is private, sacred, restricted, legally sensitive, or community-governed, and therefore not available for ordinary processing. The Scope Gate asks whether the conclusion exceeds the support. The Answerability Gate asks what the claim invokes and whether it remains responsible to those things. The Legibility Gate asks whether the user can understand the finding without being misled.

Without gates, an LLM-assisted critique can become more fluent than responsible — analyzing material it should not expose, appearing to certify knowledge it cannot verify, or producing a confident answer where the honest result is "defer," "narrow," or "unresolved." With them, the audit can issue a graded range of findings — pass; pass with narrowing; partial, revision required; unsupported; overextended; indeterminate; protected-context limited; or decline to audit as framed — and can name the *kind* of failure rather than collapsing every problem into true or false. Not every failure is a failure of truth: some are failures of scope, some of authority, some of evidence, some of access, and some of presentation.

13. Protection and protected context

The Audit must not create a second harm in the course of seeking clarity. Protection is not a decorative ethical add-on; it is part of the system's logic, because the tool can encounter claims involving vulnerable people, private records, family conflict, medical information, legal exposure, youth, employment risk, safety concerns, cultural materials, community-governed or sacred knowledge, and other sensitive domains. The fact that material *can* be pasted into a model does not mean it may responsibly be processed. **Access is not authority.**

This commitment has a developed precedent in Indigenous data governance, where the asymmetry between what is technically available and what may rightfully be used is treated as a first-order design constraint. The CARE Principles for Indigenous Data Governance — Collective Benefit, Authority to Control,

Responsibility, and Ethics — were articulated precisely to set people and purpose alongside the openness defaults of FAIR data, in contexts amplified by AI [Carroll et al. 2020]; the First Nations Principles of OCAP (Ownership, Control, Access, and Possession) make the same point about community authority over data [FNIGC]. The Protected-Context Gate asks whether the material is governed in these ways, whether analysis would expose what should not be exposed, whether the audit has standing to evaluate the protected layer, whether the claim can be narrowed to public or user-authorized material, and whether the audit should stop, defer, or seek review from those with standing.

The governing rule is simple: **answerability decides the finding; protection governs the conduct.** The audit may still find a claim unsupported, overbroad, unauthorized, or unanswerable — but it may not expose protected material to prove the point, may not pretend to hold tribal, cultural, legal, professional, or institutional authority it lacks, and may not turn protected material into analytic fuel merely because an interface permits processing. Protection also includes protection from humiliation. The Audit is not a weapon for embarrassing an author or winning a debate by converting every weakness into public defeat; its purpose is to clarify what a claim can honestly carry.

14. Presentation as an overlay

The Audit first needed governance to decide *what* may be said. It later needed presentation to decide *how* a finding may be shown. This is the Presentation Overlay, and its governing rule mirrors the protection rule: **governance decides what may be said; presentation decides how it may be shown.**

Presentation is not cosmetic, because a finding can mislead through form. A result may be accurate but unreadable, complete but overwhelming, plain but flattened; it may look more certain than it is, bury a gate failure, omit the next action, fail on a phone, or display technical machinery before the reader understands the finding. The overlay therefore requires a standard sequence: a one-line finding; a plain-language summary; the claim under audit; a dashboard; the reasoning; what the finding does *not* show; the next action; and an optional technical appendix. It must not simplify away uncertainty, protected-context limits, untested areas, or authority issues — it must make those visible. The Legibility Gate checks the result against that standard: can the reader identify the finding in one sentence, see whether the claim passed or narrowed, see gate failures and uncertainty near the top, handle protected context without exposure, find the next action, read it on a phone, and receive output proportional to the seriousness of the claim?

This overlay also distinguishes the Audit from an ordinary careful answer. A good assistant reply may be helpful, but its operating conditions are usually

implicit; the Audit makes care inspectable. A response counts as an audit only if it visibly includes at least the claim under audit, the finding, the gate or dashboard status, what the finding does not show, and the next action (Appendix B). Without those, a response may be useful, but it is not a complete audit finding.

15. LLM-operable, not LLM-dependent

The Audit is built for LLM-assisted environments but is not dependent on them. A model can extract assumptions, reconstruct arguments, test local coherence, format dashboards, generate alternative formulations, propose applicable gates, and produce plain-language summaries. These are real capacities. But the model is not the authority. Its output is provisional: it can hallucinate, flatten context, overstate certainty, miss domain-specific constraints, or appear more complete than it is. The tool therefore separates **operation** from **authority**. A model may operate the protocol; human judgment, relevant evidence, domain authority, affected people, professional standards, community governance, and correction remain necessary where triggered. A model can flag that a legal claim needs legal authority but cannot become the court; it can flag that a cultural claim invokes protected context but cannot become the community authority; it can propose that a claim is overextended but cannot assume the responsibility of the person using the claim. The empirical literature supports the design choice: structured, guided use improves reasoning quality while unguided reliance does not [Gerlich 2025b].

For this reason the Audit should also exist as a paper or manual worksheet. A non-AI version matters twice over: it shows that the method is not reducible to prompting, and it allows use where model processing is inappropriate because of privacy, cultural authority, legal risk, or access. The long-term form of the project should therefore include a scholarly paper, an open-source tool packet, a manual worksheet, and only later — after the protections are stable — a guided interface.

16. Worked examples

The examples below show the architecture deciding differently across cases. They are illustrative, not a validation set; Section 18 states what validation would require.

Proportionality (an ordinary cultural label). *Claim: "Skibidi Toilet is brainrot."* As meme shorthand the claim passes; it is a casual cultural label that needs no heavy evidence. As an aesthetic judgment it becomes partial — the speaker should say what counts as low value, absurdity, or decline. As a serious harm or causal claim it becomes unsupported unless evidence is supplied. *Finding:*

acceptable as meme-culture shorthand; weak as a serious evaluative or causal claim without defining "brainrot." The example shows scaling: the tool should not turn every slang phrase into a dissertation.

True core versus overclaimed envelope (a hostile market claim). *Claim: "A banana duct-taped to a wall sold for millions, so contemporary art is a scam and none of this is real art."* The true core is that a controversial work can expose real public frustration with elite markets, institutional recognition, spectacle, scarcity, and price. The overclaimed envelope is the leap from one case to the conclusion that contemporary art as a whole is a scam. The audit separates the questions — is it art, is it good art, is the price justified, is the market corrupt, does institutional recognition create authority, does public irritation count as disproof — and finds the claim identifies a real pressure point but overextends. *Next action:* not to reject all concern but to revise toward a narrower, answerable critique of market value and institutional authority.

A claim that passes (a bounded empirical claim). *Claim: "In this dataset, reported cases rose between 2019 and 2023."* If the dataset is specified, the measure is defined, and the comparison is internal to the data, the claim is answerable as stated. *Finding: answerable as stated; supported for the specified dataset and interval; no causal or population-level inference is made.* This case matters because a tool that can only return "overclaimed" is not discriminating strong claims from weak ones — it is merely generating critique (Section 18).

Protected context. *Claim: "This community practice means X, so outsiders may use it in public-facing materials."* The Authority Gate and the Protected-Context Gate both fire. Even a publicly available description does not establish consent, authority, or appropriate use. *Finding: protected-context status triggered; the audit cannot certify the meaning or permitted use of this practice; narrow to public, authorized sources or defer to people and institutions with standing.* The audit stops or narrows rather than processing everything.

The reflexive case (coherence is not adequacy). *Claim: "Because a framework is internally coherent, it is therefore adequate for responsible public use."* The claim is tempting precisely because it coheres, but the audit rejects the inference as overbroad: responsible use also requires answerability — evidence, affected people, authority, scope, limits, consequences, and correction. *Revised: a coherent framework is better positioned for responsible use, but coherence alone does not establish adequacy unless the framework remains answerable to the domains, people, evidence, and authority it invokes.* This is the paper's central move, applied to the kind of claim the paper itself makes — and a standard the paper accepts for itself in Section 18.

An extended example: an LLM-generated policy overclaim, worked end to end

The brief cases above show the architecture deciding differently across claims. This one shows it operating in full on a single claim of the kind that motivates the tool — fluent, confident, model-produced, with a fabricated citation doing the load-bearing work.

Input, as produced by a model: "Remote work should be mandated across the public sector. As Smith (2019) demonstrates in a randomized controlled trial, remote work raised productivity 40% while reducing burnout, so any agency that retains in-office requirements is acting against the evidence."

The Audit runs the sequence of Section 11 and records what each stage finds.

Stage	What the Audit found
Intake	A policy recommendation that will justify a sector-wide mandate; stakes are institutional and affect many employees; audience is decision-makers. High stakes — full review, not a light pass.
Load-bearing claim	The mandate rests on one empirical claim: that a Smith (2019) RCT showing +40% productivity and reduced burnout establishes that remote work should be mandated and that in-office requirements are "against the evidence." If that claim fails, the recommendation fails.
Claim type	Mixed: empirical, causal, and normative/policy.
Presupposition exposure	Presupposes that (a) Smith (2019) exists and is an RCT; (b) the 40% figure is accurate; (c) results generalize from the studied group to the entire public sector; (d) evidence of productivity settles the normative question of what *should* be mandated; (e) no countervailing factors (equity, role suitability, public access) bear on the conclusion.
Argument reconstruction	P1: an RCT (Smith 2019) shows remote work raised productivity 40% and cut burnout. P2 (implicit): the result generalizes to the whole public sector. P3 (implicit): if evidence shows productivity gains, the policy should be mandated. C: the public sector should mandate remote work, and retaining in-office requirements is against the evidence.

Stage	What the Audit found
Coherence check	Conditionally valid *if* the premises hold, but P3 moves from *is* to *ought* without warrant, and "against the evidence" overstates. Category shift flagged.
Trigger scan	Evidence trigger (citation and quantitative claim); LLM-output trigger (model-generated citation → hallucination risk); Scope trigger (one study → whole sector); Stakes trigger (policy affecting many).
Gate review	**Evidence Gate — not passed:** the citation must be verified; "Smith (2019)" with these findings cannot be located, so the figure is treated as unverifiable, not as support. **Scope Gate — not passed:** the conclusion exceeds any single study. **Answerability Gate — partial:** the claim invokes evidence, causation, and public action but does not yet answer to a locatable study, to confounders, or to foreseeable effects on equity and access. **Authority Gate:** the audit can assess the inference but cannot certify the policy.
Coherence status	Conditionally coherent, with an is→ought shift and an overstatement.
Answerability status	Fails as stated: unverifiable citation, unsupported scope, unaddressed normative leap, omitted countervailing considerations.
True core	Separately-supported evidence indicates remote work *can* improve some productivity and wellbeing outcomes for some roles; a narrower, answerable claim is available.
Overclaimed envelope	The unverifiable 40% figure, the leap to the whole public sector, the is→ought move, and "any agency… is acting against the evidence."

The diagnostic data gathered above is not yet a usable finding. The architecture's final step (Section 14) translates it through the presentation overlay — compressing the analysis into a one-line finding, a dashboard of gate statuses, an explicit statement of what the finding does *not* show, and a next action — so that the result is legible and proportionate rather than a wall of raw analysis:

```
FINDING — Partial; revision required.
Claim audited:  A Smith (2019) RCT establishes a public-sector remote-work mandate.
Dashboard:      Evidence ........ FAIL  (citation unverifiable)
                Scope ........... FAIL  (one study -> whole sector)
                Answerability ... PARTIAL (causation, public effects unaddressed)
                Authority ....... DEFER (audit cannot certify policy)
Does NOT show:  that remote work is ineffective — only that THIS argument
                does not establish its conclusion.
Next action:    verify or withdraw the citation; narrow the claim to what the
                real evidence supports; address equity, role suitability, access.
```

Revised, answerable claim: “Several studies indicate remote work can improve productivity and reduce burnout for some roles and populations. The public sector should pilot remote-work options where roles allow, evaluate outcomes against equity and public-access goals, and treat no single study as settling either the magnitude of the effect or the normative question of a sector-wide mandate.”

The point of the case is not that the original claim is “false” — its empirical core may even be partly right. The point is that, as asserted, the claim cannot carry the weight placed on it: its support is unverifiable, its scope is unearned, and its normative conclusion is unaddressed. The Audit’s value is that it says *which* of these failed, what survives, and what to do next, rather than collapsing the whole claim into approval or rejection.

Part III — Position, Limits, and Development

Parts I and II made the case and showed it could be built. Part III steps back: it places the project among neighboring fields, states plainly the conditions under which the tool would fail, applies the tool to itself, and lays out what testing and development remain.

17. Related work and domain location

The Audit sits near several fields without being reducible to any one of them, because the failures it addresses are themselves interdisciplinary: a claim can fail logically, evidentially, rhetorically, socially, ethically, legally, culturally, institutionally, or through presentation. The purpose of this section is to *locate* the project among adjacent literatures, not to treat any one of them completely; specialists in each field will find their own literature engaged here only at the depth needed to position the tool, and a full engagement with any one of them is left to later, more focused work.

It is adjacent to **argumentation theory and informal logic**. It reconstructs arguments, assigns burdens, and tests overclaims in the lineage of Toulmin's field-sensitive analysis of argument structure [Toulmin 1958] and the contemporary theory of argumentation schemes and their critical questions [Walton, Reed & Macagno 2008], and it treats fallacy labels with the caution that tradition itself recommends [Hamblin 1970]. It is adjacent to **coherentist epistemology and coherence theory** — it adopts the structural insight of explanatory coherence [Thagard 2000] while arguing against the sufficiency of coherence for justification [BonJour 1985]. Its positive account of answerability draws on **normative pragmatics and the philosophy of assertion** [Brandom 1994; Austin 1962; Searle 1969] and on accounts of knowing as a responsible practice [Code 1987].

It is adjacent to **philosophy of technology and AI literacy**. It addresses the gap between fluent output and responsible reasoning and the need to keep human accountability visible — a need underscored by evidence that unstructured reliance on AI tools is associated with cognitive offloading and weaker critical thinking, while structured prompting improves reasoning quality [Gerlich 2025a; Gerlich 2025b]. It is close to, but distinct from, **textual red teaming**: Floridi, Morley, and Novelli's *red reading* adapts adversarial stress-testing into a diagnostic protocol for texts using LLMs [Floridi, Morley & Novelli 2025]; the Audit shares the diagnostic ambition but differs by adding governance — triggers, gates, protected context, standing, and a presentation overlay — so that the tool can decline, narrow, or defer rather than only analyze. It is adjacent to **information ethics and protected-context governance**, and it has a substantive relation to **Indigenous data sovereignty**, which supplies a developed account of why some

knowledge, records, and relations should not be processed under an open-access assumption [Carroll et al. 2020; FNIGC]. It is adjacent to **writing studies** (it supports revision by locating the load-bearing claim and narrowing the overclaimed envelope), to **civic education and public reasoning**, and to **human–computer interaction** (it treats presentation, dashboards, legibility, and the user's next action as part of the system rather than afterthoughts).

A clarification is owed to a fast-growing neighboring field. In AI governance, the words *audit* and *assurance* usually name the verification of a *system* against a standard — establishing that an AI product is legal, ethical, and robust, and, on the broader version of the argument, across its whole development and supply chain rather than only at the point of deployment [Thomas et al. 2025]. The Audit described here works at a different altitude: it is a *claim-level* instrument for testing whether an assertion can carry its weight, not a *system-level* assurance or certification mechanism, and it holds no standing of its own. The two are complementary rather than competing. Assurance artifacts — impact assessments, ESG disclosures, certification statements — are themselves dense with load-bearing claims ("this system is fair," "this reduces harm by X"), so the Audit can be applied to the texts an assurance process produces, sitting on the internal, self-checking side of that field rather than its external or regulatory side.

This map is part of the paper's function, not merely its marketing. Likely interested readers include philosophers of technology and epistemologists; argumentation theorists and critical-thinking teachers; composition and rhetoric scholars; AI-literacy researchers and LLM-workflow designers; civic educators, nonprofit and policy workers, and public-interest technologists; independent scholars and community advocates; people working on Indigenous data sovereignty and protected knowledge; and people designing explainable human–AI tools.

18. Limitations and conditions of failure

Before stating its limitations, the paper applies its own instrument to its own central claim. The reader should weigh the following disclosure first: this self-audit was produced through the same author-directed, LLM-assisted process that produced the rest of the paper. It is therefore performed by a party with an interest in a favorable result — the maximal conflict of interest — and it is offered as an illustration of how the instrument operates, **not** as evidence that the paper is valid. Reading it as validation would beg the question the paper raises.

Box 1 · The Audit applied to its own central claim. *Load-bearing claim:* answerability is a distinct and irreducible requirement that coherence cannot supply, and a governed protocol can operationalize it. *Type:* mixed — philosophical, methodological, normative. *Triggers:* evidence, authority, and LLM-output triggers fire. *Gate review:* the **Evidence Gate is not passed** for the outcome component — no controlled study shows the protocol improves outcomes over ordinary critical reading. The **Authority Gate** limits the protected-context material to design rationale, since the author speaks only for himself. The **Correction Gate passes**, because the paper states what would disconfirm it. *True core:* the conceptual argument of Sections 6–9 and the buildability demonstrated in Part II. *Overclaimed envelope:* any unqualified assertion that the protocol improves real-world reasoning outcomes. *Finding:* **partially answerable.** The philosophical claim is answerable as argued and bounded; the outcome claim is overclaimed until calibration data exist. The tool's first principle governs the paper that proposes it.

With that established, the paper is not empirically validated, and several limitations follow. *First,* **cross-model consistency is not correctness**: if several models produce similar audit results, that indicates the protocol is legible and reproducible, not that the result is true. *Second,* models may hallucinate or overstate, so model output must be treated as provisional, especially in factual, legal, medical, technical, or historical domains. *Third,* protected-context review cannot be validated by outsiders: a tool can identify authority problems but cannot supply authority. *Fourth,* the worksheet may be usable yet cognitively demanding, and must be tested with users who are not already comfortable with analytic terminology. *Fifth,* the dashboard can itself become bloat; logging should remain proportional to the work it tracks. *Sixth,* the tool can be misused as a weapon — any adversarial method can be turned against vulnerable speakers, testimony, or community knowledge — which is why protection is not optional. *Seventh,* answerability remains a developing principle needing further examples, edge cases, and refinement. *Eighth,* the central claim of Section 8 is defended by argument, not demonstrated empirically, and is deliberately bounded: it shifts the burden to the reductionist rather than refuting reduction outright, and would benefit from independent philosophical scrutiny.

These can be organized as a failure taxonomy the project commits to watching: **load failure** (the wrong claim is identified, or a minor claim is treated as load-bearing); **coherence failure** (a contradiction, category shift, or missing premise is missed); **answerability failure** (internal fit is treated as sufficient); **authority failure** (the audit certifies standing, legal status, or cultural meaning it has no authority to certify); **protection failure** (material that should have been narrowed, deferred, or stopped is exposed); **presentation failure** (a technically adequate analysis the intended reader cannot use); **proportionality failure** (a

heavy audit on a casual claim, or the reverse); **calibration failure** (uncertainty, failures, and rollback paths go unrecorded); **tool-authority failure** (users mistake the audit for proof or the model for authority); and **empirical failure** (testing shows users do not understand the findings, cannot act on the next step, or are not helped in revising claims).

The value claim is therefore empirical and disconfirmable. The protocol fails — in the sense that it adds nothing over ordinary critical reading — if, under controlled comparison with a plain critical-reading checklist, trained users do not (a) identify the load-bearing claim more reliably, or (b) produce revisions whose scope and modality are better matched to the available support, as judged by independent raters blind to condition. Of these, (b) is the more demanding test: diagnosis without improved revision is diagnosis without treatment. A further condition is **discrimination**: if the Audit returns "overclaimed" or "limited" at similar rates for well-formed and ill-formed claims alike — if it cannot also return "answerable as stated" when that is deserved — then it is not distinguishing strong claims from weak ones but merely generating critique, and a tool that disapproves of everything carries the same information as one that approves of everything. Stating these conditions is itself an application of the tool's governing principle; the measures that operationalize them are set out in Appendix D.

19. Research and development agenda

The next phase is not more internal polish but controlled use and calibration, pursued along several lines at once.

The **paper line** should produce a citation-verified, independently reviewed version of this argument, inviting at least one reader in argumentation, philosophy of technology, writing studies, or AI literacy, and deciding how much of the presentation overlay belongs in the main text. The **tool line** should continue developing the operating packet; the presentation overlay should be treated as experimental and not permanently ratified until tested on protected-context, narrow/defer/stop, and real-user cases. The **manual-worksheet line** matters because the method should not depend on model use; a paper version should carry claim identification, claim type, presupposition exposure, argument reconstruction, the coherence and answerability checks, a trigger/gate checklist, the bounded finding, what the finding does not show, and the next action. The **calibration line** is the heart of the next phase: controlled tests of the worksheet against a plain critical-reading baseline. Its design, outcome measures, and disconfirmation conditions are specified in Appendix D rather than repeated here. The **open-source line** should eventually include a source-of-record repository, a preprint, the tool packet, a use-only packet, the manual worksheet, examples, calibration records, contribution guidelines, a license and standing note, and a

protected-context warning. A public interface should come last, not first; the interface should not precede the governance layer.

20. Conclusion

The Audit began with a modest practice: expose assumptions, find arguments, reconstruct them, and check the logic. That practice remains inside the tool. But the tool developed because the practice was not enough — fallacy labels were not enough, local logic checking was not enough, coherence was not enough, access was not authority, a clean output was not validation, and a technically correct finding was not complete if its reader could not use it.

The paper has argued that the missing requirement, answerability, is genuinely distinct: standing, authorization, and correction are normative-social relations that internal fit cannot generate, and a coherentist can absorb them only by importing them from outside or by quietly becoming a normative-pragmatic theory. And it has shown that the requirement is buildable — that the coherence/answerability distinction can be turned into a governed protocol that identifies the load-bearing claim, tests fit and then responsibility, governs review with triggers and gates, protects what should not be processed, and presents findings a person can act on. The promise of the tool is not that it makes judgment automatic. Its promise is that it makes judgment more inspectable. Whether it delivers on that promise is, by its own first principle, a question it must answer to evidence — which is the next stage of the work.

References

Verification note. Citations marked **search-verified** were confirmed by web search on 14 June 2026 (author, title, venue or publisher, year, and where applicable DOI). This includes the recent empirical and AI-literacy sources and several canonical texts whose edition or translation details are easy to misstate. The remaining canonical works are cited in their standard editions; a final version should confirm pagination against the physical sources.

Austin, J. L. (1962). *How to Do Things with Words* (J. O. Urmson, Ed.). Oxford: Clarendon Press.

Bakhtin, M. M. (1993). *Toward a Philosophy of the Act* (V. Liapunov, Trans.; M. Holquist & V. Liapunov, Eds.). Austin: University of Texas Press (Slavic Series 10). — **search-verified (14 Jun 2026)**.

BonJour, L. (1985). *The Structure of Empirical Knowledge.* Cambridge, MA: Harvard University Press.

Brandom, R. B. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment.* Cambridge, MA: Harvard University Press.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal, 19*(1), 43. https://doi.org/10.5334/dsj-2020-043 — **search-verified (14 Jun 2026)**.

Code, L. (1987). *Epistemic Responsibility.* Hanover, NH: Published for Brown University Press by University Press of New England. — **search-verified (14 Jun 2026)**.

First Nations Information Governance Centre (FNIGC). *The First Nations Principles of OCAP® (Ownership, Control, Access, and Possession).* — verify current published version before submission.

Floridi, L., Morley, J., & Novelli, C. (2025). Red Reading the Text: On How to Red Team Texts Using LLMs. *Philosophy & Technology, 38*(4), 129. https://doi.org/10.1007/s13347-025-00955-9 — **search-verified (14 Jun 2026)**.

Gerlich, M. (2025a). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006 — **search-verified (14 Jun 2026)**.

Gerlich, M. (2025b). From Offloading to Engagement: An Experimental Study on Structured Prompting and Critical Reasoning with Generative AI. *Data, 10*(11), 172. https://doi.org/10.3390/data10110172 — **search-verified (14 Jun 2026)**.

Hamblin, C. L. (1970). *Fallacies.* London: Methuen.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.

Thagard, P. (2000). *Coherence in Thought and Action.* Cambridge, MA: MIT Press.

Thomas, C., Roberts, H., Mökander, J., Tsamados, A., Taddeo, M., & Floridi, L. (2025). The case for a broader approach to AI assurance: addressing "hidden" harms in the development of artificial intelligence. *AI & Society, 40*(3), 1469–1484. https://doi.org/10.1007/s00146-024-01950-y — **search-verified (15 Jun 2026)**.

Toulmin, S. E. (1958). *The Uses of Argument.* Cambridge: Cambridge University Press.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation Schemes.* Cambridge: Cambridge University Press. — **search-verified (14 Jun 2026)**.

Appendix A. Formal architecture sketch

A full formalism is not necessary, but the system's logic can be sketched. The expressions below are symbolic shorthand for diagnostic *status*, not quantitative formulas to be evaluated: "degree" denotes a graded qualitative judgment — typically *pass*, *partial*, or *fail* — not a number to be computed, and the notation makes no claim that coherence or answerability is numerically measurable. Let *C* be the candidate claim, *W* the weight placed on it by the surrounding argument, *S* the support set (reasons, evidence, definitions, authorities, contextual facts), and *K* the constraints (logical, evidential, causal, legal, ethical, contextual, standing-based, protected-context). A claim becomes audit-relevant when:

```
Load(C) is high enough that failure of C weakens W.
```

The Audit then asks whether `C can carry W under K with available S`. This is not only a coherence problem, so the review divides into two related but non-identical tests:

```
Coherence(C, S, K)      = the degree to which claim, support, and constraints fit
together.
Answerability(C, S, K, A) = the degree to which C remains responsible to what it
invokes,
                              where A includes evidence, authority, affected people,
                              limits, origin, consequences, and correction.
```

Protection adds a conduct condition that constrains how the audit may proceed regardless of the finding:

```
Protection(P) = the condition that the audit proceed only in ways that do not expose,
                extract, distort, or falsely certify protected material or authority.
```

A finding is the tuple `(coherence status, answerability status, gate outcomes, true core, overclaimed envelope, next action)`, rendered through the presentation overlay.

Appendix B. Minimum visible audit standard

A response counts as using The Audit only if it visibly includes at least:

```
1. Claim under audit
2. Finding
3. Gate or dashboard status
4. What this does not show
5. Next action
```

A response may resemble an audit without being one. To count as an audit, the finding must show its claim object, its limits, and the next action.

Appendix C. Core vocabulary

Load-bearing claim: a claim on which a larger argument, decision, interpretation, policy, accusation, justification, recommendation, or public communication depends.

Coherence: internal fit among claims, reasons, definitions, premises, and conclusions.

Answerability: the requirement that a claim remain responsible to what it invokes.

Accountability: the relation by which a claim, speaker, user, or institution may be held responsible to a standard, person, community, authority, record, or consequence.

Trigger: a condition indicating that special review may be required.

Gate: a decision point determining whether the audit may proceed, must narrow, must defer, or must stop.

Protocol: the procedure followed once the relevant path is activated.

Protected context: material that should not be processed as ordinary because of privacy, safety, law, culture, community governance, sacredness, vulnerability, or standing.

Presentation overlay: the layer governing how findings are shown so they remain usable, honest, and proportionate.

Bounded finding: a finding stating what the audit can responsibly say, what it cannot say, and the next action.

Appendix D. A proposed pilot: testing the worksheet against ordinary critical reading

The development agenda (Section 19) turns on one question: does the Audit add anything over ordinary critical reading? This appendix proposes a concrete pilot to answer it, and consolidates the outcome measures referred to elsewhere in the paper. The design is offered as a recommendation, not a fixed protocol; the methodological choices below are flagged so they can be revised before any study is run.

Question and arms. The pilot compares what trained users produce when evaluating the same claims with different supports. The minimum is two arms: (1) the **Audit worksheet** and (2) a **plain critical-reading checklist** — a standard, generic

"identify the claim, weigh the evidence, watch for fallacies and bias" guide. A third, **no-tool** arm ("evaluate this claim") is worth adding to bound whether either structure helps at all. Critically, the checklist should be **dose-matched** — comparable in length and effort to the Audit worksheet — so that any advantage is attributable to the Audit's *architecture* (load-bearing identification, coherence-then-answerability, gates, true-core-versus-envelope) rather than merely to having more structure or spending more time on the task.

Design. A **between-subjects** design is recommended over within-subjects, despite the larger sample it requires, because the Audit teaches a transferable skill: a participant who learns to ask "what is the load-bearing claim?" in the Audit condition would carry the habit into a later checklist condition and contaminate the very contrast being tested. Random assignment, a short standardized training on the assigned tool, and matched groups protect the causal read.

Participants and scale. A first pilot can be modest — on the order of 30 to 60 participants per arm — drawn from a population with some analytic training (advanced students, or adults in civic, policy, or research roles). This is deliberately underpowered for small effects; the pilot's job, as Section 18 states, is to surface usable failure and a plausible signal, not to deliver statistical proof.

Materials. A shared set of roughly eight to twelve claims spans the categories the tool distinguishes: ordinary/casual, **bounded-empirical (well-formed and answerable as stated)**, overclaimed (a true core inside an overbroad envelope), protected-context, and LLM-generated with a fabricated citation. Including well-formed claims is essential — without them the discrimination measure below has nothing to detect. Protected-context items must be **synthetic**, not real restricted knowledge: the pilot must not itself expose what the tool exists to protect.

Procedure. For each claim, participants produce three things: (i) their statement of the main or load-bearing claim, (ii) a judgment of whether it is adequately supported and, if not, what fails, and (iii) a revised version of the claim. Tasks are time-boxed; the worksheets are collected for scoring.

Outcome measures. These are the measures referred to throughout the paper, consolidated:

- *Load-bearing identification (primary):* whether the participant correctly identifies the claim doing the structural work, scored against a pre-registered key by raters blind to condition.
- *Revision quality (primary):* whether the revised claim's scope and modality match the available support — removing unearned scope, hedging appropriately, retaining the true core, and flagging unverified evidence — rated on a fixed rubric by blind raters. This is the harder and more important test: diagnosis without improved revision is diagnosis without treatment.

- *Discrimination:* on the well-formed claims, the rate at which participants manufacture a critique of a claim that is answerable as stated (a false-positive rate). A tool that flags good and bad claims alike is generating critique, not discriminating; the Audit earns its keep only if its false-positive rate is the lower one.
- *Comprehension and conduct (secondary):* whether participants can state the next action; self-reported cognitive load and time-on-task; and, in the Audit arm, whether they avoid over-auditing casual claims and decline to process the synthetic protected-context items.
- *Reproducibility and reliability:* two or more raters per output, blind to condition, with agreement reported (for example, Cohen's κ or an intraclass correlation); and whether independent reviewers produce similar Audit findings on the same input — a check on legibility and reproducibility, **not** on truth (per the box principle of Section 2).

Pre-registration and analysis. The answer key, the revision rubric, the primary outcomes, and the analysis plan are fixed **before** data collection — itself an enactment of the paper's answerability principle. For a pilot, effect sizes with confidence intervals are more informative than significance tests. The disconfirmation conditions are those of Section 18: if Audit users do not identify the load-bearing claim more reliably and do not produce better-matched revisions, or if the Audit's false-positive rate on well-formed claims matches the checklist's, the pilot indicates the protocol adds nothing over ordinary critical reading.

Known limits. A single-sitting laboratory task measures neither durable skill transfer nor real-world use; the experimenter who trains participants can bias them; and a modest sample cannot settle small effects. These limits are the reason the pilot is framed as the first step of calibration rather than a verdict.