Erosion Is a Design Default, Not a Destiny: A Verification-Centered Architecture for AI-Assisted Humanistic Reading and Research

Working paper, draft v12 — version timestamp 2026-08-20T20:07:17-04:00

HYDER HUSAIN ARASTU, University of South Florida, USA
AAKASH SHAHANI, Independent, USA
TAHER ALI, Independent, USA

Generative AI's threat to critical thinking has moved from shared worry to peer-reviewed finding – an old fear resurfaced by current large-language-model developments. The same evidence increasingly locates the harm in one avoidable design choice, answer delivery without a checkable step, rather than in generative AI as such. We present a verification-centered research method for AI-assisted humanistic reading and research, and demonstrate it 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 verification architecture across iterative cycles of drafting, adversarial review, and re-verification. We argue by design analysis and the meta-project's own audit trail – no empirical result is claimed – that the method makes every offloaded step the ledgers cover a decidable check, the remainder disclosed rather than checked, while reserving interpretive judgment for the person doing the work, and we differentiate it 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: AI-assisted reading, verification, cognitive offloading, critical thinking, appropriate reliance, human-AI collaboration, digital humanities, research pipelines

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 consistent with a correlational rather than causal design, built on self-reported and assessment-based measures [Gerlich 2025]. A CHI survey of knowledge workers reports self-reported reductions in cognitive effort under generative-AI use, alongside a shift in their confidence in unaided judgment [Lee et al. 2025]. An EEG study of essay writing with and without a chatbot reports weaker neural and linguistic engagement under assistance [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 underlying 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 of the identical model produced no such

Authors' Contact Information: Hyder Husain Arastu, University of South Florida, Tampa, USA, harastu@usf.edu; Aakash Shahani, Independent, USA; Taher Ali, Independent, USA.

harm [Bastani et al. 2025]. The erosion the literature documents is not a settled property of generative AI as such, but of one common design choice – answer delivery without a checkable step – that a design is free to make or refuse.

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-centered research method for AI-assisted humanistic reading and research, and argue, by design analysis and worked demonstration rather than by empirical result, that the method gives the feared offloading a mechanically checkable form: every offloaded step the ledgers cover is a decidable check against a held, audited corpus, with paraphrase and surrounding prose outside that coverage disclosed as uncovered rather than checked, and every interpretive judgment is structurally reserved for the person doing the work. We demonstrate the method by having applied it to itself. Two parallel moderating agents ran this project: one conducted the research reported here and drafted this paper; the other, under the same gates and ledgers across the cycles that built it, served as the parallel build agent, constructing a reference implementation of the design pattern the research investigates – current as of this paper's own cycles, when the research line's read-only verification checked it against its own source without claiming it finished – drawn on below only to show how the method aids research. The project is accordingly a meta-project for the method it describes, carried out across iterative cycles of drafting, adversarial review, and re-verification – including at least one gate, drawn from a sibling manuscript, reopened and re-passed under revised grounds – and its own record of that work stands as the demonstration.

This paper makes four contributions. We **reframe** the critical-thinking-erosion debate, synthesizing causal and design-contingent evidence to argue that erosion is a property of default AI designs, not 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-centered architecture that satisfies them – a gated pipeline, dual mechanical quote and claim ledgers, adversarial review, and human ratification points (§3–§5). We **demonstrate** the method by applying it to itself: the meta-project's own audit trail – a gate reopened and re-passed in the sibling manuscript the same architecture produced, a corrected extraction artifact, adversarially reviewed drafts – shows instances of the documented failure modes that motivate the method caught and corrected by the architecture's own mechanisms: hallucinated citation, silent overclaim, false-negative extraction artifacts, and unaccountable delegation (§6). That evidence is drawn from the research line's own process – the paper's own citation, quotation, and claim checks on itself – not from a reader's use of the built reference implementation, whose correspondence to this method remains a design thesis (§3), not itself demonstrated. We **differentiate** this method from AI co-scientist systems [Ghareeb et al. 2026; Gottweis et al. 2026] on domain, corpus-as-verification-object, and terminal check, detailed in §2 (§7–§8).

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 those costs turn out to be a property of design rather than of the tool itself. Tankelevitch et al. locate metacognitive demand at three sites: framing a prompt, calibrating trust in an output, and deciding, on every occasion, whether to apply the tool at all: the "automation strategy" demand [Tankelevitch et al. 2024]. Passi and Vorvoreanu's synthesis names the failure that third demand produces, overreliance on an incorrect recommendation, and its most-cited remedy, cognitive forcing functions, while warning that 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 for that remedy 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 that LLM effects are "contingent rather than fixed" [Vendrell and Johnston 2026] and translating it into eight design principles derived from six cognitive, metacognitive, and epistemic processes, not from a trial: among them preserving cognitive friction and alternating AI-mediated with AI-free phases. Drosos et al. test a friction of this kind directly: AI-generated provocations induced critical and metacognitive thinking in a between-subjects study, shaped by task urgency, importance, 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 is reported shifting toward verification, integration, and a role workers call 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's close-reading experiment finds dose is what matters instead: one AI interpretation aids comprehension and enjoyment together, while heavy reliance trades comprehension for foreclosure [Zhi et al. 2026b].

Bae et al.'s thirteen-participant within-subjects study adds the reader as a further bound: the same unscaffolded LLM chat assistance raised critical-thinking scores for experienced student researchers and lowered them, with high individual variance, for novices [Bae et al. 2026]. Within novices, the study's own quantitative separator between improvers and decliners was sustained focused reading of the source text; self-initiated verification of the model's output surfaces only in interview data, as a disposition distinguishing experienced from novice readers and reported as divergent within the novice group itself, a result its authors, in a CHI EA short paper, call "indicative rather than definitive" [Bae et al. 2026]. Bo, Wan, and Anderson are this literature's sharpest internal critic and 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.

A further hazard sits under any credibility display: added transparency does not reliably calibrate trust. Kizilcec finds a bell-shaped relation between transparency and trust after an unexpected outcome, where raw data on top of an explanation reopens a gap the explanation had just closed [Kizilcec 2016], and 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]. Neither licenses a calibration claim for a display merely designed against, rather than tested for, that decoupling.

This literature builds and tests interventions at the interaction: a forcing function added to one task, a friction principle for one course, a dosage finding for one exercise. None asks what changes when verification is not an addition a user can decline on a given occasion, but a standing property of an architecture that a step either satisfies or does not.

2.2 AI-for-research systems

Three *Nature*-published systems now anchor the state of the art for AI-assisted research; we compare against the two nearest this paper's own structure. Google's Co-Scientist orchestrates six agents around a "scientist-in-the-loop" paradigm [Gottweis et al. 2026], screening self-generated hypotheses by an Elo-scored self-play tournament before a wet-lab terminal check; FutureHouse's Robin closes that loop end to end, ranking candidates by LLM-judged pairwise comparison and proposing, guiding, and analyzing dry-AMD experiments in a cycle that recovered a genuine drug candidate [Ghareeb et al. 2026]. We adapt a narrower, single-round instance of that selection logic: three independently drafted candidates for this paper's own scope were scored under three adversarial reader lenses, and the survivor was repaired against its sharpest critique and grafted with rival material, in place of their generative pipeline. Both systems concede what remains undone: Co-Scientist names, as future work rather than a capability, "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 a manual, one-time evaluation, not a standing gate [Ghareeb et al. 2026]. Stanford's Virtual Lab, the third, pairs a principal-investigator agent with specialist and critic agents under human-written meeting agendas and wet-lab-validated the nanobody binders that pipeline designed [Swanson et al. 2025]: architecturally the closest of the three to our own moderator/sub-agent/ratification-point pattern, though its object remains open-web science, not a closed humanistic corpus. A fourth system, lacking any *Nature* imprimatur, runs a comparable terminal check: Sakana AI's AI Scientist screens generated papers through its own simulated review [Lu et al. 2024], and its successor adds a self-review agent stage — one of three submitted manuscripts scored above the average human acceptance threshold, the first fully AI-generated paper to navigate a workshop-tier peer review [Yamada et al. 2025]. That self-review-then-human-review check is nearer in kind to the AI-review stage below (§4) than any *Nature*-published system's own.

SPIRE brings a cousin of that architecture to the humanities: seven agents realizing Unsworth's Scholarly Primitives discover and annotate evidence, and — unlike the division of labor we describe later — themselves perform citation binding and argumentative synthesis over classical Chinese and Latin corpora [Pan et al. 2026]. Its verification is generation-time and self-directed: a citation tag is enforced by the model writing the essay, and its evidence-recall metric scores that essay against a gold benchmark once, not as a check a reader could re-run against a delivered claim afterward. SPIRE's own account of the human's remaining part is that "scholarly judgement, teaching responsibility, and peer evaluation remain with human researchers" [Pan et al. 2026]. This is addressed to the scholar producing an argument, not a mechanism for the reader checking one already delivered. Fang et al. embed simulated peer agents into the reading interface, measurably shifting critical-thinking behavior [Fang et al. 2026]; the addition is dialogue, and it leaves what the reader has offloaded unchecked. Creative Reading raises the same delegation risk from the reader's own side: reading-augmentation tools that implicitly frame reading as information transmission — "reading to discard" — delegate interpretation to the machine, against which it proposes scaffolding that instead preserves a

reader's own interpretive work [Liu et al. 2026]. Its target is what reading augmentation should preserve; ours is a mechanism for checking what has already been offloaded. These aims complement each other instead of competing. Zhou et al.'s three-page poster proposes a framework that runs the inverse of our direction: it embeds close reading into every stage of a retrieval-augmented-generation pipeline to improve the pipeline's own output, rather than using retrieval to support a human's close reading [Zhou et al. 2025]. Berry names the nearest conceptual target without building toward it: time-boxed sessions structured around cognitive delegation, productive augmentation, and cognitive overhead, in which the human retains the "critical reflexivity essential to humanistic inquiry" [Berry 2025]: a method essay, with no system, load theory, or evaluation attached. Klein et al. state the same division from the discipline these systems draw on least: "Models make words, but people make meaning" [Klein et al. 2025]: a position essay in the same register.

A neighboring NLP literature runs a terminal check of this general kind, evaluated since ALCE established short-span citation attribution as this family's own benchmark standard [Gao et al. 2023b]: RARR retrieves evidence and 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 trains a model to retrieve on demand and critique its own generation against what it retrieves [Asai et al. 2024]; SciFact verifies a scientific claim as supported, refuted, or unsupported against a held evidence corpus [Wadden et al. 2020]. Each scores a model's own generation against a corpus once, at generation time; none embeds that check in a governance process over a closed, pre-manifested corpus with a disclosed false-positive taxonomy — not even a recent, human-directed, re-runnable citation-verification tool, which checks against an open, unbounded citation graph rather than a closed corpus [Haan 2025] — and none treats what a reader does with a delivered claim, rather than what a model produced to deliver it, as the object of a standing, re-runnable check. 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, that what a checker falsely flags, not what it catches, is what deployment turns on. The gate below borrows a pattern mature in software delivery and systematic-review methodology: mechanical check, severity-graded reopen-on-failure, human sign-off, applied not there but to a closed humanistic corpus at claim-and-quotation granularity, with dependency ordering and the disclosed taxonomy above.

2.3 Positioning

We differ from both bodies of work along three axes, stated at their true scope rather than as an unqualified improvement: domain, the corpus's role, and the terminal check. Domain: humanistic reading, where no terminal check exists for what a reader does with a delivered claim, as against what a model produces while annotating, on which at least one system has been evaluated [Galka and Vogeler 2026]. The corpus's role: a closed, manifested object each claim is checkable against, not fuel an agent retrieves from to generate — of the claim-ledger rows resting on sources this paper itself cites, 73 of 92 (79.3%) are linked to a quotation carrying a recorded verification status, the remainder disclosed as unlinked; the larger, project-wide figure is reported in Appendix A. The terminal check: fidelity to held source text plus a human sign-off, in place of a laboratory result or a model-judged tournament. Co-Scientist's and Robin's terminal check is stronger for their object than anything we offer for ours; SPIRE's agents perform a synthesis we withhold by design.

Our contribution is an architecture giving a checkable form to the offloaded step the erosion literature above documents and the research-systems literature above does not yet design for.

3 Design Rationale

The gaps named above translate into five desiderata a verification-centered 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. We take as a design assumption, not an empirically established reliability advantage, that a check is only as strong as what it checks against, so that material must be fixed and inspectable before any offloaded step is evaluated.

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] – evidence internal challenge is not a default a system reaches alone; deliberately surfaced provocations have been shown to restore critical engagement that 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, so that challenge never waits on a user's initiative or a model's own second look.

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]. The desideratum: a class of decisions – what an argument claims, when a source may be trusted, when a draft is ready – no mechanical check and no agent verdict may resolve in the reader's place, so the decision remains his own to make and answer for.

Cognitive load theory holds that working memory's capacity to hold and relate information at once is sharply limited [Kirschner et al. 2006]: material low in element interactivity – learnable piece by piece – is tractable where the same material assimilated as one interacting whole is not [Sweller 1994]. D1's mechanical check and D4's independent review apply this reduction twice: a task scoped narrowly, then a reviewer

inheriting only its bounded output, not the undivided problem. The load is distributed across the moderator and its sub-agents; no party ever holds an unreviewed whole.

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, the steps taken, and the probability of error [Kirsh and Maglio 1994]; closing his account of extended cognition, Clark asks "in what ways designers and users of new tools and technologies might exercise due epistemic caution" while aiming at fluid incorporation [Clark 2015]. D1 and D2 answer that question at the level of one step; D3 through D5 chain those steps so no later stage rests on one never checked.

A method built this way is easier to describe by its own architecture than by report, since a designed mechanism is easier to individuate than a found one [Smart 2022]: "engineers already know a great deal about the componential structure of the mechanisms that they themselves create" [Smart 2024]. Kept at that scope – access to what a gate, ledger, or agent role was built to check, not a reader's later use of its output – this warrants §4 and §5 describing the method by its components, not its effects.

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, exercised at different points of a single meta-project, not two designs sharing a vocabulary by coincidence. This is what the architecture is built to satisfy, not that any occasion already satisfies it; future research should test whether offloading of this kind shifts reliance behavior as the appropriate-reliance and cognitive-forcing-function literature predicts.

4 Method: 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 §6 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, shows that repository already in a modified, actively developed state — dated evidence the build side was already at work independently, not a static baseline the research side could alter. The two lines feed each other only through the director's own decisions and discrete authored handoff artifacts, never a direct channel between the agents: one documented instance is a structured specification, built from live inspection of the implementation and organized into ten evidence categories, left ready for transfer to the build side. Its authored-and-ready status is what the record documents; whether the parallel build agent applies a comparably disciplined process on receiving it is the build side's own account, not independently checked from the research side.

Within each line, delegation is tiered by task complexity: lightweight models handle metadata extraction and mechanical formatting, mid-tier models handle annotation and citation mining, and top-tier models handle synthesis, adjudication, and adversarial review. Every delegation is logged by role, model tier, and task specification, and every sub-agent's output passes to a distinct reviewing agent before any later stage

Table 1. 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

may build on it — answering the desideratum that offloading be checkable, not silent (D2): a downstream reader can ask, of any passage, which agent produced it, at what tier, and who reviewed it before adoption, the minimum condition Ji et al. [2023] and the human-AI interaction guidelines [Amershi et al. 2019] treat as necessary for trusting an automated contribution. The tiering converts a per-task metacognitive burden Tankelevitch et al. [2024] identify — deciding, task by task, which automation strategy the moment calls for — into a one-time architectural decision, fixed at design time. The warrant for dividing labor this way is the comparative-advantage framing stated under D5 [Engelbart 1962; Licklider 1960]: mechanical work belongs to the machine and judgment does not, a distinction the tiering preserves: 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 1. No gate is self-certified: a gate's status (open, passed, passed with conditions, reopened) is recorded as a fact about the project's history, not assumed from the fact that work continued. Figure 1 sets out the resulting two-agent, seven-gate architecture.

The pipeline is cyclical rather than sequential: a cycle closes on a full adversarial review and, where that review or a later ruling calls for it, reopens. Figure 2 diagrams the repeating unit: draft, gate review, and, short of a clean pass, a reopened gate feeding a revision round and a re-verification pass before the gate is attempted again. Two episodes from the pipeline's sibling manuscript — a humanities research paper the same architecture also produced, not the manuscript before the reader — illustrate this operating as a check rather than a formality: a reopened Gate D, thesis selection, under a second, independently commissioned sign-off after that paper's central argument changed materially post-pass, and a Gate F, terminal adversarial review, recorded as convened but *not* passed across two successive cycles, most recently on sixty-one findings, eleven severe. Appendix B gives both episodes, including the disclosed fact that the surviving Gate-F record is itself a reconstruction, assembled after the fact 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. The paper now before the reader carries its own, independent review trail through the same gates: two adversarial reviews of its earlier drafts, the first returning major revision at nine required changes, two of them severe, the second finding one severe defect discharged and the other substantially though not completely addressed, the residue unfinished reconciliation housekeeping; and the standing AI-review stage described below — in its in-project vehicle, an independent seven-dimension synthesis — returned fifty-one graded findings under its own verdict of major revision, run against this paper's fourth draft, six drafts earlier. The same discipline reaches the mechanical checks themselves: one verification check described in §5 was found defective, repaired, and re-run four times across the project's life, each repair evidenced by its own stated before-and-after count. A check never re-checked can accumulate exactly the silent drift the gate structure exists to prevent.

The governance record now requires a standing AI-review stage between the adversarial review and the director's own read, run under either of two sanctioned vehicles: an in-project multi-agent synthesis of independently-conducted review dimensions, run by agents distinct from the drafting agents but inside the pipeline's own tooling, or an author-commissioned audit by a reviewer independent of the drafting agents and of the pipeline's tooling — a different model family. Under an absolute no-edit rule the reviewer may inspect any project artifact but modify none; the stage's only output is a severity-graded findings report, never a revision — why the stage sits outside the seven-gate count rather than an eighth gate: even a clean report grants no ratification, and the stage cannot itself block the pipeline the way a gate's pass condition does. One audit under the external vehicle has been conducted, on this paper's third draft, under a governing prompt of 2,382 lines, by a reviewer from a different model family, intaken as a provenance-documented package recording checksums for the reviewed file and every preserved input; the delivered report ran to twenty-seven sections. That cross-family independence is now measured, not merely asserted: set against the pipeline's own seven-dimension review of the succeeding draft, the two finding registers overlap at a Szymkiewicz–Simpson coefficient of 0.35 (full breakdown in Table 6, Appendix B). Read one way, every blocking finding was independently surfaced by both reviews; read the other, from the external register's own denominator, the internal panel independently re-derived only about a third of what it held. Both readings are accurate: the misses concentrate in the minor tail, where agreement is lowest, while blocking findings were reproduced in full. Two limits bound this result — successive draft versions and no same-family replication — and every such finding enters the same disposition discipline as an adversarial-review finding: mechanical fix, scheduled work, or a decision routed to the director, none applied unilaterally.

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 rather than a hypothesis: three independently produced candidates for a piece of scope were scored by three adversarial reviewing lenses, with numeric scores recorded against each, and the surviving candidate was repaired against the flaws its sharpest critic identified and grafted with material its two rivals contributed rather than adopted unmodified (D4). The three lenses were three prompted personas run against a single model family, not three independently sourced models — a weaker independence bar than the AI-review stage above deliberately holds itself to, by drawing on a different model family. Recent work on LLM-judge panels finds that model-family diversity alone does not restore independent errors once judges share a prompt and a corpus: nine frontier models spanning seven families delivered only about two votes' worth of independent information, not nine [Kohli 2026], so

this panel's three scores are read here as three correlated readings from one model, not three independent votes. We credit both systems for this vocabulary while stating the honest difference: the panel here scores whole documents once, not machine-generated candidates across iterated rounds, and merges by textual graft rather than automated re-competition. Two adjacent mechanisms those systems run are deliberately not adopted, for reasons internal to this architecture rather than oversight: a self-critique ladder, where an agent reviews its own output before a human sees it, is excluded by the same non-self-certification rule that gives Gate F its bite; literature-grounded hypothesis evolution is excluded because this method's governing rule is its structural inverse — the model indexes and verifies but never originates a claim (§5).

Before any source may support a claim, it passes a source-intake pipeline (D3) that runs three checks in sequence. Metadata is verified against the document itself, never a filename or a citing paraphrase. The source is classified into a four-tier citability rule spanning peer-reviewed work, primary sources, work accepted for publication, and preprints, with everything outside those tiers held as background only. Finally, it is deduplicated at the work level, by hash, identifier, and normalized title-and-year, before registration in parallel machine-readable manifests that must never diverge from each other. The boundary this draws is a worked example in its own right: a well-known trade paperback on study technique was excluded as background-only, while an equally unfamous university-press monograph on an adjacent topic 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: a claim can only be checked against a source if the pipeline knows, with certainty, which sources it is allowed to check against.

Every substantive ruling that shapes the argument — a directive from the director, a moderator's resolution of an open question, a change of thesis, a change of scope — is entered in a dated, append-only record naming who made the decision, the rationale, and the row's current status. Across the one hundred seventeen dated entries this record now holds, every row carries that attribution field, and in seventy-six the field opens by naming the director rather than any agent or moderator role — accountability built into the log's own structure, not asserted about it after the fact. A superseded ruling is marked superseded, not deleted, so the record shows how the argument's shape changed over time, not only its current state, letting a reviewer verify that a reopened gate really was reopened on a logged ruling, not merely redrawn after the fact to match wherever the draft happened to land (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 however senior the model tier. This is 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).

5 Method: Verification Discipline

A pipeline that delegates reading, extraction, and drafting to language-model agents inherits the failure mode most documented in the hallucination literature: a model producing plausible, well-formed text that its own cited source does not contain [Ji et al. 2023]. Systems built for AI-assisted scientific discovery

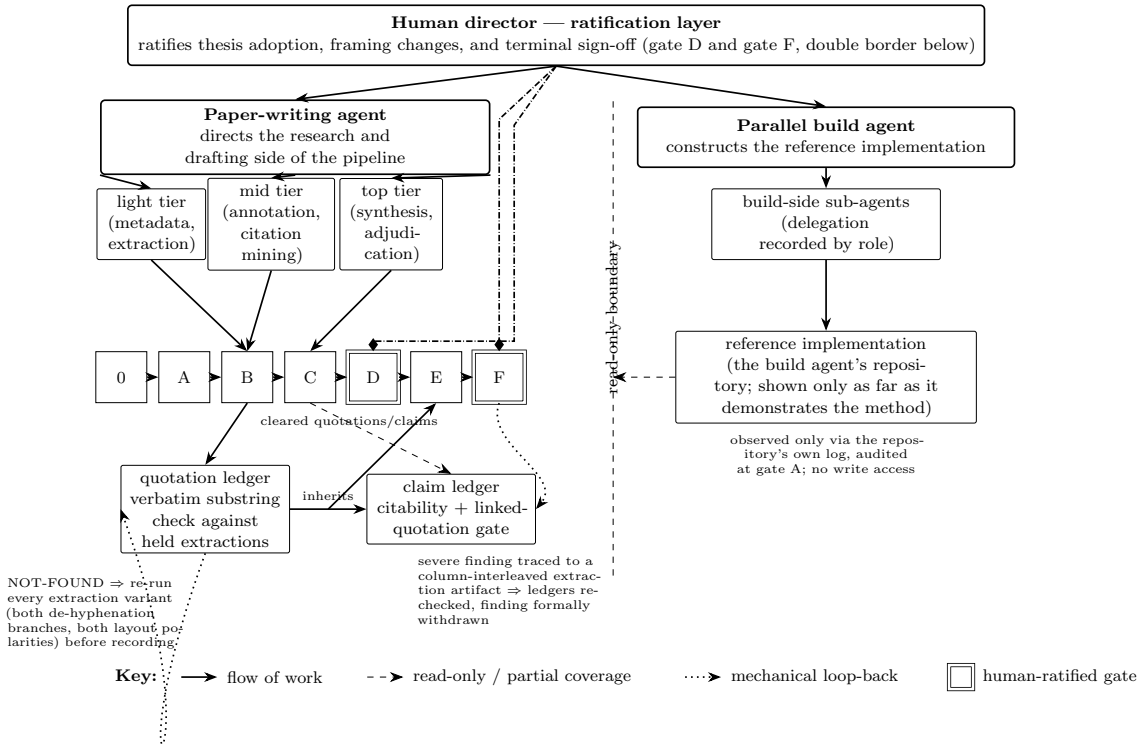

Fig. 1. The unified method's architecture. A human director (top) holds ratification authority over gate D (thesis adoption) and gate F (terminal sign-off), marked with a double border in the gate sequence. Two parallel moderating agents report to the director: a paper-writing agent, which fans work out to tiered sub-agents feeding the seven-gate sequence (gates 0 through F; each gate's checkpoint type, pass condition, and reviewing party are specified in Table 1) and the two mechanical ledgers; and the parallel build agent, which fans work out to its own sub-agents to construct the reference implementation. A dashed read-only boundary separates the two: the research side observes the reference implementation only through its own repository log, audited at gate A, and never writes to it. The quotation ledger and claim ledger sit between citation mining and the mechanical-sweeps gate; dotted arrows show two mechanical loop-backs — a not-found result triggers re-checking every extraction variant before it is recorded, and a past severe finding, traced to a column-interleaved extraction artifact, was fed back through the ledgers and formally withdrawn.

have already named a sharper version against themselves: Co-Scientist names provenance tracing back into sources only as future work (the concession quoted in §2), an admission that its "generate, debate, evolve" paradigm [Gottweis et al. 2026] does not by itself close the loop back to the literature a claim rests on. The pipeline described here inverts that priority — index, verify, gate, rather than generate, debate, evolve — with every step indexed against a held corpus, checked by machine, and gated before the next may build on it. The two mechanical ledgers and the review role below are what each gate in Figure 1 is allowed to certify, not an audit layered on afterward, and that checking is not a one-time pass: the research program has moved through five work cycles, each closing on a full ledger sweep, with Figure 2's reopening path not hypothetical here. Every project-process figure this section reports traces to the named repository record it

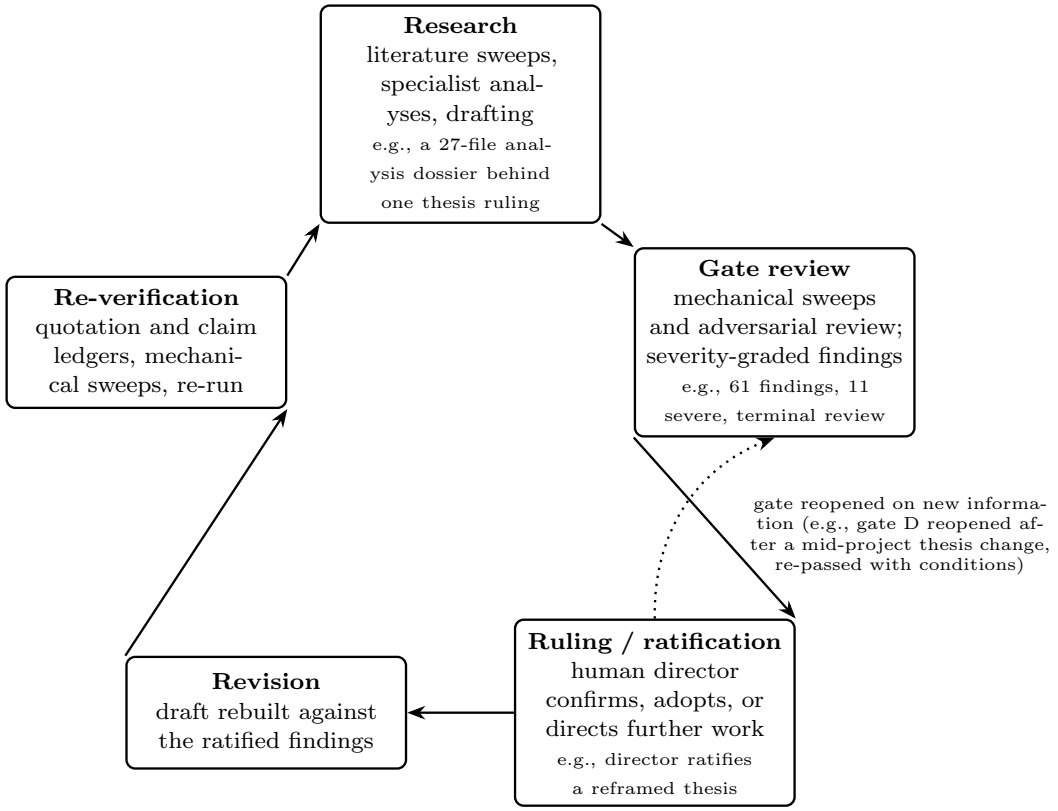

Fig. 2. The iterative cycle underlying the method. Work moves clockwise through five stages before the next research stage begins: research, gate review, ruling and ratification by the human director, revision, and re-verification. The dotted arrow shows a reopened-gate loop: a director's ruling can send work back to an earlier gate review rather than forward to revision, as happened when a mid-project change to the central argument reopened an already-passed gate and required a second, independently commissioned sign-off before drafting resumed under the new framing. The same five-stage cycle recurred across multiple project cycles rather than running exactly once.

was read from in Appendix A. The sweep tooling's own matching logic produced false failures mid-project, diagnosed and repaired four times over the project's life, two narrated below. A check that passed once is proven correct for that run only.

Three different operations share the word "check" below and are worth distinguishing once, here, instead of case by case: 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

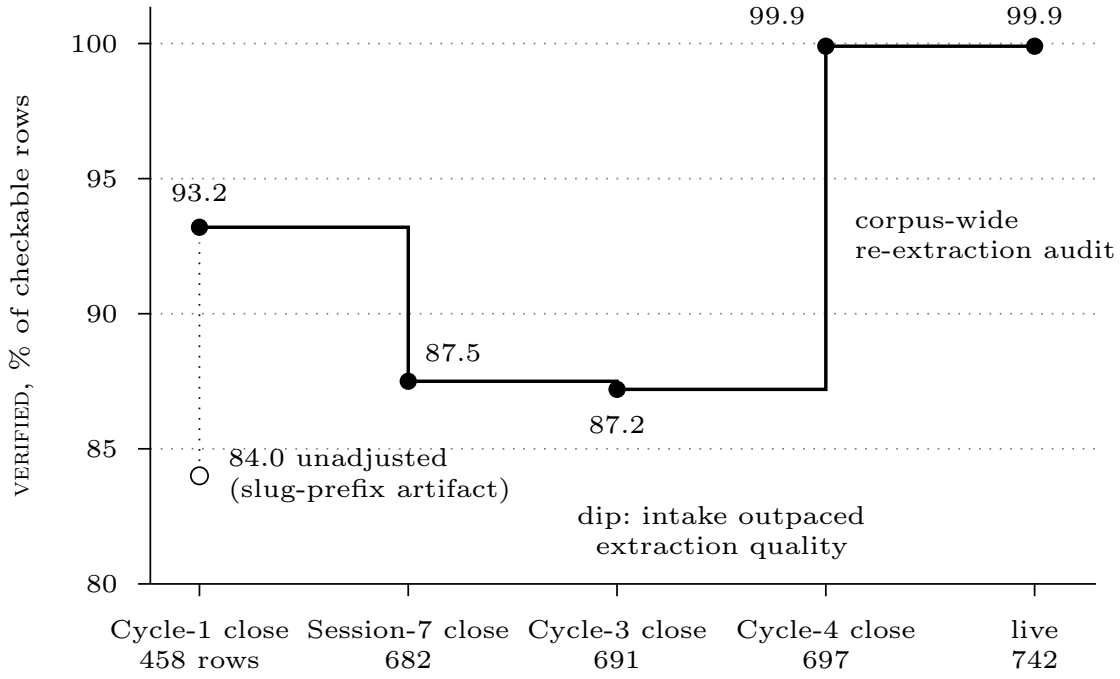

Fig. 3. Machine-verifiable quotation coverage across the demonstration corpus's committed history: the current verification machinery replayed against four committed snapshots of the quotation ledger, plus the live ledger, plotted as the VERIFIED share of checkable rows. The ledger held 458, 682, 691, 697, and 742 rows at the five points; checkable rows exclude the one row per historical snapshot graded too short to test, none live. The Session-7 point is a mid-project session close between cycles 2 and 3. The trajectory is a step function driven by identifiable repair events, not a smooth ramp: coverage dipped mid-project while source intake outpaced extraction quality, and the jump to 99.9% is the corpus-wide re-extraction audit. The Cycle-1 point is adjusted for a disclosed naming artifact of replaying today's machinery on a record written under an older naming convention — six acquisition-batch slug prefixes the current resolver does not expect; the unadjusted 84.0% is plotted as the open marker. Rates, not raw counts, carry the comparison over this growing ledger. The live point is a snapshot at the build date and decays with continued intake. One caution the replay itself forces: the ledger's own recorded verification columns were not monotone — an intermediate schema revision dropped them entirely — so the plot describes the machine-checkable state of the corpus, not the record's bookkeeping.

are. What follows reports each ledger's own outcome data at the point that datum does its argumentative work, instead of collecting it into a separate results section.

The first of two mechanical ledgers targets quotation directly, answering the exact hallucination pattern above: a plausible quoted sentence the cited source never contains [Ji et al. 2023]. It is a check run against a record outside the model's own assertion (D1), its result visible pass or fail, never a silent adoption (D2). Every candidate direct quotation is checked by script as a verbatim substring against the source's own extracted text before any claim resting on it may be used; a failing quotation is withheld outright, not used provisionally, and the check works as designed: repeatable by a third party against the same held text, not a self-report of confidence. Across the research program's 174 held sources (as of this build), the ledger carries on the order of seven hundred quotation records. A separate, author-conducted read-back pass re-checked 1,347 quotation and claim rows against source, each closed with a stated basis rather than a bare checkbox,

rejecting exactly one — a rate of author agreement with the ledger, not an independent third-party audit or a mechanical result, reported here as a count rather than a percentage. A separate audit re-ran the same verbatim-substring logic against every stored extraction variant across the source corpus: column-interleaved extraction defects, present in thirty-three of the then-manifested hundred and twenty sources, accounted for eighty-six false-negative checks, while the number of checked quotations found to assert text their source does not contain was zero.

The rejected row is also the pipeline's most informative failure, because it 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 gate review carried the finding forward 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 from both columns into the string the checker compared against. The quoted text had sat in the source the whole time; the checker had never been shown it correctly assembled. The finding was withdrawn, and the withdrawal became a standing rule: no not-found result is recorded until re-run against every plausible extraction variant: both de-hyphenation branches, since stripping a line-terminal hyphen fuses real compounds ("sub-personal" becomes "subpersonal") while keeping every hyphen fractures others, and both layout-preserving and reflowed settings, since some sources verify under only one. Not the last mechanical fault the check turned up in itself.

A second fault surfaced later, in a different part of the same check, found only because the check was re-run at a later cycle boundary rather than trusted from an earlier pass. Before comparing a quoted string to a source, the check must first resolve which held source a citation key names, by trying a fixed sequence of candidate matches — its own instructions call this the witness ladder. For one entry type, the ladder's first rung tried the cited work's title against the name of the *proceedings* it appeared in — guaranteed to fail, since no held source is titled after the volume that contains it — and stopped there instead of falling through to the citation key or the source's own title, so every citation of that shape came back indistinguishable from a genuinely broken one. Diagnosed, the witness ladder was rewritten to keep trying candidates after an empty comparison, and the fix entered the check's own standing instructions, not just that run's output: one batch it touched moved from fifty-three verified, twelve unmatched, one failure, to sixty-six verified and zero failures, all thirteen prior failures traced to the ladder fault, not any citation.[1] Running the check again at a cycle boundary — Figure 2's reopening path — is what caught a fault a single run could not surface.

This shape of problem is not new: an automated integrity tool needing scrutiny of its own is the structure of the statcheck debate in meta-science, where a script built to re-derive reported p-values and flag inconsistencies at scale [Nuijten et al. 2016] was shown, because its parser cannot represent a legitimately corrected p-value, to flag compliant papers as inconsistent while certifying papers that omit a needed correction as correct — manufacturing both false-positive and false-negative findings at once [Schmidt 2016]. A dedicated search of the AI-research-integrity and document-parsing literatures did not locate the specific chain here — extraction and matching artifacts producing false severe findings, diagnosed and withdrawn twice, each time converted into a standing rule the pipeline's later cycles enforce against itself — credited to statcheck as the nearest lineage.

[1]That before/after is not an artifact of the run that recorded it: re-implementing both the pre-repair and repaired ladders from the repair note's specification — a re-implementation from spec, not a re-execution of the original harness — and running both over the full quotation ledger reproduces the fix at scale: sixty-nine spurious no-key failures under the old ladder, zero under the new, every flipped row landing verified, no row outside the repair's documented scope changing at all.

What the quotation check can and cannot catch has itself been measured, and its blind spot is structural, so it goes first: a substring check verifies presence, not completeness. It answers whether the quoted words appear in the source in that order, not whether qualifying words were silently dropped, so a truncated quotation passes as verified. Deterministic fault injection over fifty-one verified quotation rows found the check caught every meaning-altering corruption tested: thirteen of thirteen word transpositions, thirteen of thirteen polarity flips, twelve of twelve inserted clauses, and zero of thirteen truncations. The ledger's elision-marking convention and the twenty-five-word ceiling on direct quotation exist to police by hand exactly the cut points the script cannot. That blind spot compounds with a risk this paper has not tested and names as an open threat for future work, not a closed one: if confidence in a VERIFIED badge grows faster than the badge's demonstrated reliability against this failure mode, the badge could leave a reader calibrated worse than none at all, since a meaning-altering truncation is exactly the corruption the check is proven not to catch.

A verbatim-substring check cannot catch quotation hallucination's harder sibling: a paraphrase drifting past what its source supports, with no quoted string for a script to compare. A second, parallel 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 verification status — admitting a claim only once both are met. The ledger carries 688 claim records (as of this build); 506 (about 74%) resolve by inheriting a linked quotation's verification status. Of those, 490 resolve under a plain single-quotation join — 473 (68.8%) inheriting from quotation rows marked verified by the full witness-ladder check, 17 inheriting through at least one row marked text-verified (a text-match check, not the full witness-ladder verification) — while the remaining 16 carry a semicolon-separated multi-value quotation link the plain join cannot resolve by design, a disclosed field-format defect rather than one folded into the headline count. Forty-one of the 506 additionally carry an independent claim-level verification stamp; the remaining 182 claim records carry no verbatim quotation and are flagged outside the mechanical check, a disclosed scope limit, not one absorbed into a summary number.

Both ledgers feed a structurally separate review role (D4), held by an agent distinct from whoever produced the work under review and instructed to find defects, not confirm quality — the design answer to a question users of multi-agent research tools have raised about whether such a tool will ever push back or only agree [Adavi et al. 2026]. Its output is a severity-graded findings list, every finding keyed to exact file-and-line evidence, closed with an explicit verdict: pass, revise, major revision, or sign-off granted or denied. The meta-project runs this same gate machinery over two manuscripts sharing one reference implementation: a humanities research paper and the paper here. One gate review of that sibling manuscript, convened over its full draft, returned major revision across five argument dimensions, sixty-one findings, eleven severe — nine fixed, one left open (a missing front-matter item), one only partially addressed, and the gate is recorded as convened but not passed, sign-off withheld pending that manuscript's own director read-through. That review's record was itself written after the fact: reconstructed from working logs once the contemporaneous artifact the process should have produced was found never to have been committed to the repository. We disclose this rather than leave a reader to find it by tracing the citation. It is evidence the gate architecture has run under real stakes and caught and fixed real defects, not evidence that the paper before you has passed that review. This paper's own review trail: a first adversarial pass returned major revision (nine required changes, two severe); a focused follow-up re-checked those two severe fixes against primary source

text, discharged one outright, found the other substantially, though not completely, addressed, and returned another-pass-needed; a subsequent full-manuscript audit, synthesizing seven independently-conducted review dimensions, returned major revision: fifty-one graded findings, two blocking. Oversight of this kind is built to be more than the false security a purely procedural step can provide [Passi and Vorvoreanu 2022]. A pipeline that only ever returns "pass" is hard to distinguish from one that never checked; the gate review sits inside the same reopening loop as the ledgers, in Figure 1 and Figure 2 alike, built to prefer a defensible negative verdict every cycle. The checking mechanism itself — not merely the material it checks — was caught and repaired four times over the project's life; the column-interleaved extraction defect and the witness-ladder fallthrough narrated above are two of the four, the other two resolved earlier in the same repair trail. Table 5 (Appendix B) tabulates all four episodes: thirty-five false-failure verdicts retired, every one traced to the checker and none to a quotation misrepresenting its source, alongside the one weakness that survives all four repairs — a surname-and-year fallback rung that could still resolve an unmanifested conference paper against the wrong held work, unexercised on the current corpus but disclosed rather than closed. The same record admits a less comfortable reading than rigor: a checking mechanism corrected four times also fits the profile of an immature, still-drifting apparatus, and four caught faults are no guarantee against a fifth. What answers that reading is not the count of repairs but their shape: four distinct, independently diagnosed causes, each entered into the check's own standing instructions rather than patched once and left to recur — an answer to the worry that the apparatus corrects itself only in appearance, though it does not rule out an undiscovered fifth fault, and the record makes no claim that it does. Replaying the current checking machinery against the ledger's committed snapshots extends that record to project scale: all thirty-one checkable failures at the first cycle's committed snapshot, eighty-four of eighty-five at the mid-project session's, and eighty-six of eighty-eight at the third cycle's verify today against the repaired pipeline and re-extracted corpus, none proved to be a citation defect. The survivors: one quotation straddling a printed page break, which a single-pass substring test cannot span — a documented checker limitation the ledger records as such, failing identically at every snapshot — and, at one snapshot only, a malformed committed ledger row whose underlying quotation verifies. That retirement decomposes: of the forty-four failures the first cycle recorded, thirteen clear under today's checker against that cycle's unchanged corpus, a machinery gain, while the other thirty-one needed the later corpus-wide re-extraction, a witness gain — why the extractor-polarity and witness-ladder repairs are this section's most distinctive contribution.

6 The Method in Action: A Meta-Project

The method described above is not offered as an abstraction. The project that produced this paper ran it: one paper-writing agent dispatching parallel sub-agent analyses to build the argument above, and one parallel build agent that, across the cycles that built it, directed the construction of a reference implementation the research program studies as it currently stands, not as a finished artifact, both under the same human director. Neither line proceeds in isolation; findings cross between them, and the research line checks what crosses against the build line's own record rather than taking it on report – a check run at each cycle boundary and running one way, not a claim that the build line audits the research line's record in return. What follows traces one such cycle end to end, because the cycle is itself the demonstration this paper offers in place of a deployment study: the project this paper reports ran the method on the method's own subject matter.

The cycle begins on the research side, where the appropriate-reliance and cognitive-load literature converged on a design requirement: an unfamiliar source needs a background judgment assembled from independent signals, not a single confidence number a reader has no way to weigh [Kizilcec 2016; Pareek et al. 2024]. That finding crossed to the build line as a binding specification and returned as a credibility structure separating publication rigor, creator expertise, host provenance, evidence strength, relevance, and pedagogical value into six independently displayed dimensions, with popularity held apart as a reported fact that never moves a score. That panel inherits the same hazard automation-bias research documents for any credibility display: an explanation or verdict, once shown, raises reliance whether or not it is correct, and stacking discrete signals beside a score does not by itself escape that decoupling [Kizilcec 2016; Pareek et al. 2024] — a risk this panel is designed with in view, not tested against. The traffic runs the other way too: when a draft proposed describing the reading plan's personalization as progressive withdrawal of support, the claim did not enter the paper on the drafting agent's say-so — it was checked, read-only, against the build line's own source, and the check found something 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 section's central point: treating a built capability as unverified until read against its own source, mechanically and independent of who reports it, is the same discipline the quotation ledger applies to a citation, extended here from what a source says to what a system does. The method is checking its own subject: the artifact under study is held to the verification standard the research program applies to everything else it claims, by an agent distinct from the one that built it, before either line may assert the capability as fact. The build line runs a parallel mechanical discipline of its own: a battery of nine cross-entity consistency checks, each independently tested apart from the system applying them, and a full test-suite run reported by category rather than a single pass rate. By the build line's own account, one such run covered four hundred fifty-eight cases: four hundred forty-five passed outright, eight flagged as flakes or intentional skips, and five failed, each individually annotated with its own cause. Neither line takes the other's report on faith as a rule; this run, though, is the build line's self-report, not yet independently re-run by the research line, and stands as exactly that until it is.

The artifact enters this account only as far as that discipline lets it, and only where it aids the research it supports. A retrieval layer runs over a corpus the reader has supplied and licensed, eliminating what a query returns down to what is relevant; fabricated citations and factual claims lacking a grounded quotation are excluded before either line may treat them as evidence, though the generated prose surrounding a grounded citation is not itself checked sentence by sentence against what the citation says [Ji et al. 2023]. A discovery layer surfaces material a reader's own citations never named, across ten relation categories and twelve query lanes, phrased as discovery of reception and secondary literature, not a completed influence map. Each capability is described here exactly as it entered the research record: as a check the build line's own source passed, not a feature demonstrated in use.

One such check is worth walking through. 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 ("Defines akrasia") and explanation; a candidate whose claimed quote does not occur in the block at all, by contrast, 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: the interface that displays it 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. This card carries the same automation-bias exposure noted above for the credibility structure, answered the same way: under D5's reservation of interpretive judgment for the reader [Engelbart 1962; Licklider 1960], an annotation's supporting evidence stays contestable at the point of contact, not merely inspectable after the fact. Whether a reader in practice uses that contest, or the card's transparency invites the deference Kizilcec and Pareek document elsewhere, is left open; the design commits to reserving the decision for the reader, without showing that readers exercise it.

The discovery layer is also where the method's stake in an understanding-versus-information distinction becomes concrete. A surfaced connection is not asserted and left; it carries the supporting quotation and the relation category that grounds it, so adopting the finding means relating a new claim to a source already held, not accepting an unexplained assertion. Being told that legs and hypotenuse satisfy $a^2 + b^2 = c^2$ delivers a fact; being shown the relation that makes it true delivers the structure experts, not novices, use to organize a class of problems [Chi et al. 1981], and presenting material is not what changes what a recipient can subsequently do with it: explaining it is [Kirschner et al. 2006]. This is the same requirement the moderators impose on each other: a finding crossing from one line to the other is adopted only once it carries its own grounds. Cognitive-psychology names that discipline self-explanation — stating, in one's own words, why a step follows from what came before, not simply re-reading it — and treats it as the difference between restudy and comprehension [Dunlosky et al. 2013]. A capability report absorbed without that step is information passed along; one re-verified against its source and adopted with its grounds attached is the method doing, on itself, what its own literature says distinguishes understanding from being told.

The cycle's footprint across the sibling manuscript, a humanities research paper, is part of the record, and it ran at real scale against dated, committed checkpoints: Table 4 in Appendix B tabulates the corpus and manuscript at each committed cycle close, with two cells left visibly in conflict rather than reconciled after the fact. Figure 3 plots the trajectory that scale produced: replaying the current verification machinery against four committed snapshots of the quotation ledger shows coverage moving from an adjusted 93.2% of checkable rows at the first cycle's close — the slug-naming adjustment is disclosed in the figure's caption — through a mid-project dip to 87.2% as intake outpaced extraction quality, to 99.9% after the corpus-wide re-extraction audit: a step function driven by identifiable repair events, not a smooth ramp. Nor was the record itself monotone: one interim schema revision dropped the ledger's verification columns entirely before a later cycle reinstated them, so the machine-checkable state improved across the project while the recorded state, for a stretch, regressed.

The paper before the reader shows the same signature at draft scale. Where the same review instrument was applied to it twice, severity fell within one round: the first adversarial review raised two severe findings and one major; the follow-up raised zero severe or major, only three moderate and one minor. The mechanical sweep battery's hard-failure count fell to zero by the second draft and held through the fifth — one failure

raised by a newly added sweep was fixed within the same pass — even as the battery grew from seven sweeps to ten and its statistics audit from nineteen to thirty checked figures. Four qualifications travel with that trajectory. A later, deliberately exhaustive review instrument returned fifty-one findings against this paper; that jump measures the instrument change, not draft decay, and the comparable top-severity band held at two. The zero hold did not survive the wave of test-derived material that followed: its additions broke the typography and figure sweeps, and counts re-derived from still-growing records went stale between passes — failures the battery again raised against the paper's own newest material. A moderate-and-minor tail refreshes at every round rather than converging, and three of the four new findings the follow-up adversarial review raised were introduced by the preceding fix cycle's own edits.

The project this paper reports is, on this evidence, a meta-project: it applied the method it describes to verifying the very artifact the method was built to study, and it kept, in committed form, the non-monotone record of that application.

7 Discussion

We also concede the appropriate-reliance negative result and the Co-Scientist/Robin differentiation, both already stated in §2 [Bo et al. 2025]. Our claim is architectural: a citation that resolves against the held corpus or returns not-found is a decidable check, not a soft cue confidence can talk a reader out of noticing. That decidability has a measured referent: Lau et al.'s thirteen-item Critical Thinking in AI Use scale, validated across six studies ($N = 1{,}341$), identifies Verification, alongside Motivation and Reflection, as the construct's three constitutive factors, and names "checking cited references" explicitly in the Verification factor's definition [Lau et al. 2026]. The construct is a dispositional tendency of users, and the scale's authors themselves allow that verification can be "initiated and supported by automatic processes" through cognitive offloading. The bridging step from that factor to a design property is not automatic and needs stating rather than assuming: what the architecture makes standing and low-cost is one behaviour the factor's own definition names, checking a citation against the held corpus. That is a claim about the occasion on which the behaviour is available at no cost, not a claim about any reader's disposition or score.

That gap invites a substitution objection sharp enough to run here. Lau defines Verification as the behaviour "through which users assess" AI-generated content; in this architecture a script performs the check, not the reader, so a resolved citation could as easily read as offloading that behaviour 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] – the mechanical check settles only whether a citation resolves, never whether the reader should believe what it supports. The reader side of that reservation is a designed action, not an assumption: every kept annotation carries Verify, Dispute, and Reject controls on its card (§6), so accepting a check's result is itself something a reader does, open to contest, not a default he falls into because the check ran. Bae et al.'s self-initiated-verification marker (§2) is that same factor's hedged behavioral referent, not evidence that this architecture moves anyone's score [Bae et al. 2026].

A further objection deserves a direct answer: why should a verification pipeline auditing its own research process count as an HCI contribution, when no reader outside the project appears in the evidence? It counts as the mechanism-level evidence a systems-and-infrastructure contribution offers in place of interaction data – a primary-contribution type HCI venue guidance recognizes. The two lines this paper reports, one

building the reference implementation and one holding every claim about it to the same ledger discipline the implementation applies to sources, are one architecture instanced twice; the self-application is that instancing made visible, not a substitute for a reader study.

The architecture also carries a precondition: it concentrates expert judgment at the pipeline's human ratification points and at the reader's own contest actions on generated annotations, and that judgment stays required throughout. A reader with no competency to exercise there is left with a checkable record whose use still depends on the checking a competent reader would do.

Three further limits belong here, not in a footnote. Paraphrased material inherits its claim-ledger status only where the ledger's inheritance rule reaches it; drift outside that reach goes unchecked. No controlled comparison against an answer-delivering baseline exists or is claimed here. And the strongest causal anchor for the design-contingency claim in §1–§2, Bastani et al.'s guardrail result, is drawn from high-school mathematics [Bastani et al. 2025]; carrying it to humanistic reading proceeds by domain analogy, not direct evidence.

One question we deliberately do not argue is constitutive: whether an annotation the reference implementation generates can count as part of the machinery realizing a reader's own capacity, rather than a tool that capacity uses. No reader was measured here, the design contribution neither needs nor licenses an answer, and the extended-cognition literature this paper does draw on enters only where it does design work (§3).

The gate sequence in §4 did not always pass first: a thesis change forced a passed gate to reopen under an independent sign-off, and the terminal adversarial review returned major revision – sixty-one findings, eleven severe – recorded as convened but not passed until a human read the draft. That figure belongs to a different manuscript. This research program also produces a humanities research paper on the same underlying software, run under the same gates, ledgers, and moderator but arguing a different central claim; the reopened gate and the sixty-one-finding, eleven-severe review are that paper's own Cycle-4 history, not this one's. Its record is itself a reconstruction: the review artifact due at that gate was never written to the repository as the work happened, and was assembled afterward from the sub-agent logs the work left behind. This paper's own review trail, separate and independently numbered, is given in full in §4 (adversarial passes, one discharge, one partial finding) and named again below (the external audit's fifty-one findings); the disposition record that follows picks up where that trail leaves off.

That disposition record convicts itself. The round's headline discharge figure – forty-two of fifty-one findings closed – proved unverifiable: no file the project holds re-derives it. One task-ledger row asserting no findings left outstanding is contradicted on at least four findings by the project's own decision memo – a live instance of the overclaim class these reviews keep catching. The verifiable floor is severity-weighted: both top-severity findings closed and independently re-derived; of the eleven at the next severity, ten closed, one partial. And remediation reliably lengthened the manuscript, a later trimming pass recovering only part of the overage. These negatives are reportable at all only because the closure figures were themselves audited.

The ledgers and decision record in §5 show the same discipline: a severe integrity finding this machinery raised against itself was diagnosed as an extraction artifact and formally withdrawn, not dropped.

A skeptical reading of that record deserves an explicit answer: a near-zero rejection rate could, on its own, mean the check rarely fires rather than that it always succeeds. Four instances answer that reading with what the record kept, not only what it reports. Re-running the citation-resolution check after diagnosing its own witness-ladder fault moved thirteen prior failures to verified against the identical spans, each traced to

the checker's own logic, not to a citation. The severe finding above was withdrawn rather than dropped once diagnosed. A further quotation flagged as a near-miss was held at that status through an author-audit pass rather than silently upgraded, closing out only once an independent re-extraction confirmed it verbatim. And four source files – a clipped scan, an OCR character substitution inside a load-bearing passage, a garbled byline, a missing text layer – were each caught and replaced before reaching a cited passage. The correcting reaches the record-keeping apparatus itself, not only its citations: a later pass over the project's own task log found entries carrying fabricated descriptions from an automated summarization step – twelve by the correcting task's own count, fourteen by a companion log entry, a discrepancy the correction never resolved – and logged that finding as a new, dated correction. Whether correction is dated rather than silent is itself a measured rate here, not an assumption: a cell-level census of the two research ledgers and the source manifest, over their full committed history, matched 94.4% of 5,520 historical cell changes to a dated record (58.8% to exact per-cell accounting), with the failure concentrated in one commit predating the project's records tooling – 293 cells rewritten and all nine historical row deletions, on 2026-07-23, unrecorded in any dated log – and a documented rate of at least 99.6% for every commit after that tooling was installed. The census found that counterexample before any reviewer could. The same instances support a second, less flattering reading: a mechanism that keeps needing repair — a witness-ladder fault, a near-miss held rather than upgraded, four defective source files, an unresolved task-log discrepancy — could as easily read as an immature, still-drifting mechanism as a self-correcting one. That reading does not survive the record's other half: a system that never surfaces its own faults would be more suspicious, not more reassuring, and each fault above was caught, dated, and logged by the same machinery it exposes, not found afterward by an outside reader. Immaturity that goes uncaught looks identical, from outside, to a system with nothing left to catch; this record is of one that keeps finding its own.

A check that rarely fires is not the more economical reading of a record with this many fired, diagnosed, and separately logged corrections. None of it is third-party verification: every check here was designed, run, and mostly reviewed by agents inside the project it audits. The same discipline now reaches the reviewers themselves: the standing AI-review stage puts the manuscript the pipeline's own adversarial reviewer already cleared in front of a reader from a different model family, outside the pipeline's tooling.

8 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 §6 demonstrates. Each line answers to the same gates, ledgers, and human ratification points in §4–§5, 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 accordingly 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: wherever a corpus can be held, a check can be made decidable, and a human is willing to be the point where judgment, not confidence, closes the loop, the index–verify–gate pattern is available, in principle, to another human-directed, corpus-bounded research program.

Disclosure

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 in §4–§5. 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.

On accessibility (AI-04): this PDF declares its language, and every figure in this paper carries descriptive alt text; none of this paper's five tables carries machine-marked header cells, for the same reason. Full PDF/UA machine tagging was investigated and found not cleanly achievable with the project's current build toolchain – `acmart`'s tagging support requires the `tagpdf` package, which this project's TeX distribution does not have installed – and is deferred to a future revision rather than attempted as an unsupported patch.

Acknowledgments

The generative-AI disclosure required by the ACM Policy on Authorship is given in full in the Disclosure section immediately above.

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

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

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

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

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

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

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

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

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

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

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

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

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

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 §5 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 §4–§6 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 working tree descended from repository commit `2fdd369` (commit hash; this file itself carries uncommitted edits against that commit at snapshot time, so the commit that first contains these exact figures is that commit's immediate successor on the local history – check out that successor to check these figures once it lands, or its parent, 2fdd369 itself, to see the working-tree state they were derived from); the ledgers grow during active intake, so "as of this build" counters decay within hours, and a later reader must recount against the tree 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 the audit script at `tools/locator_audit.py`, which ships in the repository tree at that path and is included in the same commit that carries this manuscript revision, 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 word-budget trim pass relocated several process-evidence blocks out of the files this script originally covered (extending it to the two files that relocation created, logged in this file's own header comments), reports 40 rows, 50 pins, zero mismatches. Two prose figures are deliberately absent: the nine-judges/seven-families statistic and its two-votes corollary (§4) are a cited literature finding, not a project record, and §6's akrasia walk-through is labeled an illustration in the text itself. The cycle-close summary table and the coverage-trajectory figure that §6 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 2. Provenance of project-process figures in §5 (verification discipline), plus one paper-scoped coverage figure asserted in §3's Positioning discussion. MV = `method_verification.tex`; MP = `method_pipeline.tex`; DM = `demonstration.tex`; RW = `related_work.tex`. "Live" = re-derived from the working tree this pass.

Claim in prose	Location	Live-record value (source)	Status
Five completed work cycles, sweep-closed	MV:3	Task ledger (`agent_task_ledger.csv`, re-derived this pass via `csv.DictReader` at 230 rows total, the ledger's frozen terminal count) carries the five closed cycles' task rows; the four earlier closes are committed (`94829e5`, `0ed9252`, `b918f62`, `5fead12`) and the fifth's closing sweep is `draft_v5_sweep_report.md`; 16 `C6.*` rows (15 Complete, one `C6.CHAIN` In Progress) and 14 `C7.*` rows (11 Complete, one `C7.ARTIFACT-PACKAGE` BLOCKED-ON-B.9, two – `C7.CAP-TRIM`, `C7.FINAL-CLOSE` – In Progress) belong to the cycle in flight. This cell was re-pinned at every wave that appended to the ledger rather than assumed stable between passes; the ledger is now FROZEN for the remainder of this workflow (no further appends), so 230/16/14 is the terminal figure, not a further-movable live count	MATCH
Sweep tooling repaired four times	MV:3, 33; MP:111–116	`gate_e_skill_repair3.md` l.112 ("the fourth time"); four dated episodes, `process_data_inventory.md` §1.1	MATCH
174 held sources (as of this build)	MV:7	`source_manifest.csv` and `.json` both re-counted this pass, 174 rows/entries (CSV = JSON; prior stale values 167/168/170 re-pinned); MV:7's rendered prose was directly re-read this pass and itself now states "174 held sources," re-pinned by the C7.SOURCE-ACCURACY wave since the prior D1/POST-TRIM notes above were written	MATCH
On the order of seven hundred quotation records	MV:7	`quote_ledger.csv`: 742 rows (714 VERIFIED, 27 TEXT-VERIFIED, 1 VERIFIED_PAGE_STRADDLE)	MATCH
1,347 rows re-checked, exactly one rejected	MV:7	`verification_audit_data.md` §1 r.11: 65+10+69+51+553+599 = 1,347; 1,346 Y / 1 N	MATCH

continued on next page

Table 2 continued from previous page

Claim in prose	Location	Live-record value (source)	Status
Column-interleave in 33 of then-120 sources	MV:9	`EXTRACTION_INTEGRITY_REPORT.md` l.77; scope l.3 (all 120)	MATCH
86 false-negative checks	MV:9	Same report ll.30–31 (86 rows across 16 sources)	MATCH
Zero quotations asserting absent text	MV:9	Same report l.30 ("genuine is zero")	MATCH
Batch 53/12/1 → 66/0; all 13 failures ladder faults	MV:13	`gate_e_skill_repair3.md` §3 before/after table; l.111	MATCH
Replay: 69 spurious no-key → 0, every flip verified	MV:13 fn.	`test2_sweep6_replay.md` §3 (736-row corpus; re-implementation from spec, not re-execution); ledger since grown to 742 by intake outside the replay's scope	MATCH
Fault injection: 51 rows; swaps 13/13, flips 13/13, insertions 12/12 caught; truncations 0/13	MV:18	`test3_mutation_test.md` §2: 51 deterministic samples (every 14th VERIFIED row of the then-736-row ledger), all 51 baselines VERIFIED; per-type catch table	MATCH
Project-wide, six-cycle, cross-track inventory: 688 claims; 506 (≈74%) inherit — 473 strict, 17 broad (490 single-linked); 16 multi-linked excluded from the split; 41 also claim-stamped; 182 unlinked	MV:24	`claim_ledger.csv` joined to `quote_ledger.csv` live: 688 / 506 (73.5%) linked, of which 490 carry a single `quote_id` — 473 (68.8%) inherit from all-VERIFIED quote rows (strict), 17 through at least one TEXT-VERIFIED row (broad) — and the remaining 16 carry a semicolon-separated multi-value `quote_id`, a field-format defect the plain single-value join cannot resolve by design and that `method_verification.tex` now excludes from the split pending its own routing; 41 claim-stamped (24 VERIFIED + 17 TEXT-VERIFIED); 182 unlinked — corrects this row's prior 489-strict figure, which had folded those same 16 multi-value rows into strict. This is every claim row entered across all six project cycles and both the research and build tracks, not a figure scoped to this paper's own citations; see the next row for that narrower figure	MATCH

continued on next page

Table 2 continued from previous page

Claim in prose	Location	Live-record value (source)	Status
Paper-scoped coverage: of the claim-ledger rows resting on sources this paper itself cites, 73 of 92 (79.3%) are linked to a quotation carrying a recorded verification status	RW:88	Independently re-derived this pass (2026-08-20) by a cite-key cross-join: every \cite{} key live in main.tex plus the eight body sections (appendix files excluded) resolved against claim_ledger.csv slugs, with the same 6 documented slug/cite-key aliases related_work.tex l.68–74 states — 52 distinct cite keys, 29 with at least one claim-ledger row, 92 claim rows total, 73 resolving (via quote_ledger.csv slug+quote_id join, multi-value quote_id split on ";") to a non-blank verification_status: 73/92 = 79.3%, confirming related_work.tex's own figure exactly, not merely copied from its report	MATCH
Sibling Gate-F: 5 dimensions, 61 findings, 11 severe	MV:29; MP:83–91	gate_f_cycle4_review.md §1 (a disclosed after-the-fact reconstruction)	MATCH
9 severe fixed / 1 open / 1 partial	MV:29	Same file l.66; process_data_inventory.md §1.2	MATCH
First adversarial pass: 9 required changes, 2 severe	MV:31; MP:103–108	draft_v2_review.md (MAJOR-REVISION)	MATCH
Second pass: 1 discharged, 1 partial; another pass needed	MV:31; MP:103–108	draft_v3_review.md l.95	MATCH
AI audit: 51 graded findings, 2 blocking	MV:31; MP:110–113; DM:215–218	c5_v4_ai_review_r1.md register AI-01..AI-51 (its header's "twelve major" differs from the register's 11 S1 by one; the prose states the register-true form)	MATCH
Four repair episodes; 35 false failures retired, none a citation defect	MV:33	test5_repair_ablation.md §3: 6+3+13+13 = 35, re-derived from each repair report's own before/after table	MATCH
Snapshot replay: 31/31, 84/85, 86/88 verify today; of cycle 1's 44 recorded failures, 13 machinery gain, 31 witness gain	MV:41	test6_historical_replay.md §4–§5: survivors are one page-straddle (all snapshots) and one malformed committed row; decomposition from the §5 cross-tab	MATCH

Table 3. Provenance of project-process figures in §4 (pipeline architecture) and §6 (demonstration). DM = `demonstration.tex`; FG = `figures.tex`; AP = `appendix_process_evidence.tex` (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 (`charter_reference.md`); structural claim consistent with the governance log	MATCH
Handoff spec in ten evidence categories	MP:15–20	`graph_rebuild_prompt.md` l.45 ("exactly one of ten categories")	MATCH
Seven named gates, 0 through F	MP:50–52	Gates 0, A–F = 7 (Table 1; charter gate structure)	MATCH
Reopened Gate-D: twelve conditions, four blocking	AP:163–166	`gate_d_signoff_cycle4.md` l.17	MATCH
Governing review prompt of 2,382 lines	MP:134–136	`wc -l` on the preserved prompt = 2,382; preserved-inputs table	MATCH
Delivered report of twenty-seven sections	MP:141–142	27 numbered top-level sections (pattern count re-run this pass)	MATCH
Register overlap 0.35; blocking findings mutually surfaced; external unique yield exceeds best internal dimension	MP:146, 157–165	`test8_reviewer_diversity.md` §2c/§3: Szymkiewicz–Simpson $10.5/\min(51, 30) = 0.35$; external unique register yield 11 vs. best single internal strictly-new yield 6	MATCH
Three scope candidates, three scored lenses, one model family	MP:187–199	`SCOPE.md` §8 via `method_adaptations.md` §2.1 (numeric scores recorded)	MATCH
Three intake checks; four citability tiers	MP:202–210	Intake-skill pipeline record; tier set fixed by logged ruling (USER-24 row)	MATCH
117 dated governance entries, all attributed; 76 name the director	MP:257–260	`DECISIONS.md`, re-derived this pass by direct parse of all rows whose first pipe-delimited cell matches the date pattern (not naive grep, since free-text cells elsewhere in the row can themselves contain dates): 117 dated rows, zero blank Made-By cells, 76 whose Made-By field opens "USER." This is an honest re-check against a now-FROZEN file (no further DECISIONS.md appends occur anywhere in this workflow), not a fresh derivation under continued growth – 117/76 is the file's terminal state, confirmed to match the frozen fixed-point value stated in this task's own brief exactly, and prior 114/75, 110/72, 100/63, 97/62, and 89/59 are each true only of an earlier, still-growing moment	MATCH

continued on next page

Table 3 continued from previous page

Claim in prose	Location	Live-record value (source)	Status
Six credibility dimensions; popularity never moves a score	DM:25–30	`app_capability_verification.md` l.122 (six named dimensions; popularity assertion unit-tested)	MATCH
Five fixed stages; known items review-only	DM:34–40	Same file ll.21–36 (verdict PARTIAL, scoped to exactly this narrower wording)	MATCH
Nine consistency checks, each independently tested	DM:54–56	Same file l.111; `process_data_inventory.md` §1.5	MATCH
458 cases: 445 passed, 8 flakes/skips, 5 failed	DM:57–60	Reference app's own status file: 445 passed, 4 flaky-recovered + 4 intentional skips (= 8), 5 failed, each annotated; the build line's self-report, as the prose labels it	MATCH
Ten relation categories, twelve query lanes	DM:71–74	`app_capability_verification.md` ll.136–137, 168	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 four cycle-close commits (`94829e5`, `0ed9252`, `b918f62`, `5fead12`); pages and terminal-review cells per the close-commit messages, the project memory's phase record, and the named review artifacts	MATCH
Coverage trajectory: 93.2 / 87.5 / 87.2 / 99.9 / 99.9% of checkable rows (84.0 unadjusted at cycle 1)	DM:182–189; FG:329–333, 356–371	`test6_historical_replay.md` §2, machine replay against committed snapshots; the interim column's 682 rows re-derived this pass from commit `6657486`; the former in-text coverage table is merged into the figure, whose caption carries the denominators	MATCH
Cycle example: a 27-file analysis dossier behind one thesis ruling	FG:209	`ls` count of `research/thesis_reframe/`: 27 analysis files, listed and re-counted this pass (2026-08-20); the directory is named as the Cycle-4 thesis-reframe deliverable set in the project memory file. Replaces an untraceable "roughly sixty" — the ledger carries ˜60 only as free text and a close commit says ˜40; no countable record yields either	MATCH
Sweep battery grew 7 to 10 sweeps, statistics audit 19 to 30; hard failures held at zero through the fifth draft, one new-sweep failure fixed in-pass	DM:209–214	`test7_review_trajectory.md` §1/§5 (v1–v5 sweep-report table: 7/9/10 sweeps; Sweep-10 scope 19/21/30; v3's single new-sweep FAIL fixed same pass)	MATCH

continued on next page

Table 3 continued from previous page

Claim in prose	Location	Live-record value (source)	Status
Draft-scale severity fell within one round: 2 severe + 1 major, then 0 severe or major, 3 moderate + 1 minor; 3 of 4 follow-up findings self-inflicted	DM:195–199, 222–225	`draft_v2_review.md` and `draft_v3_review.md` severity registers, as re-tallied in `test7_review_trajectory.md` §2	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 §6 as evidence the demonstrated cycle ran at real scale – is tabulated here in full so §6 need only point to it. Table 4 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.

Table 4. 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 4). 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

§5 narrates two of the four dated repair episodes to the mechanical quote-verification sweep in full; Table 5 gives all four, including the two (episodes 1 and 2) resolved earlier in the same repair trail and referenced there only by count.

B.3 Reviewer-Diversity Corroboration

§4 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 6 gives the granular breakdown behind that coefficient.

Table 5. 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)

Table 6. 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 `research/test8_reviewer_diversity.md` §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

§4 illustrates the pipeline's cyclicality with two episodes from its sibling manuscript, a humanities research paper the same architecture also produced, not the manuscript before the reader, 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.