A Verification-Gated Architecture for AI-Assisted Humanistic Reading and Research Amid the Risk of Epistemic Erosion

Working paper, draft v13d — version timestamp 2026-08-22T20:09:51-04:00
HYDER HUSAIN ARASTU, University of South Florida, USA
AAKASH SHAHANI, Independent, USA
TAHER ALI, Independent, USA

This paper presents a verification-gated architecture for AI-assisted humanistic reading and research, built amid – not against – the risk that unchecked AI assistance erodes critical thinking, rather than the certainty that it does. That risk has moved from shared worry to peer-reviewed finding, and the evidence increasingly locates the harm in one avoidable design choice – answer delivery without a checkable step – not in generative AI as such. This paper's contribution is architectural, not evaluative: a systems and infrastructure contribution over a closed, pre-manifested corpus. A mechanical, script-run check settles whether a quotation resolves against that corpus; a rule-based join settles whether a claim inherits its status; neither settles paraphrase fidelity or source trust, decisions structurally reserved for a human or an independent reviewing agent at every result, not only failures. We demonstrate the architecture by applying it to itself: a meta-project in which two parallel moderating agents – one conducting this research, the other building a reference implementation of the design pattern under study – worked under one shared architecture across cycles of drafting, adversarial review, and re-verification against historical, committed snapshots. No empirical result about a reader's thinking is claimed; the evidence is the meta-project's own audit trail, and a narrower, single-project methodological claim rests on that same evidence. We differentiate the architecture from AI co-scientist systems by domain, corpus role, and terminal check.

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; *Interaction paradigms*; • **Applied computing** → *Arts and humanities*.

Additional Key Words and Phrases: critical thinking, verification, epistemic erosion, AI-assisted scholarship, humanistic reading

Working paper, draft v13d — version timestamp 2026-08-22T20:09:51-04:00.
Public archival copy; byline restored.

1 Introduction

The fear that new tools erode a reader's own thinking has, once again, moved from shared unease into the peer-reviewed record. A mixed-methods study of 666 participants finds frequent AI-tool use correlating with lower critical-thinking scores, an association its own design keeps correlational rather than causal [Gerlich 2025]. An EEG study of essay writing with a chatbot reports weaker neural and linguistic engagement [Kosmyna et al. 2025], a finding a subsequent methodological commentary presses – sample size, EEG-analysis reproducibility, reporting completeness – without overturning the underlying concern [Stankovic et al. 2026].

None of this is new in kind: the worry that a technology will do a mind's work for it and leave the mind weaker recurs across the history of writing technologies; large-language models are only its latest occasion.

What the recent literature adds is design-contingency. In a large field study across high-school mathematics classrooms, the same language model raised or lowered post-test performance depending only on its guardrails: an unguarded chat interface let students copy answers, leaving later unassisted performance worse than a no-AI control, while a guardrail-equipped version produced no such harm [Bastani et al. 2025]. The erosion the literature documents is a property of one common design choice – answer delivery without a checkable step – not of generative AI as such, surveyed next. This paper names that design-contingent property *epistemic erosion*: the hypothesis that an AI-assisted workflow degrades a reader's own critical thinking when it substitutes answer delivery for a checkable step, not a demonstrated general effect of AI assistance [Bastani et al. 2025; Gerlich 2025].

Philosophy of cognitive science already poses the same problem as an open design question. Closing an argument about which digital resources count as part of a person's own cognitive apparatus, Andy Clark poses, as a question for future research rather than a settled requirement, how the designers and users of new tools might exercise due epistemic caution [Clark 2015]. Clark draws the historical line himself: Plato's *Phaedrus* already gives "a clear statement of the fear that new-fangled inventions such as reading and writing will have catastrophic effects on human memory," and on the same page carries that fear forward to the present case [Clark 2025, p. 1]. The question this paper takes up is Clark's: not whether such a tool can erode a reader's thinking – the literature above already answers that, for one class of designs – but how its design might instead make that thinking checkable.

We present a verification-gated architecture for AI-assisted humanistic reading and research: a gated pipeline holding a closed, pre-manifested corpus, exposing AI-offloaded quotation and citation-linkage operations to three explicitly typed checks – a mechanical, script-run verbatim-substring match, a rule-based claim-inheritance join, and a severity-graded human-and-agent review. A human-ratification point is reserved for every result, not only failures, and each check is re-verified against historical, committed snapshots rather than trusted from a single run. The first check settles whether a quotation resolves; the second, only whether a claim inherits that status, never whether its paraphrase is a faithful entailment; the third is explicitly interpretive, not mechanical. This is a design and demonstration claim, not an empirical one.

We demonstrate the architecture by applying it to itself. Two parallel moderating agents ran this project: one conducted the research and drafted this paper. The other, under the same gates and ledgers, served as the parallel build agent, constructing a reference implementation of the design pattern under study – checked by the research line's read-only verification against its own source, drawn on below only to show how the architecture aids research. The project is accordingly a meta-project for the architecture it describes, carried out across iterative cycles of drafting, adversarial review, and re-verification. What this design and demonstration claim does not license is stated in one place next (§3).

This paper makes four contributions. We **reframe** the critical-thinking-erosion debate, arguing from causal and design-contingent evidence that erosion is a structural design risk, not a demonstrated property of AI use as such (§1–§2). We **derive** numbered design desiderata from the appropriate-reliance, cognitive-load, and cognitive-offloading literature and **specify** a verification-gated architecture that satisfies them, typed to the three checks above (§4–§6). We **report** a worked feasibility case: applying the architecture to its own production surfaced the documented failure modes that motivate it – hallucinated citation, silent

overclaim, false-negative extraction artifacts, unaccountable delegation – each caught by the architecture's own mechanisms on this one project, not established as a general property (§7). That evidence is the research line's own process, not a reader's use of the reference implementation, whose correspondence to this architecture remains a design thesis (§4), not itself demonstrated. We **differentiate** this architecture from AI co-scientist systems [Ghareeb et al. 2026; Gottweis et al. 2026] on domain, corpus role, and terminal check, detailed in §2.

2 Related Work

2.1 Erosion and design-contingent enhancement

A CHI-native literature documents generative AI's metacognitive costs and the conditions under which they are a property of design, not the tool itself. Tankelevitch et al. locate metacognitive demand at framing a prompt, calibrating trust, and deciding whether to apply the tool at all [Tankelevitch et al. 2024]. Passi and Vorvoreanu name the failure that last demand produces, overreliance on an incorrect recommendation, and its most-cited remedy, cognitive forcing functions, while warning an appeal to human oversight "can provide a false sense of security" where nothing enforces it [Passi and Vorvoreanu 2022]. Buçinca et al. supply the founding evidence and its cost in one result: forcing functions reduced overreliance, and participants rated the designs that worked best least favorably to use [Bucinca et al. 2021]. Vendrell and Johnston carry the design-contingency claim into higher education, arguing LLM effects are "contingent rather than fixed" [Vendrell and Johnston 2026]; Drosos et al. find AI-generated provocations induced critical thinking, shaped by task urgency and user expertise [Drosos et al. 2025].

Lee et al.'s survey of knowledge workers is the sharpest hook for our argument: confidence in generative AI predicts less enacted critical thinking, while what remains shifts toward verification and stewardship [Lee et al. 2025]. Two studies bound how far that shift can be designed for. Zhi, Kumar, and Lee find a temporal reversal: early access helps under time pressure and hurts with time to spare [Zhi et al. 2026a]. Zhi, Long, So, and Lee find dose is what matters instead: one AI interpretation aids comprehension, while heavy reliance trades it for foreclosure [Zhi et al. 2026b]. Bae et al.'s thirteen-participant study adds the reader as a further bound: the same unscaffolded assistance raised critical-thinking scores for experienced researchers and lowered them, with high variance, for novices, a result its authors call "indicative rather than definitive" [Bae et al. 2026]. Bo, Wan, and Anderson bound every later claim: across four hundred participants, interventions built to shape reliance "generally fail to improve appropriate reliance" [Bo et al. 2025]. Design intent is not, on its own, a warrant for a cognitive outcome.

Added transparency does not reliably calibrate trust – trust tracking a system's true capabilities, in Lee and See's founding sense [Lee and See 2004]. Kizilcec finds a bell-shaped relation between transparency and trust after an unexpected outcome [Kizilcec 2016]. Pareek et al. find any explanation attached to a credibility verdict raises reliance on it whether or not it is correct, because "users could not discern whether the AI directed them towards the truth" [Pareek et al. 2024]. A CHI'26 controlled study on a claim-evidence interface for scholarly Q&A sharpens the same hazard at our own object: displaying claim-evidence provenance lowered trust without changing behavioral reliance, and cost usability besides [Martin-Boyle et al. 2026]. None of these findings license a calibration claim for a display merely designed against, not tested for, that decoupling; our architecture makes no reader-facing trust or reliance claim for this reason. This literature

builds and tests interventions at the interaction — a forcing function, a friction principle, a provenance panel, each for one occasion. None asks what changes when verification is a standing property of an architecture, not an addition a user can decline.

2.2 AI-for-research systems

Three *Nature*-published systems anchor the state of the art for AI-assisted research. Google's Co-Scientist orchestrates six agents around a "scientist-in-the-loop" paradigm, screening self-generated hypotheses by tournament before a wet-lab terminal check [Gottweis et al. 2026]; FutureHouse's Robin closes that loop end to end, recovering a genuine drug candidate [Ghareeb et al. 2026]. Stanford's Virtual Lab pairs a principal-investigator agent with specialist and critic agents, architecturally closest of the three to our own moderator/sub-agent/ratification pattern, though its object is open-web science, not a closed corpus [Swanson et al. 2025]. All three pursue a different goal than ours, not a lesser one, and each concedes what remains undone for a claim's provenance. Co-Scientist names, as future work, "the development of agents with enhanced provenance capabilities to trace claims to specific figures or data within a source" [Gottweis et al. 2026]. Robin's hallucination check was manual and one-time, not a standing gate [Ghareeb et al. 2026]. A fourth system without *Nature*'s imprimatur runs a comparable check: Sakana AI's AI Scientist screens generated papers through its own simulated review [Lu et al. 2024], and its successor produced the first fully AI-generated paper to navigate a workshop-tier peer review [Yamada et al. 2025].

SPIRE brings a cousin of that architecture to the humanities: seven agents realizing Unsworth's Scholarly Primitives discover, annotate, and — unlike the division of labor we describe later — themselves bind citations and synthesize argument over classical Chinese and Latin corpora [Pan et al. 2026]. Its verification is generation-time and self-directed, scoring an essay against a gold benchmark once rather than as a check a reader could re-run. Its own account of the human's remaining part, "scholarly judgement, teaching responsibility, and peer evaluation remain with human researchers" [Pan et al. 2026], addresses the scholar producing an argument, not a reader checking one delivered. Narrower DH-domain terminal checks exist too: an LLM-assisted named-entity and cross-lingual annotation-projection check over a TEI/XML critical edition reaches over 97% projection accuracy [Galka and Vogeler 2026], and a framework embeds close reading into a retrieval-augmented-generation pipeline to improve its own output, not the reverse [Zhou et al. 2025]. Both target what a model produces while processing a source, not what a reader does with a delivered claim afterward, the distinction this paper's own object turns on. Fang et al. embed simulated peer agents into the reading interface, shifting critical-thinking behavior but leaving what the reader has offloaded unchecked [Fang et al. 2026]. Creative Reading raises the delegation risk from the reader's side, against tools that frame reading as "reading to discard" [Liu et al. 2026]; its target is what augmentation should preserve, ours a mechanism for checking what is already offloaded, and the two complement rather than compete. Berry names the nearest conceptual target without building it [Berry 2025]; Klein et al.: "Models make words, but people make meaning" [Klein et al. 2025].

A neighboring NLP literature runs a terminal check of a related kind, evaluated since ALCE established short-span citation attribution as this family's own benchmark standard [Gao et al. 2023b]. RARR edits unsupported spans out of a model's own output [Gao et al. 2023a]; FActScore decomposes a generation into atomic facts and scores the fraction a held source supports [Min et al. 2023]. Self-RAG critiques its own generation against what it retrieves [Asai et al. 2024]; SciFact verifies a scientific claim against a held

evidence corpus [Wadden et al. 2020]; AVeriTeC extends the same task to real-world claims checked against open-web evidence, reporting substantial agreement on its own verdicts [Schlichtkrull et al. 2023]. Each scores a model's own generation against a corpus once, at generation time — not even a recent, human-directed, re-runnable citation-verification tool checking an open, unbounded citation graph [Haan 2025] — embeds that check in a governance process with a disclosed false-positive taxonomy over a closed, pre-manifested corpus. SciTrue is closer still: a peer-reviewed, human-evaluated claim-verification interface that nonetheless checks a submitted claim once against the open Semantic Scholar index, not as a standing property of a closed, pre-manifested corpus [Tan et al. 2026]. It frames itself, in its own words, as "an assistive tool, supporting rather than replacing expert judgment, close reading, and peer review" [Tan et al. 2026]. That family's checkers are now benchmarked: across thirteen citation-metadata verifiers, a preprint benchmark finds "the false-positive rate, not recall, decides whether a verifier is deployable" [Reizinger and Brendel 2026] — independent support for our own disclosed false-positive defect taxonomy's premise.

2.3 Provenance-tracing and claim-adjudication architectures

A distinct, more recent literature builds general-purpose provenance machinery, not a domain-specific check; none engaged in this paper before. F(AI)^{2}R extends W3C PROV-O with a seven-rung, human-gated verification ladder whose top two rungs no AI agent can grant, reporting its own invalidation mechanism as designed, not shipped [Krebs 2026]. LEDGER reconstructs a layered, typed-edge trace graph from an agent's coding session for a human to traverse, but verifies no claim and discloses its own graph construction "is not fully deterministic" [Kim et al. 2026]. Evidence-Ledger Adjudication classifies a claim against evidence into one of four relations at drafting time, benchmarked against 2,335 external rows, but the check is a single LLM judgment run once, supported claims never routed to a human [Chen et al. 2026]. ScientistOne's Chain-of-Evidence is closest on typed multiplicity, running four mechanized integrity checks and reporting zero hallucinated references, but its checks run on code and logs for one paper-writing session, with no per-claim human gate and no reader-supplied corpus [Meng et al. 2026]. ProVe verifies a knowledge-graph triple against Wikidata text with a tunable threshold and curator-in-the-loop escalation, an older, non-agentic lineage for the mechanical, human-escalated pattern our own check descends from, applied to an open graph, not a closed corpus [Amaral et al. 2024]. Nearest of all is Jing's Traceable Scholarship, a normative framework for AI-assisted humanistic inquiry built on one Kant knowledge base. Its NO_EVIDENCE marker — an explicit refusal to answer past a scope boundary rather than fall back on the model's own knowledge — is the closest single mechanism in this literature to our own not-found disposition, independently arrived at [Jing 2026]. It discloses, though, no invalidation mechanism, no executed quantitative evaluation, and one case study authored and judged by its own builder. The W3C PROV family, the Web Annotation Data Model's quotation selector, and the Micropublications claim-evidence ontology supply the vocabulary layer this literature builds on, not cited in our account until now [Clark et al. 2014; Lebo et al. 2013; Moreau and Missier 2013b; Sanderson et al. 2017].

2.4 Positioning

Both bodies of literature above answer a different question than ours: what a system *makes possible* for a class of tasks, evaluated by what a reader or curator can subsequently do with it, is a distinct and legitimate contribution type from a measured behavioral effect [Olsen 2007; Wobbrock and Kientz 2016]. It

Table 1. Six components of a verification architecture, scored against the nearest systems in this literature. Y = present as the column asks; P = present but narrower or caveated; N = absent, including designed-but-not-shipped; – = not applicable (a vocabulary standard). Values are each system's own stated text, not inference.

System	Corpus closed	Exact-ness	Pred. typing	Invali-dation	Human gate	PROV-aligned
Our architecture	Y	Y	Y	N	Y	N
F(AI)^{2}R [Krebs 2026]	N	N	P	P	Y	Y
LEDGER [Kim et al. 2026]	N	N	N	N	Y	N
Evidence-Ledger Adj. [Chen et al. 2026]	N	N	P	N	P	N
ScientistOne/CoE [Meng et al. 2026]	N	P	Y	N	N	N
SPIRE [Pan et al. 2026]	P	N	N	N	P	N
ProVe [Amaral et al. 2024]	N	P	P	P	Y	N
SemanticCite [Haan 2025]	N	N	N	P	N	N
SciTrue [Tan et al. 2026]	N	N	P	N	N	N
ALCE/RARR/FActScore group [Gao et al. 2023a,b; Min et al. 2023]	N	N	N	N	N	N
PROV/Web Annot./Micropub. [Clark et al. 2014; Lebo et al. 2013; Sanderson et al. 2017]	–	P	–	Y	N	Y

is the type this paper claims. Read across Table 1, every comparator matches our own architecture on at most a few of six columns. Our own architecture rests on a closed, pre-manifested, dedupe-checked corpus — the research program's 215-entry manifest as of this build — checked by a mechanical, deterministic, verbatim-substring predicate with machine-derived locator pins. That check is typed apart from two other, differently-decidable checks rather than folded onto one ladder or one relation label, and gated by a human-ratification point reserved for every result, not a subset. Every result is also re-verified against historical, committed snapshots rather than trusted from a single run. PaperTrail is not a row here: it measures a display's effect on a reader's trust and reliance, not a verification architecture's own components, and its finding is addressed in Discussion instead [Martin-Boyle et al. 2026]. Stated plainly rather than elided: F(AI)^{2}R is stronger on the human-ratification ceiling and on standards alignment, both properties our account does not yet claim; ScientistOne's Chain-of-Evidence is stronger on typed multiplicity; SciTrue is stronger on peer-reviewed, human-evaluated attribution accuracy, a property outside this table's six columns. Jing's Traceable Scholarship is nearest on domain; Co-Scientist's and Robin's terminal check is stronger for their object than ours; SPIRE performs a synthesis we withhold by design. A closed corpus, a re-runnable mechanical check, and a reserved human gate, applied together to humanistic reading, is the combination this literature has not yet put together; our contribution is that combination, applied to the offloaded step it does not yet check as one.

3 Scope and Non-Claims

The architecture and demonstration above license less than they might seem to. Ten things are not claimed, collected here rather than left scattered:

No reader-facing evaluation. No reader's critical thinking, reliance calibration, or erosion was measured, moved, or prevented by the reference implementation.

No semantic or entailment verification. The claim ledger checks citability and an inherited quotation's mechanical status only, never whether a surrounding paraphrase is a faithful entailment of what the quotation says.

No dependency-invalidation propagation. Re-verification is a full-ledger resweep at each cycle boundary, not selective, change-triggered propagation from a recorded dependency edge.

No independent, cross-model verification. Every check is designed, run, and mostly reviewed by agents inside the project it audits.

No demonstrated implementation correspondence. That the reference implementation corresponds to this architecture is not itself demonstrated; it remains a design thesis (§4).

No statistical generalization from fault-injection results. Catching hand-designed corruptions is evidence about, not proof against, corruption types never sampled.

No causal warrant specific to humanistic reading. The design-contingency evidence's strongest causal anchor (§1) is drawn from K-12 mathematics; extending it here proceeds by domain analogy, not direct evidence.

No agent-line independence claim. The research and build agent lines are not independent in a statistical or epistemic sense; both share one director, one project, and one corpus.

No edition- or variant-aware verification. A quotation check confirms that a string occurs in the specific held source; it does not represent that a passage reads differently across editions, translations, or witnesses, and it offers no adjudication between contested readings.

No OCR, non-Latin-script, or facsimile reliability assessment. The verification checks depend on the held source's extraction fidelity. No assessment is made of the checking machinery's reliability on OCR'd sources, materials in non-Latin scripts, or facsimile images; such reliability is neither claimed nor measured.

4 The Verification-Gated Architecture

The gaps named above translate into five desiderata a verification-gated architecture must satisfy, stated as design commitments; none is a claim about what any single reading occasion demonstrates.

D1. Verification must be mechanical, not confidence-based. Self-reported reductions in critical effort under generative-AI use track confidence in a system's output, not its correctness [Lee et al. 2025]; trust calibration is itself a metacognitive demand a design satisfies structurally or leaves for the user [Tankelevitch et al. 2024]. The desideratum: a check run against a record outside the model's own assertion, never a score it reports about itself.

D2. Offloading must be checkable, not silent. A model unable to signal its own uncertainty is a documented hallucination failure mode [Ji et al. 2023], and interface guidance for human-AI systems calls for making clear what a system can and cannot do [Amershi et al. 2019]. The desideratum: every offloaded step carries a visible pass or fail against an outside record, with a not-found result treated as a first-class outcome.

D3. The corpus must be closed, manifested, and audited, not the open web. Multi-agent systems built for autonomous synthesis over an open literature [Pan et al. 2026] answer a different question: not what can be synthesized from everything available, but whether a fixed, itemized, deduplicated source set supports a specific claim. A check is only as strong as what it checks against, a design assumption rather than an established reliability advantage. "Closed" means closed at each audited snapshot, not fixed for the project's life: the held corpus has grown at every intake cycle (215 sources as of this build), under the same gate as everything checked against it.

D4. Challenge and contest must be designed in, not hoped for. Users of a multi-agent tool have asked, unprompted, whether it will ever push back on them [Adavi et al. 2026]; surfaced provocations have been shown to restore critical engagement unprompted AI assistance otherwise erodes [Drosos et al. 2025]. The desideratum: a reviewing role independent of the work it reviews, with a defined severity scale and a verdict that can withhold sign-off.

D5. Interpretive judgment must be structurally reserved for the human. Human-machine task allocation by comparative advantage – mechanical work to the machine, judgment to the person – is the original framing this architecture extends [Engelbart 1962; Licklider 1960]. This is narrower than a claim that only two kinds of decision exist; most work in §5 blends mechanical and interpretive elements. For one class specifically – what an argument claims, when a source may be trusted, when a draft is ready – no mechanical check and no agent verdict, at any model tier, may resolve the question in the reader's place.

Cognitive load theory holds working memory's capacity to hold and relate information at once sharply limited [Kirschner et al. 2006]: material low in element interactivity is tractable where the same material assimilated as one whole is not [Sweller 1994]. D1 and D4 apply this reduction twice, an inference from the design, not a measurement.

Offloading needs a construct, not only a name: an epistemic action is "a physical action whose primary function is to improve cognition by" cutting the memory a reasoner must hold and the probability of error [Kirsh and Maglio 1994]. Clark asks "in what ways designers and users of new tools and technologies might exercise due epistemic caution" [Clark 2015]. D1 and D2 answer that at one step; D3 through D5 chain the steps so no later stage rests on one never checked.

A narrow predicate registry. These commitments are enforced by named, individually scoped checks, not one undifferentiated "verification." Each is typed as one of three kinds, the distinction §6 applies throughout. *Exact* means a script alone resolves it; *structural* means a script resolves it by a fixed rule over a relation a human designed; and *human-gated* means no script resolves it, because what is decided is not the kind of thing a rule decides. Table 2 lists ten such checks – a designed mechanism is easier to individuate this way than a found one [Smart 2022, 2024]. Its last column carries equal weight to its operator column: what a pass does *not* establish is part of the check's specification.

Each check returns one of a small fixed label set, never a continuous score (`VERIFIED`/`TEXT-VERIFIED` for a match, `FUZZY`/`NOT_FOUND` for a failure, `NOT-VERIFIABLE`, `NO-KEY`, `MISMATCH`, `AUTHOR-RULED`). Three decision points are reserved for the director alone, closeable by no script or agent at any tier: adopting a thesis, changing a framing that alters an argument's central claim, and signing off on a finished draft as ready to stand behind.

Borrowed vocabulary, not a borrowed standard. The registry's two most load-bearing relations already have names in an existing standard: a quotation's link to its source is PROV-O's `wasQuotedFrom`; a claim's reliance on a linked quotation is `hadPrimarySource` [Lebo et al. 2013]; a locator pin's pointer to prose lines is the Web Annotation Data Model's `TextQuoteSelector` [Sanderson et al. 2017]. Neither standard checks the relation it names – PROV-O asserts a quotation without verifying it is verbatim; `TextQuoteSelector`

Table 2. A narrow predicate registry: ten of the checks the verification-gated architecture runs, typed by what decides each one. No row claims semantic or entailment verification – the registry contains none.

Check	Input	Decided by	Type	Passing does not prove
Quotation match	a quoted span	verbatim substring match (>0.99 similarity); a 0.95–0.99 near-match is disposed `FUZZY`, reported separately, never counted as `VERIFIED`	exact	completeness; entailment of the surrounding paraphrase
Witness resolution	a citation key	a five-rung fallback ladder	structural	the resolved source is right beyond string agreement
Extraction selection	a resolved source	a ten-tier variant preference order	structural	the chosen variant is itself defect-free
Claim inheritance	a claim's evidence link	a plain-string join to quotation status	structural	the claim's own prose paraphrases faithfully
Citability tier	source metadata	a four-tier publication rule	structural, agent-run	the tier assignment itself is error-free
Manifest dedupe	a hash, DOI, title-year	three-way exact match	structural	two entries are not the same work under a different edition
Locator placement	a (claim, location) pair	structural plus clause-boundary check	structural	the prose at that location says what the row claims
Fault-injection catch rate	sampled verified rows	four corruption families rerun through the match check	structural, sampled	robustness against corruption types never sampled
Gate and AI-review finding	a full draft	severity-graded reading, no fixed procedure	human-gated	correctness beyond that reviewer's own judgment at that reading
Director ratification	a proposed decision	a dated, human-authored record	human-gated	the ratified decision is correct beyond who decided

locates a passage without verifying it is faithful. The match and placement checks add the confirmation each vocabulary presupposes but does not supply.

Invalidation is not implemented; a full resweep is. The registry recognizes one dependency relation, unpersisted: a claim cannot inherit a status until its quotation resolves, but no field records which claim rows depend on which extraction-variant file, the way a build system's task graph would [Mokhov et al. 2018]. What is implemented, operating in §6, is a full-ledger resweep at every cycle boundary – the least efficient point in that literature's own rebuild-strategy space, with no dirty bit and no verifying trace. A genuine `wasInvalidatedBy`-style mechanism [Moreau and Missier 2013a] would need a persisted, per-quotation content hash, checked before re-running the match; no live schema carries one (§6 reports a research-only reference-tool test of this gap, twelve of twelve fixtures, same-author caveat noted there).

We take this composition as a design thesis, a mechanism-to-construct correspondence only. The research method itself – checkable, dependency-ordered offloading across narrowly scoped, reviewed agents – is a specific case of the function the reference implementation is being developed to accomplish for a reader, run by a second moderating agent instead. One commitment governs both, rather than two designs sharing a vocabulary by coincidence.

5 Pipeline Architecture

A single human director oversees two parallel top-level agents, each authorized to delegate its own task-specific work to further sub-agents. The paper-writing agent conducts the research program and drafts the paper the reader now holds; the parallel build agent constructs the reference implementation that §7 draws on solely to demonstrate how the method aids research. Figure 1 situates this division inside the gate structure below. The boundary between the two lines is read-only and was checked, not merely declared: a version-control snapshot of the implementation repository, taken before research work began, already shows that repository in a modified, actively developed state — dated evidence the build side was already at work on its own schedule. Both lines share one director, one project, and one corpus; nothing below claims independence beyond that boundary. The two lines feed each other only through the director's own decisions and discrete authored handoff artifacts: one documented instance is a structured specification, built from live inspection of the implementation, left ready for transfer to the build side. Its authored-and-ready status is what the record documents; whether the build agent applies a comparably disciplined process on receiving it is the build side's own account, not independently checked here.

Within each line, delegation is tiered by task complexity: lightweight models handle metadata extraction and formatting, mid-tier models handle annotation and citation mining, top-tier models handle synthesis, adjudication, and adversarial review. Every delegation is logged by role, tier, and task specification, and every sub-agent's output passes to a distinct reviewing agent before any later stage builds on it — answering the desideratum that offloading be checkable, not silent (D2, §4) [Amershi et al. 2019; Ji et al. 2023]. The tiering converts a per-task metacognitive burden Tankelevitch et al. [2024] identify into a one-time architectural decision, warranted by the comparative-advantage framing under D5 [Engelbart 1962; Licklider 1960]: even the top-tier synthesis agents remain sub-agents under human-ratified review, not autonomous authors.

Work does not move forward on a sub-agent's own say-so. It clears a sequence of seven named gates, Gate 0 through Gate F, specified in Table 3. No gate is self-certified: a gate's status (open, passed, passed with conditions, reopened) is a recorded fact about the project's history, not assumed from the fact that work continued.

The pipeline is cyclical, not sequential: a cycle closes on a full adversarial review and, where that review or a later ruling calls for it, reopens (Figure 1). Two episodes evidence this operating as a check, not a formality (full detail, including the disclosed fact that one surviving Gate-F record is itself a reconstruction, in Appendix B). The sibling manuscript the same architecture also produced saw a reopened Gate D under a second, independently commissioned sign-off after its central argument changed post-pass, and a Gate F review recorded as convened but *not* passed across two cycles, most recently on sixty-one findings, eleven severe. This paper's own trail runs through the same gates: three adversarial reviews of earlier drafts (the second discharging one severe defect and substantially addressing another; the third, on the tenth draft, returning twenty-two findings under major revision) and the standing AI-review stage's in-project vehicle, run twice under major revision. Against the fourth draft (nine drafts earlier) fifty-one graded findings accumulated; against the tenth (three drafts earlier) thirty-five. The mechanical checks answer to the same discipline: one check in §6 was found defective, repaired, and re-run four times, each evidenced by its own before-and-after count – a check never re-checked can accumulate exactly the drift the gate structure exists to prevent.

Table 3. The seven-gate structure. No gate is self-certified: each pass condition is either a mechanical check or a review role distinct from the work's own author.

Gate	Checkpoint type	Pass condition	Reviewing party
0	Scaffold	Directory structure, status tracker, and version control initialized	Moderator
A	Reference audit	Reference material fully inspected with no write access exercised	Moderator
B	Source intake	Every candidate source metadata-verified, tiered for citability, and deduplicated before manifest entry	Intake agent; moderator-reviewed
C	Literature review	Each disciplinary review verified against its search brief	Discipline-review agents
D	Thesis selection	Central argument selected and ratified by the human director; independent sign-off granted	Human director; independent reviewing agent
E	Mechanical sweeps	Instruction-leak, provenance-leak, tense, effect-verb, quotation, and typography checks all report pass	Automated sweep tooling
F	Adversarial review	Severity-graded review returns a verdict of pass, or sign-off is explicitly granted	Adversarial-review agent; human director

The governance record requires a standing AI-review stage between the adversarial review and the director's own read, run under either an in-project multi-agent synthesis or an author-commissioned audit by a reviewer independent of the drafting agents and tooling — a different model family, under an absolute no-edit rule. The stage's only output is a severity-graded findings report, never a revision, which is why it sits outside the seven-gate count: even a clean report grants no ratification. One external-vehicle audit has run, on this paper's third draft, under a governing prompt of 2,382 lines, intaken as a provenance-documented package with checksums for every preserved input. Its findings, set against the pipeline's own seven-dimension review of a later, fourth draft, overlapped at a Szymkiewicz–Simpson coefficient of 0.35 (full breakdown in Table 11, Appendix B). That comparison spans a version gap it cannot correct for, and one that biases the measured overlap downward, since an external-only finding may target text the later draft had already cut. One reading: every blocking finding was raised by both reviews; the other: the internal panel independently re-derived only about a third of the external register's findings. Both readings are accurate: the misses concentrate in the minor tail.

Where AI co-scientist systems select among machine-generated hypotheses through an Elo-scored self-play tournament [Gottweis et al. 2026] or an LLM-judged pairwise ranking [Ghareeb et al. 2026], this project applies the same competitive-selection logic to a document. Three prompted personas, run against one model family, not three independently sourced models, scored three candidates for a piece of scope; the surviving candidate was repaired against its sharpest critic's flaws and grafted with material its rivals contributed. Model-family diversity alone does not restore independent errors once judges share a prompt and a corpus: nine frontier models spanning seven families deliver only about two votes' worth of independent information [Kohli 2026]. So this panel's three scores are read as correlated readings from one model, not three independent votes — weaker than the AI-review stage above holds itself to. A self-critique ladder is excluded by the same non-self-certification rule that gives Gate F its bite; hypothesis evolution is excluded

because this method's rule is its structural inverse — the model indexes and verifies but never originates a claim (§6).

Before any source may support a claim, it passes a source-intake pipeline (D3, §4) running three checks in sequence. The first verifies metadata against the document itself, never a filename or citing paraphrase; the second classifies it into a four-tier citability rule (peer-reviewed work, primary sources, work accepted for publication, preprints), with everything outside those tiers held background-only. The third deduplicates at the work level, by hash, identifier, and normalized title-and-year, before registration in parallel manifests that must never diverge. A well-known trade paperback on study technique was excluded as background-only while an equally unfamous university-press monograph was retained as fully citable — the line runs by publication tier, not fame or topical fit. The resulting corpus is closed, manifested, and audited rather than assembled ad hoc from the open web, the condition the next section's verification claims depend on.

Every substantive ruling that shapes the argument — a directive from the director, a moderator's resolution of an open question, a change of thesis or scope — is entered in a dated, append-only record naming who decided, the rationale, and the row's current status. As of this build the record holds one hundred fifty-seven dated entries, one hundred five opening by naming the director rather than any agent or moderator role; the task ledger separately holds three hundred rows tracking individual delegated units of work. A superseded ruling is marked superseded, not deleted, so a reviewer can verify a reopened gate really was reopened on a logged ruling, not redrawn after the fact (D4).

None of this closes the loop mechanically. A small number of checkpoints are held out, by design, as points a mechanical pass or an agent's own synthesis cannot clear alone: adopting a thesis, changing a framing that alters the argument's central claim, and signing off on a finished draft as ready to stand behind. Each requires the director's own dated confirmation, distinct from an agent's verdict at any model tier — the structural answer to what Passi and Vorvoreanu describe as oversight functioning as false security [Passi and Vorvoreanu 2022]: the highest-stakes calls are routed through the director by construction, so the pipeline's terminal state cannot self-certify (D5). One request's path through all seven gates is traced in Figure 2.

6 Evaluation and Results

Three operations share the word "check" below. A quotation check is a mechanical, script-run verbatim-substring match against held text. A claim check is a rule-based, non-mechanical test of whether a claim inherits a linked quotation's status. A gate review is a severity-graded, human-and-agent judgment call, not itself decidable in the sense the first two are (§5 details the gate sequence).

6.1 Method: A Frozen Evaluation Snapshot

The results below run against a fixed *frozen evaluation snapshot* (Figure 3): a pre-registered, hash-verified evaluation protocol, frozen 2026-08-22 (RNG seed `20260822`), before any measurement ran. That protocol fixes the two mechanical ledgers and the source manifest at 748 quotation-ledger rows, 693 claim-ledger rows, and 193 manifest entries, along with the thresholds, sample sizes, and exclusion rules the results below draw on. The corpus's live size differs – 215 manifested sources, 771 quotation-ledger rows, 715 claim-ledger rows as of this build – ordinary intake growth, not a ledger repair. A companion 37-case fixture suite tests the checking machinery itself, not the corpus: 33 passed, zero failed, four informational disclosures, among

them that zero of the 193 frozen manifest entries carry a non-empty `title` field, a data-quality census, not a pass/fail result.

6.2 Results: Seeded Fault Injection

A pre-registered mutation catalogue corrupted a fresh, randomly seeded sample of the frozen ledger's verified rows and re-ran them through the quotation check (Table 4). It reports Wilson 95% intervals rather than a bare fraction, since boundary cases here (thirty of thirty, one of thirty) are exactly where a normal-approximation interval misleads. Two cheaper baselines – trust a citation because it resolves to a manifested source; trust the ledger's recorded status without re-inspecting the quoted string – ran against the identical sample and caught nothing at every class, confirming that detection requires inspecting the quoted text.

Table 4. Seeded fault injection against the frozen ledger, capability and negative-control classes ($n = 30$ drawn per class, seed 20260822). Wilson 95% confidence intervals; both baselines (citation-exists-only; trust-the-stored-status) caught 0.0% at every class, Wilson upper bound under 12%, confirming neither substitutes for re-inspecting the quoted string.

Class	n caught/valid	Rate	Wilson 95% CI
Fabricated quotation	30/30	100.0%	[88.6, 100]%
Polarity reversal	29/29	100.0%	[88.3, 100]%
De-hyphenation toggle[a]	29/30	96.7%	[83.3, 99.4]%
Curly-quote toggle	29/30	96.7%	[83.3, 99.4]%
Truncation (negative control)	**1/30**	**3.3%**	**[0.6, 16.7]%**

A truncated quotation is, by construction, still a genuine substring of the source. A check answering whether quoted words occur in order cannot distinguish a complete quotation from a shortened one, so the one-in-thirty catch here is noise around a structural zero, not a near-miss. It is reported at the same prominence as the higher rates above rather than left to recede into an aggregate. That the checker reliably catches transpositions, polarity reversals, and normalization edge cases says nothing about whether it catches a more complex or compound corruption never sampled. The coupling-effect assumption mutation testing relies on is itself untested here, an open premise, not a proven one.[1]

A second, hand-built set of classes – a fabricated attribution, a source-version mismatch, a citation left unchanged while its claim is strengthened or reversed, a source reclassified after a claim rests on it – all pass silently. None of the three checks tests whether a claim's prose is entailed by what it cites. Two structural classes behave as designed: deleting a claim's evidence link reclassifies it from linked to unlinked with no false retention (thirty of thirty); removing a manifest source resolves every dependent citation to unresolved. A third surfaces a population artifact, not a checker regression: only ten of thirty randomly drawn manifest entries carried a live citation key at all, nine of which resolved. That low raw rate reflects how much of the manifest a single paper cites, not a failure among entries actually cited.

[1]The de-hyphenation-toggle class as generated tests a quote-side proxy for the branch property; a dedicated predicate-suite fixture above tests the property itself.

6.3 Results: Dependency Invalidation Is Proposed, Not Implemented

The pipeline's re-verification is a full-ledger resweep at every cycle boundary, never a selective, dependency-triggered re-check of only the records an upstream change touched – the machinery keeps no persisted record, per source or per quotation, of which upstream input a check last ran against. In the vocabulary a build-systems formalism gives this design space [Mokhov et al. 2018], the posture sits at the least efficient, simplest point available: unconditional always-rebuild, with neither a dirty bit nor a verifying trace. What a genuine mechanism requires is stated precisely by the provenance standard's own definition of invalidation [Moreau and Missier 2013a]: a persisted record of which downstream entity an upstream change invalidates. None of that record exists in the live schema.

To give that gap a concrete artifact, a research-only reference tool (never the deployed pipeline) was built against eleven synthetic topologies (Appendix B) and exact-matched the expected affected set on twelve of twelve fixtures (Table 5; same-author fixtures and tool, so this confirms a from-specification implementation, not independent validation). The deployed pipeline's own affected set is the empty set on every fixture, by construction – a full resweep invalidates nothing selectively.

Table 5. Dependency-invalidation replay against a research-only reference tool, never the deployed pipeline. Twelve fixtures span eleven topology classes (one class split into two variants).

Metric	Result
Reference tool, exact match	12/12 (0 missed, 0 spurious)
Deployed pipeline's own affected set	$\emptyset$ on every fixture, by construction

6.4 Results: Historical Replay as Ecological Corroboration

Replaying the current checking machinery against six committed states of the ledger and corpus is ecological corroboration of prior, already-governed findings, not a controlled evaluation. Checker logic, extraction availability, corpus size, and status vocabulary all change together, so a rate change between two points cannot be attributed to one confound without separating them. One confound is separated: the witness-resolution repair's own before/after count retired 0, 50, 75, and 77 spurious unresolved citations at the four snapshots, while extraction and intake volume moved independently. The resulting trajectory (Table 6) is non-monotone, dipping mid-series before recovering – directionally reproducing an already-published finding on an independent re-implementation, not a re-run of the original tool, which was never committed.

Table 6. Historical replay, current-checker machine-checkable rate against each snapshot's own commit-era corpus. Prior-published rates (final column) come from an earlier, differently-implemented replay and are not expected to match exactly; the direction – a dip mid-series, then recovery – is what this replay directionally reproduces, not exact-value corroboration.

Snapshot	Rows	This pass	Prior-published
Cycle 1	458	84.0%	93.2%
Session 7	682	81.5%	87.5%
Cycle 3	691	81.3%	87.2%
Cycle 4	697	94.4%	99.9%
This evaluation's own snapshot	748	92.1%	—

6.5 Results: Claim-to-Quotation Coverage

A second ledger tracks argumentative claims against two conditions: the citability of the source a claim rests on, and, where the claim depends on a linked quotation, that quotation's own status. As of this build the claim ledger carries 715 rows project-wide; 533 (74.5%) are linked to at least one quotation. Of those, 517 carry a single quotation link and resolve under a plain join – 480 from an exactly-checked quotation, 37 from a weaker text-verified one. The remaining 16 linked claims carry a semicolon-separated multi-value link the plain join cannot resolve by design; this is a disclosed defect excluded from the split above, and a natural target for a future multi-value join schema rather than only a present limitation. The remaining 182 claims carry no quotation link at all.

6.6 No Semantic-Support Claim Is Made

Every check reported above answers whether a quoted string occurs in a held source, whether a citation resolves to the correct document, or whether a claim inherits a linked quotation's mechanical status – none tests whether the surrounding paraphrase is a faithful, entailed reading of what the quotation actually supports. A four-way annotation codebook for exactly that judgment (supported / contradicted / not-addressed / mixed) is specified and ready to run once qualified, independent human annotators are available; none is available this cycle, and no agent-generated label stands in for one. No semantic-support performance claim is made anywhere in this paper, for any claim–quotation pair, because no independently-labeled entailment judgment exists for any of them.

6.7 The Prior Self-Test Suite

Nine self-tests from earlier cycles remain the project's longest-running evidence base, full detail in Appendix B. They are a provenance-table audit, a witness-ladder replay retiring sixty-nine spurious failures, a fault-injection sample reproduced at larger n in Table 4, a claim-provenance coverage census, and a repair-episode ablation. The remaining four are a historical replay extended by one point in Table 6, a review-trajectory analysis that caught a fabricated records claim, a reviewer-diversity analysis, and a silent-edit census. None is superseded here.

6.8 One Failure, Diagnosed

The most informative failure this checking machinery has produced did not start out as one: a quotation check flagged a real, previously verified source as failing to contain text a passage had quoted from it, and an adversarial review escalated it as a severe citation-integrity defect. The diagnosis was not a citation defect. The source is set in two columns, and the extractor had scanned across the page in one horizontal pass, interleaving unrelated lines into the compared string – the quoted text had sat in the source the whole time, never shown to the checker correctly assembled. The finding was withdrawn, becoming a standing rule: no not-found result is recorded until re-run against every plausible extraction variant – both de-hyphenation branches, both layout-preserving and reflowed settings – since some sources verify under only one. A parallel audit found the same defect in thirty-three of the then-manifested hundred and twenty sources, responsible for eighty-six false-negative checks and zero genuine hallucinations.

An automated integrity tool needing scrutiny of its own mirrors the statcheck debate in meta-science, where a re-derivation script produced both false-positive and false-negative findings from its own parsing limits [Nuijten et al. 2016; Schmidt 2016] – the nearest lineage, not identical. Table 10 (Appendix B) tabulates this episode alongside the three others diagnosed and repaired over the project's life. Thirty-five false-failure verdicts were retired total, none a quotation misrepresenting its source, and one weakness survived all four, disclosed rather than closed – a surname-and-year fallback rung that could still resolve an unmanifested conference paper against the wrong held work, unexercised on the current corpus.

7 Bounded Self-Application Case

The architecture above is our contribution; the reference implementation below illustrates it and is not itself an evaluated system. The project that produced this paper ran the architecture on its own production: one paper-writing agent building the argument above, one parallel build agent directing, across the same cycles, the construction of a reference implementation the research program studies as it currently stands, not as a finished artifact – both under one human director. The two lines exchange findings but do not take each other's report on faith: a finding crossing from the build side is checked, read-only, against the build line's own source before the research side may use it, run at each cycle boundary and one-way, not a claim that the build line audits the research line's record in return. What follows is one bounded case of that discipline, a worked illustration, not independent validation across other teams, corpora, or domains.

One exchange shows the check running against the direction convenient to the paper. A draft proposed describing the reading plan's personalization as progressive withdrawal of support; the claim did not enter on the drafting agent's say-so. Checked, read-only, against the build line's own source, the finding came back narrower: five fixed stages, material re-ranked toward review as it is rated known, nothing removed and nothing new introduced as mastery grows. The paper reports the narrower finding because the wider one did not survive the check.

That check is the case's central point: a built capability is held unverified until read against its own source, mechanically and independent of who reports it – the same discipline the quotation ledger applies to a citation, extended from what a source says to what a system does. An agent distinct from the one that built a capability must verify it before either line may assert it as fact.

Where the reference implementation appears further in this account, it appears only as far as that discipline lets it, and only where it aids the research it supports. Three further capabilities pass the same discipline in brief. A credibility structure separates six independently displayed signals – publication rigor, creator expertise, host provenance, evidence strength, relevance, pedagogical value – rather than one confidence number, inheriting the same automation-bias exposure any such display carries [Kizilcec 2016; Pareek et al. 2024]. A retrieval layer excludes fabricated citations and claims lacking a grounded quotation, though the generated prose around a grounded citation is not itself checked sentence by sentence against what it says [Ji et al. 2023]. A discovery layer surfaces citation relations a reader never named, each carrying its own supporting quotation, so that adopting it means relating a new claim to a source already held, not accepting an unexplained assertion – the difference cognitive psychology calls self-explanation [Chi et al. 1981; Dunlosky et al. 2013; Kirschner et al. 2006]. Each capability is described here exactly as it entered the research record: a check the build line's own source passed, not a feature demonstrated in use.

One check is worth walking through in full, because it is the case's clearest instance of the architecture checking itself. An annotation layer reads a block of the primary text and asks a model to explain it, and the build line enforces, mechanically, that an annotation's supporting quote must be a verbatim substring of the block it annotates before the annotation is kept; an annotation whose quote does not occur in the block is dropped before it reaches a reader. Consider, as an illustration of that rule rather than a case drawn from the test suite, a block reading "Akrasia is weakness of will, acting against one's own better judgment." A candidate annotation whose supporting quote is "weakness of will" – a string the block actually contains – would be kept, carrying forward its summary and explanation; a candidate whose claimed quote does not occur in the block at all would be dropped. The same substring test the quotation ledger runs on a citation runs here on the tool's own reading of its own source. An annotation that survives that gate still reaches a reader as contestable, not settled: the interface carries Verify, Dispute, and Reject controls on the same card, so a reader who judges a genuinely grounded explanation wrong on its merits has a recorded way to say so – D5's reservation of interpretive judgment for the reader [Engelbart 1962; Licklider 1960] held at the point of contact, not merely inspectable after the fact. Whether a reader in practice uses that contest, or the card's transparency instead invites the deference Kizilcec and Pareek document, is left open; the design commits to reserving the decision for the reader, without showing that readers exercise it.

The case ran at project scale, not only in these two exchanges. Table 9 (Appendix B) tabulates the sibling manuscript's corpus and manuscript at each committed cycle close, with two cells left visibly in conflict rather than reconciled after the fact. The same checking machinery, replayed against four committed ledger snapshots, shows coverage moving from an adjusted 93.2% at the first cycle's close through a mid-project dip to 87.2% as intake outpaced extraction quality, to 99.9% after a corpus-wide re-extraction audit – a step function driven by identifiable repair events, not a smooth ramp: one interim schema revision dropped the ledger's verification columns entirely before a later cycle reinstated them. The paper before the reader shows the same signature at draft scale, tabulated in Appendix B: the mechanical sweep battery's hard-failure count fell to zero and held across four further drafts even as the battery grew from seven checks to ten and its statistics audit from nineteen to thirty checked figures, while a later, deliberately exhaustive review instrument returned fifty-one graded findings against the same paper – a jump measuring the instrument change, not draft decay, top-severity band held at two throughout. None of this is a claim that the check generalizes past the one project, team, and corpus that ran it, a limit section 8 states directly.

The project this paper reports applied its own architecture to verifying the artifact it was built to study, keeping, in committed form, the non-monotone record of that application rather than smoothing it after the fact – the case this section offers in place of a deployment study, bounded to one project, one team, and one corpus, across the cycles that ran it.

8 Discussion

Our claim is architectural: a citation that resolves against the held corpus or returns not-found is a mechanical, exact-match check – the predicate type §4's registry calls P1, not the weaker structural join a claim's inheritance gets or the interpretive judgment a gate review gives. It is decidable in that narrow sense, not a soft cue confidence can talk a reader out of noticing. That decidability has a measured referent: Lau et al.'s Critical Thinking in AI Use scale ($N = 1{,}341$) names "checking cited references" inside Verification, one of the construct's three constitutive factors, allowing verification to be initiated and supported by automatic

processes through cognitive offloading [Lau et al. 2026] – a rational, metacognitively-governed strategy, not a capability deficit [Risko and Gilbert 2016]. What the architecture makes standing and low-cost is one behavior that factor's own definition names – a claim about when the behavior is available at no cost, not about any reader's disposition or score.

That gap invites a substitution objection: a script performs the check here, not the reader, so a resolved citation could as easily read as offloading the behavior as exercising it. D5 answers at the design level – whether a source may be trusted and whether an argument claims what it says stay reserved for the human [Engelbart 1962; Licklider 1960]. Every kept annotation carries Verify, Dispute, and Reject controls on its card (section 7). Accepting a check's result requires an explicit reader action, open to contest, rather than following automatically from the check having run – a design affordance, not a claim about what a reader in practice does with it. Bae et al.'s self-initiated-verification marker is that same factor's hedged behavioral referent, not evidence this architecture moves anyone's score [Bae et al. 2026]. We also concede in full, rather than repeat here, the appropriate-reliance negative result and the Co-Scientist/Robin differentiation already stated in Related Work [Bo et al. 2025].

A verification pipeline auditing its own research process counts as an HCI contribution the way mechanism-level evidence counts in place of interaction data for a systems-and-infrastructure contribution [Olsen 2007; Wobbrock and Kientz 2016]. The two lines this paper reports are one architecture instanced twice, and the self-application makes that instancing visible rather than substituting for a reader study. The architecture presupposes a reader's competence; it does not substitute for it. Judgment concentrates at the pipeline's human ratification points and the reader's own contest actions on generated annotations; a reader with no competency to exercise there is left with a checkable record whose use still depends on judgment the architecture cannot supply.

Internal validity. Every check reported here – on the reference implementation and on this paper's own claims alike – was designed, run, and mostly reviewed by agents inside the one project, under the one director, that it audits: a common-mode risk no check in this record independently rules out. The checking machinery itself needed repair four times over the project's life, each diagnosed and converted into a standing rule rather than patched once – evidence the discipline catches its own faults, not evidence an outside party caught them.

External validity. No claim is made, or evidenced, that this architecture transfers past the one project, one director, and one corpus that ran it, which the research and build lines share. The fault-injection tests exercised five seeded corruption classes at $n = 30$ per class (Table 4), not the wider space of compound corruptions the Coupling-Effect assumption in mutation testing takes on faith. The strongest causal anchor for the design-contingency claim is drawn from secondary mathematics, not humanistic reading; carrying it here proceeds by domain analogy, not direct evidence. The held corpus is Latin-script, text-layer PDFs; OCR quality, facsimile-only sources, and non-Latin scripts are untested, and the extraction-variant defect class §6 documents was diagnosed and repaired entirely inside that same scope.

Construct validity. "Decidable" and "mechanical" name three qualitatively different checks. The first is an exact, verbatim-substring match at similarity above 0.99, with the 0.95–0.99 near-match band reported separately as `FUZZY` and never counted as exact. The second is a rule-based, join-only inheritance test deciding linkage and inherited status but never entailment. The third is an explicitly interpretive human-and-agent review – judgment, not measurement. No independently-labeled human judgment exists anywhere in this

corpus for whether a cited quotation actually entails the claim built on it – a semantic-support gap the claim ledger's citability-and-status check does not close. A related, structural gap (§6): the checking machinery runs a full resweep of the entire ledger at each cycle boundary rather than selective, dependency-triggered re-verification – the least efficient point in that design space [Mokhov et al. 2018]. It is kept because it is what caught two of the four checker repairs above, not because a selective mechanism was built and rejected. A related interface finds that provenance display can lower a reader's trust without lowering reliance on the system behind it [Martin-Boyle et al. 2026] – a caution against assuming this architecture's own contestable-annotation card moves reliance behavior as its design intends, since that card has not itself been tested against readers.

Future work should test, with independent readers, whether the standing, low-cost availability of these checks measurably changes verification behavior, rather than assuming the disposition literature's own hedge extends automatically to a designed occasion. No such reader study has been run, and this paper claims none.

The review trail behind this paper is part of that same record, reported plainly rather than smoothed. One task-ledger entry claiming no findings remained outstanding did not hold: independent audit found it contradicted on at least four findings. The verifiable floor is severity-weighted: both top-severity findings closed and independently re-derived; of eleven at the next severity, nine closed, one partial, one open by design. A cell-level census of the two ledgers and the source manifest, over their full committed history, matched 94.4% of 5,520 historical cell changes to a dated correction record under a broad match, 58.8% under a strict exact-cell accounting. The shortfall concentrates in one commit predating the project's records tooling; every commit after that tooling was installed shows a rate of at least 99.6%. The standing AI-review stage answers the internal-validity gap above partway, putting the manuscript in front of a reviewer from a different model family, outside the pipeline's own tooling.

We also do not argue a further, constitutive question: whether a generated annotation counts as part of the machinery realizing a reader's own capacity, rather than a tool that capacity uses. No reader was measured here, and the extended-cognition literature above enters only where it does design work.

Ethics and anonymity. This paper reports no human-subjects data: no reader was recruited, observed, or measured, and the held corpus is licensed material its owner controls, not scraped or redistributed. This manuscript was previously circulated in anonymized form for peer review; this public copy restores the identifying information. The data-provenance appendix (Appendix A) traces every project-process figure to a described project record rather than to a literal filename, commit identifier, or other searchable internal marker.

9 Conclusion

This paper describes one unified method, not two adjacent projects. Two parallel moderating agents carried it out: one conducted the research and drafting reported here; the other built, from the same design desiderata, the reference implementation section 7 describes. Each line answers to the same gates, ledgers, and human ratification points, and each depends on the other: the research line's own verification discipline instances the architecture it argues for, and the implementation is the object that discipline holds accountable. The project is a meta-project for the method it presents: how the paper itself was produced is a second, load-bearing case of the pattern, alongside the reading tool it describes. Whether the pattern generalizes past this one instance

is not a transfer we have tested, only a conjecture the design supports. The conjecture holds two conditions together. First, wherever a corpus can be held, a citation or quotation check can be made mechanically decidable in the narrow, exact-match sense argued above – not claim-inheritance or gate review, which stay structural and interpretive respectively. Second, a human is willing to close the loop with judgment rather than confidence. Where both hold, the index–verify–gate pattern is available to another human-directed, corpus-bounded research program.

Acknowledgments

Generative language models supported the moderating agents throughout this pipeline: literature search, drafting and revision of prose under human-set briefs, and the mechanical and adversarial checks described above. Two model families were used: Claude (Anthropic), across three capability tiers assigned by task type – routine and mechanical work, literature synthesis and drafting, and moderator-level verification judgments – for the pipeline's own agents; and Codex (OpenAI), as an external, different-model-family reviewer for one independent pre-submission audit, outside the pipeline's own tooling. The human director set every research and framing decision, ratified every gate the project's own record shows ratified, and bears responsibility for this paper. No generative model is listed as an author.

Data availability. A supplementary artifact package accompanies this paper, containing the predicate registry, the evaluation protocol, the frozen evaluation inputs, the analysis scripts, and the resulting run outputs. The package targets the *Artifacts Available* standard: no party independent of the authors has reproduced its results as of this writing. Two categories of material are deliberately excluded, and the exclusion is disclosed in the package's own limitations document rather than left implicit: the extracted-text corpus the quotation-verification check reads against, withheld because it is derived, near-full-text extraction from copyrighted third-party academic sources; and the project's full version-control history, which one evaluation component reads from directly and cannot be partially included. Where a component depends on either excluded resource, the package instead ships that component's already-executed output, the script that produced it, and a stated account of what an independent reviewer can and cannot confirm from the package alone; one component, the dependency-invalidation-replay fixture suite, is fully self-contained and reproduces its committed output byte-for-byte from the package's own contents.

On accessibility: this PDF declares its language, and every figure carries descriptive alt text; none of this paper's eleven tables carries machine-marked header cells, for the same reason. Full PDF/UA machine tagging was investigated and found not achievable with the project's current build toolchain – `acmart`'s tagging support requires the `tagpdf` package, absent from this project's TeX distribution – and is deferred to a future revision rather than attempted as an unsupported patch.

The full artifact package that accompanies the public release of this working paper is available separately, alongside its complete supplementary materials.

References

Krishna Akhil Kumar Adavi, Pratik Ghosh, Richard Banks, Advait Sarkar, and Sian E. Lindley. 2026. "Will This Tool Ever Push Back or Challenge Me?": Reflections on a Multi-agent LLM Tool for Perspective Seeking. In *Proceedings of the 5th Annual Symposium on Human-Computer Interaction for Work* (Linz, Austria) *(CHIWORK '26)*. Association for Computing Machinery, New York, NY, USA, 16 pages. doi:10.1145/3808045.3808058

Gabriel Amaral, Odinaldo Rodrigues, and Elena Simperl. 2024. ProVe: A pipeline for automated provenance verification of knowledge graphs against textual sources. *Semantic Web* 15 (2024). doi:10.3233/SW-233467

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. doi:10.1145/3290605.3300233

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. SELF-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511* (2024). doi:10.48550/arXiv.2310.11511

Seunghyun Bae, Hyungwoo Song, Hyeon Jeon, Taesup Kim, and Bongwon Suh. 2026. Supporting or Hindering? Understanding the Divergent Effects of LLM Assistance on Critical Thinking in Academic Reading. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems* (Barcelona, Spain) *(CHI EA '26)*. Association for Computing Machinery, New York, NY, USA, 5 pages. doi:10.1145/3772363.3798788

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics. *Proceedings of the National Academy of Sciences* 122, 26, Article e2422633122 (2025). doi:10.1073/pnas.2422633122

David M. Berry. 2025. AI Sprints: Towards a Critical Method for Human-AI Collaboration. *arXiv preprint arXiv:2512.12371* (2025). doi:10.48550/arXiv.2512.12371

Jessica Y. Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. 24 pages. doi:10.1145/3706598.3714097

Zana Bucinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1, Article 188 (2021). doi:10.1145/3449287

Gengyu Chen, Yongjie Yu, and Weiling Wang. 2026. Evidence-Ledger Adjudication for Claim-Evidence Traceability. arXiv:10.48550 [cs.CL]

Michelene T. H. Chi, Paul J. Feltovich, and Robert Glaser. 1981. Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science* 5, 2 (1981), 121–152. doi:10.1207/s15516709cog0502_2

Andy Clark. 2015. What 'Extended Me' Knows. *Synthese* 192 (2015), 3757–3775. doi:10.1007/s11229-015-0719-z

Andy Clark. 2025. Extending Minds with Generative AI. *Nature Communications* 16, Article 4627 (2025). doi:10.1038/s41467-025-59906-9

Tim Clark, Paolo N. Ciccarese, and Carole A. Goble. 2014. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. Journal of Biomedical Semantics 5:28 (2014). doi:10.1186/2041-1480-5-28

Ian Drosos, Advait Sarkar, Xiaotong (Tone) Xu, and Neil Toronto. 2025. "It Makes You Think": Provocations Help Restore Critical Thinking to AI-Assisted Knowledge Work. *arXiv preprint arXiv:2501.17247* (2025). doi:10.48550/arXiv.2501.17247

John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. 2013. Improving Students' Learning with Effective Learning Techniques: Promising Directions from Cognitive and Educational Psychology. *Psychological Science in the Public Interest* 14, 1 (2013), 4–58. doi:10.1177/1529100612453266

Douglas C. Engelbart. 1962. *Augmenting Human Intellect: A Conceptual Framework*. Technical Report AFOSR-3223. Stanford Research Institute.

Xinrui Fang, Anran Xu, Chi-Lan Yang, Ya-Fang Lin, Sylvain Malacria, and Koji Yatani. 2026. LLM-based In-situ Thought Exchanges for Critical Paper Reading. In *31st International Conference on Intelligent User Interfaces* (Paphos, Cyprus) *(IUI '26)*. 22 pages. doi:10.1145/3742413.3789069

Selina Galka and Georg Vogeler. 2026. Annotating, Projecting, and Interpreting Named Entities in Digital Scholarly Editions with LLMs. *International Journal of Digital Humanities* (2026). doi:10.1007/s42803-025-00114-8

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. doi:10.18653/v1/2023.acl-long.910

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)* (Singapore). Association for Computational Linguistics, 6465–6488. doi:10.18653/v1/2023.emnlp-main.398

Michael Gerlich. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* 15, 1, Article 6 (2025). doi:10.3390/soc15010006

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. 2026. A Multi-agent System for Automating Scientific Discovery. *Nature* 655 (2026), 497–505. doi:10.1038/s41586-026-10652-y

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, Jose R. Penades, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2026. Accelerating Scientific Discovery with Co-Scientist. *Nature* 655 (2026), 487–496. doi:10.1038/s41586-026-10644-y

Sebastian Haan. 2025. SemanticCite: Citation Verification with AI-Powered Full-Text Analysis and Evidence-Based Reasoning. *arXiv preprint arXiv:2511.16198* (2025). doi:10.48550/arXiv.2511.16198

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12, Article 248 (2023). doi:10.1145/3571730

Deyu Jing. 2026. Traceable Scholarship: Page Anchors and Ariadne's Thread for Humanistic Inquiry in the Age of Generative AI. arXiv:10.48550 [cs.CL]

Daehong Kim, Haichao Miao, and Shusen Liu. 2026. LEDGER: Claim-to-Evidence Trace Graphs for Auditing LLM Agents. arXiv:10.48550 [cs.CL]

Paul A. Kirschner, John Sweller, and Richard E. Clark. 2006. Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist* 41, 2 (2006), 75–86. doi:10.1207/s15326985ep4102_1

David Kirsh and Paul Maglio. 1994. On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science* 18, 4 (1994), 513–549. doi:10.1207/s15516709cog1804_1

René F. Kizilcec. 2016. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, CA, USA). 2390–2395. doi:10.1145/2858036.2858402

Lauren Klein, Meredith Martin, André Brock, Maria Antoniak, Melanie Walsh, Jessica Marie Johnson, Lauren Tilton, and David Mimno. 2025. Provocations from the Humanities for Generative AI Research. *arXiv preprint arXiv:2502.19190* (2025). doi:10.48550/arXiv.2502.19190

Guneet Kohli. 2026. Nine Judges, Two Effective Votes: Correlated Errors Undermine LLM Evaluation Panels. *arXiv preprint arXiv:2605.29800* (2026). doi:10.48550/arXiv.2605.29800

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. *arXiv preprint arXiv:2506.08872* (2025). doi:10.48550/arXiv.2506.08872

Florian Krebs. 2026. F(AI)2R: Who Did What, and Who Checked? Verifiable AI Provenance as an Executable Skill. arXiv:10.48550 [cs.CL]

Gabriel R. Lau, Wei Yan Low, Louis Tay, Ysabel A. Guevarra, Dragan Gašević, and Andree Hartanto. 2026. Understanding Critical Thinking in Generative Artificial Intelligence Use: Development, Validation, and Correlates of the Critical Thinking in AI Use Scale. *Computers in Human Behavior Reports* 22, Article 101103 (2026). doi:10.1016/j.chbr.2026.101103

Timothy Lebo, Satya Sahoo, and Deborah McGuinness. 2013. *PROV-O: The PROV Ontology*. Technical Report. W3C Recommendation. https://www.w3.org/TR/

Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, 22 pages. doi:10.1145/3706598.3713778

John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80. doi:10.1518/hfes.46.1.50_30392

J. C. R. Licklider. 1960. Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics* HFE-1 (1960), 4–11. doi:10.1109/THFE2.1960.4503259

Sophia Liu, Sarah Abowitz, Yijun Liu, Sarah Sterman, Shm Garanganao Almeda, and Max Kreminski. 2026. Creative Reading: Scaffolding Reading for Transformation. In *37th ACM Conference on Hypertext (HT '26)* (London, United Kingdom) *(HT '26)*. Association for Computing Machinery, New York, NY, USA, 8 pages. doi:10.1145/3800935.3830891

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint arXiv:2408.06292* (2024). doi:10.48550/arXiv.2408.06292

Anna Martin-Boyle, Cara A. C. Leckey, Martha C. Brown, and Harmanpreet Kaur. 2026. PaperTrail: A Claim-Evidence Interface for Grounding Provenance in LLM-based Scholarly Q&A. In *Proceedings of the 2026 CHI Conference on Human*

Factors in Computing Systems (CHI '26). doi:10.1145/3772318.3791101

Rui Meng, Bhavana Dalvi Mishra, Jiefeng Chen, Chun-Liang Li, Palash Goyal, Mihir Parmar, Yiwen Song, Yale Song, Rajarishi Sinha, Parthasarathy Ranganathan, Burak Gokturk, Jinsung Yoon, and Tomas Pfister. 2026. ScientistOne: Towards Human-Level Autonomous Research via Chain-of-Evidence. arXiv:10.48550 [cs.CL]

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. doi:10.18653/v1/2023.emnlp-main.741

Andrey Mokhov, Neil Mitchell, and Simon Peyton Jones. 2018. Build Systems a la Carte. In *Proceedings of the ACM on Programming Languages, Volume 2, Issue ICFP, Article 79*. doi:10.1145/3236774

Luc Moreau and Paolo Missier. 2013a. *PROV-DM: The PROV Data Model*. Technical Report. W3C Recommendation. https://www.w3.org/TR/

Luc Moreau and Paolo Missier. 2013b. *PROV-N: The Provenance Notation*. Technical Report. W3C Recommendation. https://www.w3.org/TR/

M. B. Nuijten, C. H. J. Hartgerink, M. A. L. M. van Assen, S. Epskamp, and J. M. Wicherts. 2016. The Prevalence of Statistical Reporting Errors in Psychology (1985–2013). *Behavior Research Methods* 48, 4 (2016), 1205–1226. doi:10.3758/s13428-015-0664-2

Dan R. Olsen, Jr. 2007. Evaluating User Interface Systems Research. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (UIST '07)*. doi:10.1145/1294211.1294256

Yating Pan, Jiajun Zhang, Jun Wang, and Qi Su. 2026. Extending AI for Research to the Humanities: A Multi-Agent Framework for Evidence-Grounded Scholarship. *arXiv preprint arXiv:2605.30947* (2026). doi:10.48550/arXiv.2605.30947

Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2, Article 383 (2024). doi:10.1145/3686922

Samir Passi and Mihaela Vorvoreanu. 2022. *Overreliance on AI: Literature Review*. Technical Report. Microsoft Aether (AI Ethics and Effects in Engineering and Research).

Patrik Reizinger and Wieland Brendel. 2026. HALLMARK: Diagnosing Three Failure Modes in LLM Citation Verifiers. *arXiv preprint arXiv:2607.18360* (2026). doi:10.48550/arXiv.2607.18360

Evan F. Risko and Sam J. Gilbert. 2016. Cognitive Offloading. *Trends in Cognitive Sciences* 20, 9 (2016), 676–688. doi:10.1016/j.tics.2016.07.002

Robert Sanderson, Paolo Ciccarese, and Benjamin Young. 2017. *Web Annotation Data Model*. Technical Report. W3C Recommendation. https://www.w3.org/TR/

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *NeurIPS 2023 (37th Conference on Neural Information Processing Systems), Datasets and Benchmarks Track*.

T. Schmidt. 2016. Sources of False Positives and False Negatives in the STATCHECK Algorithm: Reply to Nuijten et al. (2016). arXiv:1610.01010. https://arxiv.org/abs/1610.01010

Paul R. Smart. 2022. Toward a Mechanistic Account of Extended Cognition. *Philosophical Psychology* 35, 8 (2022), 1107–1135. doi:10.1080/09515089.2021.2023123

Paul R. Smart. 2024. Extended X: Extending the Reach of Active Externalism. *Cognitive Systems Research* 84, Article 101202 (2024). doi:10.1016/j.cogsys.2023.101202

Milos Stankovic, Ella Hirche, Sarah Kollatzsch, and Julia Nadine Doetsch. 2026. Commentary on Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., ... & Maes, P. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. *arXiv preprint arXiv:2601.00856* (2026). doi:10.48550/arXiv.2601.00856

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. 2025. The Virtual Lab of AI Agents Designs New SARS-CoV-2 Nanobodies. *Nature* 646 (2025), 716–723. doi:10.1038/s41586-025-09442-9

John Sweller. 1994. Cognitive Load Theory, Learning Difficulty, and Instructional Design. *Learning and Instruction* (1994). doi:10.1016/0959-4752(94)90003-5

Neşet Özkan Tan, Minghao Li, and Mark Gahegan. 2026. SciTrue: Evidence-Grounded Scientific Claim Verification and Traceable Summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026), Volume 3: System Demonstrations*. doi:10.18653/v1/2026.eacl-demo.27

Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 24 pages. doi:10.1145/3613904.3642902

Mireia Vendrell and Samantha-Kaye Johnston. 2026. Scaffolding Critical Thinking with Generative AI: Design Principles for Integrating Large Language Models in Higher Education. *Computers and Education: Artificial Intelligence* 10, Article 100572 (2026). doi:10.1016/j.caeai.2026.100572

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. doi:10.18653/v1/2020.emnlp-main.609

Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. ACM Interactions 23(3). doi:10.1145/2907069

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv preprint arXiv:2504.08066* (2025). doi:10.48550/arXiv.2504.08066

Jiayin Zhi, Harsh Kumar, and Mina Lee. 2026a. Investigating the Effects of LLM Use on Critical Thinking under Time Constraints: Access Timing and Time Availability. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (Barcelona, Spain) *(CHI '26)*. 21 pages. doi:10.1145/3772318.3791796

Jiayin Zhi, Hoyt Long, Richard Jean So, and Mina Lee. 2026b. What Does AI Do for Cultural Interpretation? A Randomized Experiment on Close Reading Poems with Exposure to AI Interpretation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (Barcelona, Spain) *(CHI '26)*. 18 pages. doi:10.1145/3772318.3791727

Jing Zhou, Li Si, and Wenjun Hou. 2025. Humanities-in-the-Loop: Using Close Reading as a Method for Retrieval-Augmented Generation (RAG). *Proceedings of the Association for Information Science and Technology* 62, 1 (2025), 1747–1749. doi:10.1002/pra2.1529

A Data-Provenance Table

As §6 states, every project-process figure that section reports is traced to the record it was read from here; this appendix is that trace, offered as verification support for the paper's checkability claim rather than as the claim's performance. Every project-process number asserted in §5–§7 is traced to the named record it was read from, so a third party holding the repository can recount each one. All values are *point-in-time* snapshots re-derived at final build (2026-08-20), on the project state finalized for this paper; the ledgers grow during active intake, so "as of this build" counters decay within hours, and a later reader must recount against the project's own records at their own read time rather than reuse these figures. An earlier run of this audit caught four such counters one intake batch stale; they were re-pinned, not reconciled silently. Source-file line numbers refer to the LaTeX source at this snapshot, and every one of them is checked, not merely asserted: each locator pin in the two tables below is derived by an audit script included with this paper's supplementary materials, so a reader can re-run it rather than take the pin on trust. The script's structural check confirms a pin's lines exist, stay in bounds, and never fall on a blank or comment-only line; its boundary check confirms a clause genuinely closes at or before a pin's start and within its span, reconstructed from the live section file rather than read off a single physical line, since these files wrap prose across line breaks. Its most recent run, after the Cycle-8 restructuring wave rewrote §3–§8 wholesale and this appendix's own tables were re-derived against the rewritten files (logged in this file's own header comments), reports 29 rows, 29 pins, zero mismatches — fewer than the pre-rewrite 39 rows, 50 pins because several claims the pre-rewrite draft made are no longer stated anywhere in the rewritten sections and were retired rather than re-pinned to content that no longer exists; that retirement is itself logged, not silent (this file's own header). Two prose figures are deliberately absent: the nine-judges/seven-families statistic and its two-votes corollary (§5) are a cited literature finding, not a project record, and §7's akrasia walk-through is labeled an illustration in the text itself. The cycle-close summary table and the coverage-trajectory figure that §7 cites are traced here by their headline series; individual pages and terminal-review cells rest on the

named close-commit messages and review artifacts, and the two cells where the committed records conflict with each other are disclosed in the table's own caption rather than reconciled here.

Table 7. Provenance of project-process figures in §6 (evaluation and results). MV = the Evaluation and Results section (§6). "Live" = re-derived from the working tree this pass. A prior version of this table also carried a paper-scoped citation-coverage figure sourced from §2's Positioning discussion; that figure is not currently stated anywhere in the rewritten manuscript and the row is retired rather than left pointing at content that no longer exists (this file's own header comments).

Claim in prose	Location	Live-record value (source)	Status
Frozen evaluation snapshot: 748 quotation-ledger rows, 693 claim-ledger rows, 193 manifest entries, sha256-pinned	MV:14	the frozen evaluation protocol record §3 (protocol checksum-pinned, confirmed against the evaluation gate report §2); the frozen-inputs checksum manifest, all four files `OK`	MATCH
Live corpus, as of this build: 215 manifest entries; 771-row quotation ledger; 715-row claim ledger	MV:14	Re-derived this pass against the quotation ledger (771: 722 VERIFIED / 48 TEXT-VERIFIED / 1 VERIFIED_PAGE_STRADDLE), the claim ledger (715), and the source manifest (215), all by direct `csv.reader` parse (2026-08-22, superseding the pre-evaluation preparation report's earlier-cycle figures)	MATCH
Predicate fixture suite: 37 cases, 33 passed, 0 failed, 4 informational	MV:14	the predicate fixture suite record, independently re-run and confirmed byte-identical apart from timestamp; the evaluation gate report §3.1	MATCH
Seeded fault injection ($n = 30$/class): fabrication 30/30, polarity 29/29, de-hyphenation toggle 29/30, curly-quote toggle 29/30 (Wilson lower bound $\geq$83%); both baselines 0.0% at every class	MV:28	the evaluation gate report §7 Table A, independently re-run and diffed byte-for-byte against the committed run, §3.2 (223/223 mutants, zero mismatches across all 15 classes)	MATCH
Truncation, negative control: 1/30 = 3.3%, Wilson [0.6%, 16.7%], reported at full prominence	MV:52	Same source, Table A	MATCH
Known-gap classes: deleted-evidence-edge 30/30 correctly reclassified linked→unlinked; omitted-corpus-item 9 of 10 resolved among the cited-key sub-population (30 raw draws, only 10 carrying a citable key)	MV:56	the evaluation gate report §7 Table B; §2 (H4 denominator-scoping finding: a population artifact, not a witness-ladder defect)	MATCH

continued on next page

Table 7 continued from previous page

Claim in prose	Location	Live-record value (source)	Status
Dependency-invalidation reference tool: 12/12 exact match (11/11 protocol rollup); live pipeline's own affected set $\emptyset$ on every fixture, by construction	MV:64	the evaluation gate report §7 Table C, §5; the verification-protocol record §10, §13	MATCH
Historical replay, witness-ladder repair isolated: spurious NO-KEY retired 0, 50, 75, 77 at the four historical snapshots	MV:85	the historical-replay record §4 (OLD-vs-NEW ladder isolation, per-snapshot NO-KEY counts 50/50, 50/0, 76/1, 77/0)	MATCH
Historical replay, six-point trajectory: this pass 84.0/81.5/81.3/94.4/92.1%; prior-published 93.2/87.5/87.2/99.9%	MV:85	the evaluation gate report §7 Table D	MATCH
Witness-ladder replay: 69 spurious no-key failures retired to zero	MV:133	the witness-ladder replay record §3 (736-row corpus; re-implementation from spec, not re-execution)	MATCH
Claim-to-quotation coverage, project-wide: 715 total, 533 linked (74.5%) — 517 single-`quote_id`, all 517 resolving (480 strict, 37 broad) — 16 multi-value excluded from the split, 182 unlinked	MV:112	Re-derived this pass by direct `csv.reader` join of the claim ledger against the quotation ledger, same join logic as the verification-protocol record §4/§8. Sanity check: 480+37=517; 517+16=533; 533+182=715	MATCH
Column-interleave in 33 of then-120 sources	MV:133	the extraction-integrity report l.77; scope l.3 (all 120)	MATCH
86 false-negative checks	MV:133	Same report ll.30–31 (86 rows across 16 sources)	MATCH
Zero quotations asserting absent text ("zero genuine hallucinations")	MV:133	Same report l.30 ("genuine is zero")	MATCH
Four repair episodes; 35 false failures retired, none a citation defect	MV:139	the repair-ablation record §3: 6+3+13+13 = 35, re-derived from each repair report's own before/after table	MATCH

Table 8. Provenance of project-process figures in §5 (pipeline architecture) and §7 (bounded self-application case). DM = the Bounded Self-Application Case section (§7); FG = the paper's own figures; AP = the Supplementary Process Tables appendix (Appendix B), where the word-budget trim pass relocated several self-contained process-evidence blocks out of MP/DM/FG.

Claim in prose	Location	Live-record value (source)	Status
One director, two top-level agents	MP:4–7	Charter record (the project charter); structural claim consistent with the governance log	MATCH
Handoff spec in ten evidence categories	MP:11–15	the reference-application handoff specification l.45 ("exactly one of ten categories")	MATCH
Seven named gates, 0 through F	MP:50–52	Gates 0, A–F = 7 (Table 3; charter gate structure)	MATCH
Reopened Gate-D: twelve conditions, four blocking	AP:168–171	the Gate-D sign-off record l.17	MATCH
Governing review prompt of 2,382 lines, provenance-documented package with checksums	MP:82–84	`wc -l` on the preserved prompt = 2,382; preserved-inputs table	MATCH
Register overlap 0.35 (Szymkiewicz–Simpson); read one way every blocking finding was raised by both reviews, read the other the internal panel independently re-derived about a third of the external register	MP:84–85	the reviewer-diversity analysis record §2c/§3: Szymkiewicz–Simpson $10.5/\min(51, 30) = 0.35$. Re-worded this pass from a one-sided "exceeds best internal dimension" framing to state both readings explicitly, per the external cross-family review's R09 disposition (the disposition table for that review, row R09) — §5's own prose now carries this softened form	MATCH
Three prompted personas, one model family, three candidates scored	MP:87	the scope record §8 via the method-adaptation record §2.1 (numeric scores recorded)	MATCH
Three intake checks; four-tier citability rule	MP:94	Intake-skill pipeline record; tier set fixed by logged ruling (USER-24 row)	MATCH

continued on next page

Table 8 continued from previous page

Claim in prose	Location	Live-record value (source)	Status
157 dated governance entries, one hundred five naming the director; task ledger 300 rows	MP:105–109	the project's decision log and the task ledger, re-derived this pass (2026-08-22, superseding the prior same-cycle pass this row reported): the decision log parsed for rows whose first pipe-delimited cell matches the date pattern — 157 dated rows, independently confirmed exactly; the task ledger — 301 lines = 300 data rows (header-excluded), independently confirmed exactly. The naming-the-director sub-count is re-derived this pass by a field-level parse of the decision log's own "Made By" column (values beginning `USER`) rather than the prior pass's coarser any-cell scan: 105 of 157 rows. Supersedes the prior pin's 143/94/264, stale relative to this cycle's continued governance growth (Wave-10 intake and remediation)	MATCH
Six credibility dimensions; popularity never moves a score	DM:38–42	the reference-application capability-verification record l.122 (six named dimensions; popularity assertion unit-tested)	MATCH
Five fixed stages; known items re-ranked toward review, nothing removed or newly introduced	DM:23–26	Same file ll.21–36 (verdict PARTIAL, scoped to exactly this narrower wording)	MATCH
Cycle-close table: manifest 55/88/120/128; quote rows 458/683/691/697; draft 30 to 54 pages	AP:70–72	Manifest and quote-ledger series re-derived this pass by CSV parse of the project's four cycle-close commit records; pages and terminal-review cells per the close-commit messages, the project memory's phase record, and the named review artifacts	MATCH

continued on next page

Table 8 continued from previous page

Claim in prose	Location	Live-record value (source)	Status
Coverage trajectory (§7's own, narrower three-point form): 93.2% at cycle 1's close, dipping to 87.2% mid-project, recovering to 99.9% after re-extraction	DM:93–100	the coverage-trajectory replay record §2, machine replay against committed snapshots. **Narrowed this pass:** the prior pin's five-point figure (93.2/87.5/87.2/99.9/99.9%) and its FG co-pin no longer match §7's own text — §7's rewrite states only three points inline, and the `fig:trajectory` figure that once carried the fuller series was removed in this cycle's figure rebuild (confirmed by grep, zero hits for these percentages anywhere in the paper's figures; the new `fig:eval-design` figure that replaces it depicts the evaluation *design*, not this result). Re-pinned to §7's prose alone, narrowed to the three points it actually states, rather than asserting a five-point MATCH the live file no longer supports	MATCH
Sweep battery grew seven to ten checks, statistics audit nineteen to thirty; hard-failure count held at zero across four further drafts; a later exhaustive review returned fifty-one graded findings; top-severity band held at two throughout	DM:95–103	the review-trajectory analysis record §1/§5 (v1–v5 sweep-report table: 7/9/10 sweeps; Sweep-10 scope 19/21/30) and the early-draft review records' severity registers, re-tallied in the review-trajectory analysis record §2. **Merged this pass:** §7's rewrite states the sweep-battery-growth fact and the draft-scale-severity fact in one sentence where the prior pin held them as two separate rows at two separate locations; merged here to match, not duplicated	MATCH

B Supplementary Process Tables

B.1 Cross-Cycle Corpus and Manuscript Growth

The sibling manuscript's growth across its own committed cycles – referenced in §7 as evidence the demonstrated cycle ran at real scale – is tabulated here in full so §7 need only point to it. Table 9 gives the corpus and manuscript at each committed cycle close, with two cells left visibly in conflict rather than reconciled after the fact: the committed manifest count against the project's own memory file, and, at one cycle close, the manifest's own CSV against its JSON. Both conflicts are left standing on purpose: no source read for this table settles which committed record is correct, and silently picking one would misrepresent a genuine, unresolved disagreement in the historical record as a settled fact – disclosing the conflict is more honest than reconciling it without evidence that would justify the choice.

Table 9. The demonstration at scale: corpus, manuscript, and terminal review at each committed cycle close of the sibling project, a humanities research paper, each figure parsed from that cycle's close commit of the shared record. [a]The manifest row is parsed from the manifest CSV at each cycle's close commit. Two conflicts are disclosed rather than reconciled: the project's committed memory file at the cycle-2 close records 60 and 89 entries for cycles 1 and 2 against the CSV's 55 and 88, and at that same commit the manifest's two committed serializations themselves disagree: 88 CSV rows against 89 JSON entries. No source read for this table resolves either conflict, so the figures are reported, not adjusted. [b]The cycle-4 severity mix survives only in a disclosed after-the-fact reconstruction of that review report (section 5). Word counts were never recorded for this manuscript at any cycle close; pages are the only recorded length measure.

	Cycle 1	Cycle 2	Cycle 3	Cycle 4
Source manifest, entries	55[a]	88[a]	120	128
Quotation ledger, rows	458	683	691	697
Draft manuscript, pages	—	30–42	44	54
Terminal review of the cycle	9 discipline reviews verified	major revision, 9 findings; then accepted with revisions	97-item external correction pass	major revision: 61 findings, 11 severe[b]

B.2 Repair-Episode Ablation

§6 narrates one of the four dated repair episodes to the mechanical quote-verification sweep in full (episode 1, the column-interleaved-extraction case); Table 10 gives all four, including the three (episodes 2–4) referenced there only by count.

Table 10. The four dated repair episodes to the mechanical quote-verification sweep. Each row is that repair report's own before/after delta, measured against the manuscript as it stood at that run — span totals differ because the manuscript changed between runs, and only episode 4 ran against the final-cycle build. Across the four episodes, 35 false-failure verdicts were retired and zero proved to be genuine citation defects: every failure traced to the checker, none to a quotation misrepresenting its source. Two scope disclosures: a fifth tooling-repair report on disk repairs a different sweep (the effect-verb gate) and a different defect kind, and is excluded here; and one known checker weakness survives all four repairs — the surname-and-year fallback rung can still resolve an unmanifested conference paper to the wrong work — disclosed rather than fixed, and exercised by no span in the current build.

Episode	Date	Checker-side defect	Spans	False failures retired
1	2026-07-26	Column-interleaved extraction read as stored; citation key assumed to equal filename stem	45	6
2	2026-07-27	Witness slug missed by key-based resolution; accented characters deleted rather than Unicode-folded	45	3
3	2026-07-27	Chapters resolved to the wrong same-editor volume; elision marker stripped before the elision test ran	83	13
4	2026-07-27	Unmatched proceedings title dead-ended the witness ladder for conference-paper entries	69	13
Total retired (of which genuine citation defects)				35 (0)

B.3 Reviewer-Diversity Corroboration

§5 states the topline Szymkiewicz–Simpson coefficient (0.35) between this paper's external (cross-family) and internal (seven-dimension) AI-review registers, and both readings of that coefficient, in body prose. Table 11 gives the granular breakdown behind that coefficient.

Table 11. Reviewer-diversity corroboration between the external (cross-family) AI-review audit of this paper's third draft and the in-project seven-dimension AI-review synthesis of its fourth draft. Figures re-derived from the project's reviewer-diversity self-test report, §2b, §2c, §3, §5.

Metric	Value
Szymkiewicz–Simpson overlap coefficient	0.35
Blocking findings corroborated	All; one pair numerically identical; two external blockers register on the internal side only as narrative observations, not formal findings
Minor-tail agreement rate	Below 1 in 5
Cross-family reviewer's unique yield vs. best single internal dimension	Exceeds it
Internal panel's independent re-derivation of the external register	About 1/3
External register findings the internal panel missed outright	About 2/3
External-to-internal strict rate (same numerator/denominator as the Simpson coefficient, read the other way)	35.0% (10.5/30)

B.4 Sibling-Manuscript Gate Episodes

§5 illustrates the pipeline's cyclicality with two episodes from its sibling manuscript, a humanities research paper the same architecture also produced, not this paper itself, and gives both in full here.

Gate D, thesis selection, was not a one-time hurdle there: when that paper's central argument changed materially after Gate D had already passed once, the pipeline did not carry the earlier ratification forward. It reopened Gate D, required a second, independently commissioned sign-off under the new framing, and recorded a new pass state — twelve conditions, four blocking — before later work could depend on the revised argument.

Gate F, the terminal adversarial review, shows the same discipline in reverse: across two successive cycles it returned major revision, most recently across five argument dimensions with sixty-one findings, eleven severe, and was recorded as convened but *not* passed, sign-off withheld pending the director's own read-through — a ruling the director then confirmed explicitly, not by default. The surviving record of that episode is itself a reconstruction, assembled after the fact from sub-agent logs once the contemporaneous review report turned out never to have been written, disclosed here because surfacing gaps in its own record is the architecture's own standing claim.

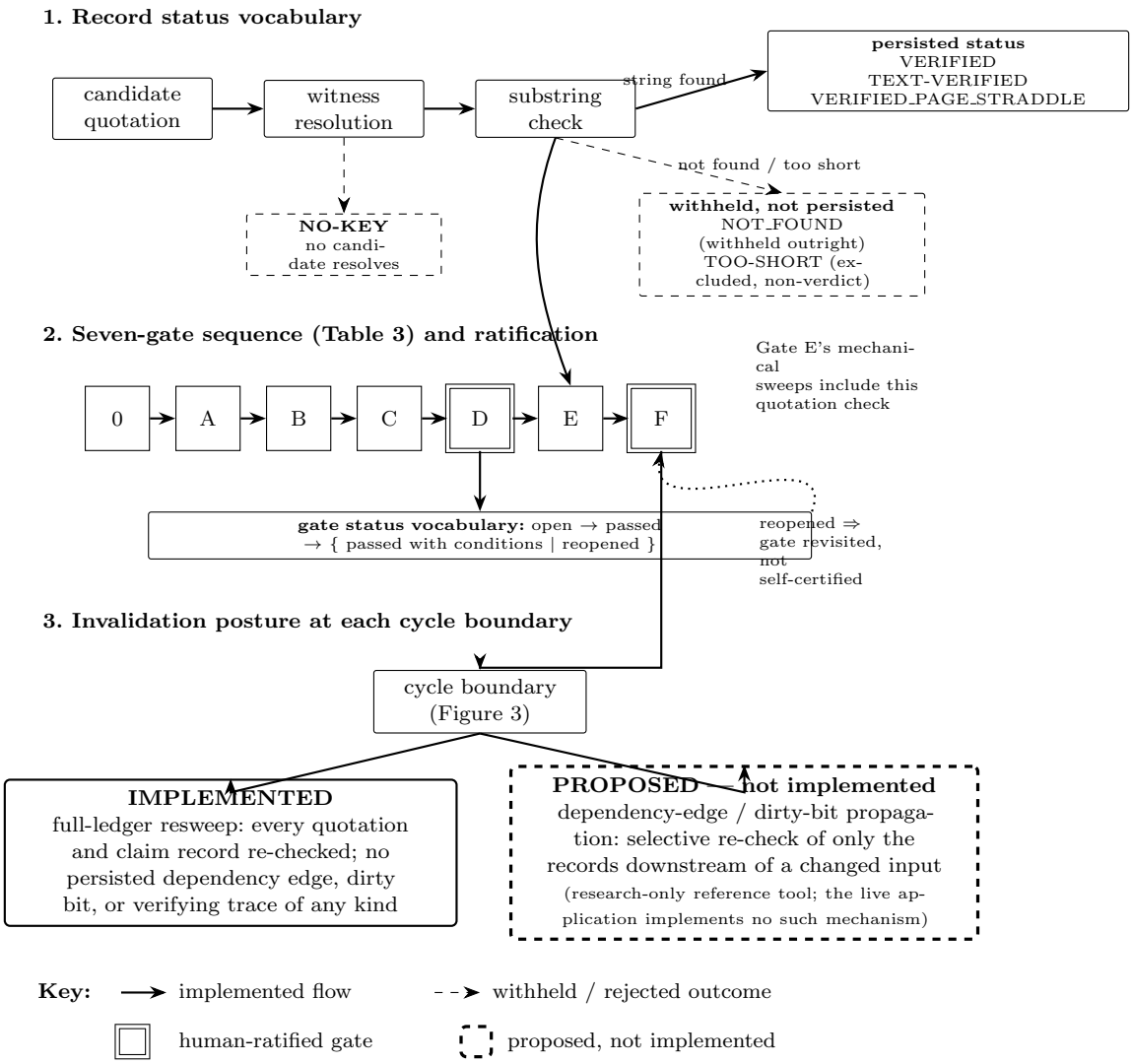

Fig. 1. Contract state and invalidation flow. Band 1 shows the record status vocabulary a checked quotation can reach: a candidate quotation passes through witness resolution and a substring check; a resolved, matching check persists as one of three accepted statuses (VERIFIED, TEXT-VERIFIED, or the documented page-straddle limitation VERIFIED_PAGE_STRADDLE), while a failed match is withheld outright (NOT_FOUND) or excluded as a non-verdict (TOO-SHORT) rather than persisted, and an unresolved witness returns NO-KEY. Band 2 places this check inside the seven-gate sequence of Table 3, with the gate-level status vocabulary — open, passed, passed with conditions, or reopened — and the two human-ratification points, gate D (thesis selection) and gate F (terminal adversarial review), drawn with a double border. Band 3 shows the two invalidation postures at a cycle boundary: the architecture's IMPLEMENTED posture is a full-ledger resweep with no persisted dependency edge, dirty bit, or verifying trace, re-checking every record rather than only the ones a change affects; a PROPOSED selective-invalidation mechanism, drawn with a heavy dashed border and labeled explicitly rather than left to line style alone, exists only as a research-only reference tool (Figure 3) and is not part of the live application. No dependency-invalidation propagation is implemented in the reference implementation this paper describes.

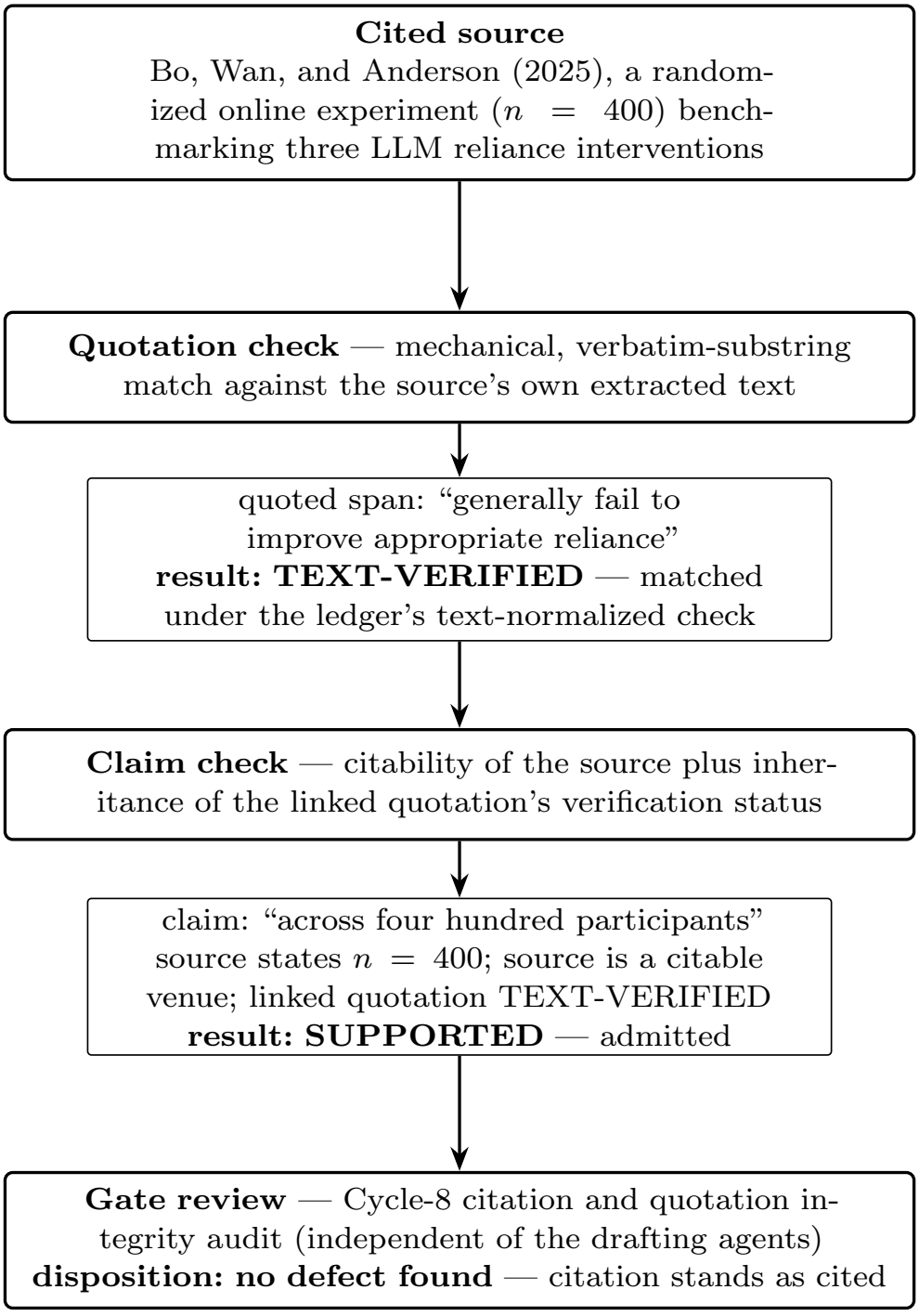

Fig. 2. One worked quotation-to-claim-to-gate path, illustrated with an already-cited, TEXT-VERIFIED example from the Cycle-8 citation and quotation integrity audit rather than a constructed illustration. A quotation from Bo, Wan, and Anderson (2025) is checked by a mechanical, verbatim-substring match against the source's own extracted text and returns TEXT-VERIFIED, the quotation ledger's status for a match confirmed under its text-normalized extraction path rather than the stricter default VERIFIED tier; a claim resting on that source's reported sample size inherits the quotation's TEXT-VERIFIED status and the source's citable classification, and is admitted as SUPPORTED under the claim check's rule-based inheritance test, which treats either match tier as sufficient; the pair then passes through the gate-review layer, here the Cycle-8 audit that independently re-checked every quoted span and its attendant claims against the primary source and found no citation-integrity defect in this pair. Each of the three predicates — the mechanical substring match, the rule-based inheritance test, and the severity-graded gate review — is explicitly typed apart from the other two, per this architecture's central design commitment, rather than folded onto one ordinal ladder.

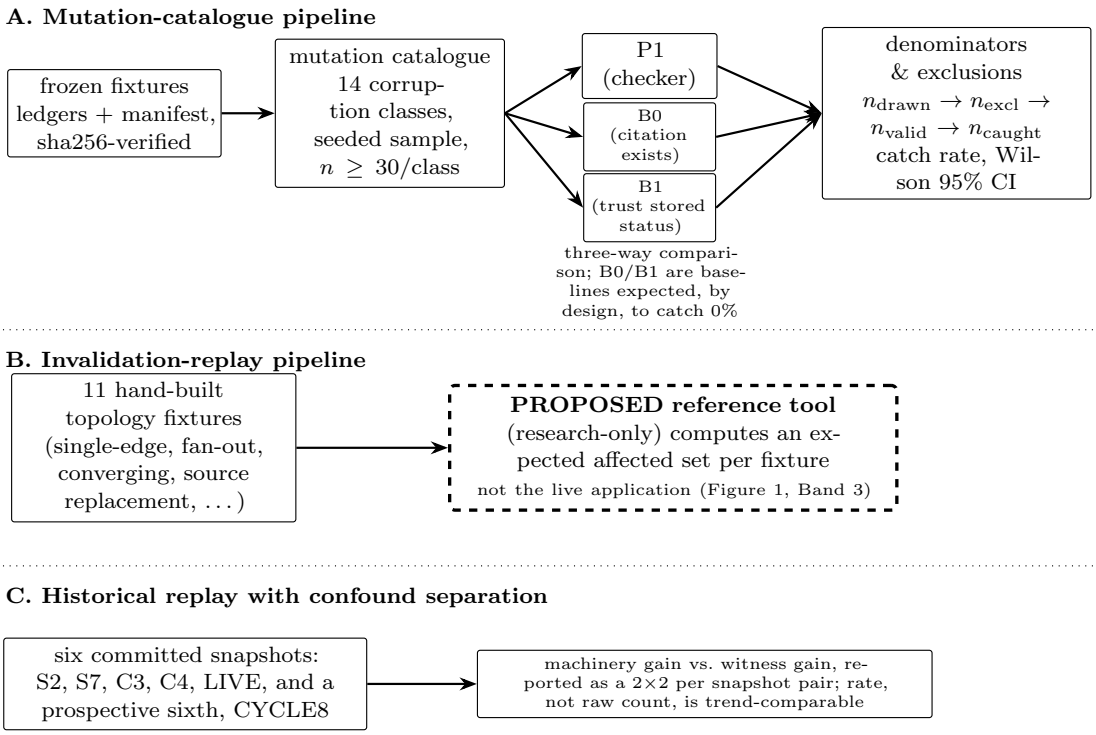

Fig. 3. The evaluation design this project's Cycle-8 protocol specifies, pre-registered before any measurement ran; results against this same frozen design are reported in Table 4 (Row A), Table 5 (Row B), and Table 6 (Row C), so this figure depicts the design those results were measured against rather than duplicating them. Row A: a seeded, sha256-verified frozen snapshot of the ledgers and manifest feeds a fourteen-class mutation catalogue (fabricated quotations, truncations, polarity reversals, and structural gaps such as attribution error and stale status among them), sampled at a minimum of thirty draws per class; the corrupted sample is scored under a three-way comparison — the real checker (P1) against two baselines, citation-exists-only (B0) and trust-the-stored-status (B1), both of which are expected by design to catch nothing — and every catch rate is reported with its denominators, its logged exclusions (normalization-equivalent, too-short, or malformed constructions), and a Wilson 95% confidence interval rather than a bare fraction. Row B: eleven hand-built graph topologies, each representing a distinct dependency shape a source, quotation, or claim edit could propagate through, are scored against a PROPOSED, research-only reference tool that computes the expected affected set per topology; this tool is explicitly not the live application, which implements no invalidation propagation at all (Figure 1, Band 3). Row C: the protocol extends the project's existing five-snapshot historical replay with a prospective sixth point and requires any future run to separate machinery gain from witness gain as a full two-by-two per snapshot pair, and to compare rates rather than raw ledger-row counts, since the ledger itself grows across the series independent of any quality change.