Adversarial Self-Correction for Computational Research on Undeciphered Corpora: A Harness–Firewall–Refutation Architecture and Its Record

Tim Walsh
Principal Researcher, DireLabs, LLC
tim@direlabs.com direlabs.com ti.mims.ms

Methods preprint v3 — 17 July 2026

Abstract

Computational study of undeciphered or otherwise ground-truth-free corpora is unusually prone to plausible-but-wrong conclusions: with no key against which to check a reading, "it comes out looking like language/a cipher/a number system" is nearly unfalsifiable, and the analyst's degrees of freedom are vast. We describe an operating architecture that couples three disciplines to blunt this: a synthetic *ground-truth harness* (every method must separate known positives from matched null/meaningless controls before it touches the real object); a *firewall* (every reported number is produced by deterministic, versioned code that records its own script, commit, inputs, and parameters—nothing is estimated or recalled); and a standing *clean-context adversarial refutation pass* in which an independent skeptic, working *without* the original analysis's context, is tasked to build the strongest case against each finding, its brief preserved in a versioned archive. We do not claim to have proven that this discipline "works": the evidence is a single program's record and there is no controlled counterfactual. What we can report, over an eleven-iteration program on the Voynich Manuscript (Beinecke MS 408), is a firewall-verified *record* in which over a dozen first-pass conclusions—several asserted by the program's *own analysis code*, and two of them *negatives*—were narrowed, withdrawn, or retracted-then-partly-restored across successive iterations of the working preprint. Two of those episodes show the discipline declining to over-read a *negative* as readily as a positive, and one shows it curbing its own correction ("retraction, not reversal"); in that recursive case, however, the adjudication came partly from new experiments, not the refutation pass alone. In the sharpest episode, engaging a concurrently-published cipher designed to imitate the manuscript, the pass reversed a first-pass result that *flattered* the program's prior conclusion—and a further pass caught a recalled, unsourced number in the retraction write-up itself. We situate the architecture against pre-registration, multiverse analysis, severe testing, adversarial collaboration, and reproducible-research practice, and offer it—and an adoption checklist—as a *disciplined default worth adopting and testing*, not a demonstrated cure. Consistent with the method it describes, this paper was itself subjected to the same refutation pass, whose brief is archived with the program.

Keywords: adversarial validation; reproducible research; severe testing; undeciphered scripts; red-teaming; researcher degrees of freedom; evidence grading.

1 Introduction

The computational study of undeciphered corpora—the Voynich Manuscript, the Indus script, Linear A, rongorongo—has a recurring pathology: confident, mutually incompatible "solutions" are produced faster than they are refuted [Reddy and Knight, 2011, Bowern and Lindemann, 2021]. The root cause is structural. Where there is no ground truth, a result of the form "under transform T the text resembles class C" cannot be checked against a key; it can only be checked against *itself*, and a sufficiently flexible analysis will nearly always find some T that makes some C look plausible. This is the "garden of forking paths" [Gelman and Loken, 2013] and the researcher-degrees-of-freedom problem [Simmons et al., 2011] in an acute form: not merely many analyses of one dataset, but many analyses of an object whose correct analysis is unknown even in principle. The field's own literature supplies the cautionary archetype: the debate over whether block-entropy statistics show the Indus script to be linguistic [Rao et al., 2009] was settled methodologically by the demonstration that those very statistics *cannot* separate linguistic from non-linguistic symbol systems as used [Sproat, 2014] —a measure that does not discriminate the hypotheses it is invoked to decide.

Our response inverts the usual order of work. Rather than seek a finding and then defend it, we make the *validation machinery* the primary object and treat every candidate finding as guilty until an adversary fails to convict it—an operational stance close in spirit to severe testing, on which a claim is corroborated only insofar as it has passed a test it would probably have failed were it false [Mayo, 2018]. This paper describes that architecture (§2), explains why each component catches a distinct class of error (§3), and —its empirical core—reports the architecture's *record*: the catalogue of the program's own conclusions that the discipline overturned before the program's final synthesis (§4). We then situate the approach against existing practice (§5), state its limitations reflexively (§6), and give an adoption checklist (§7). The case study is an eleven-iteration program on the Voynich Manuscript; the numbers we quote from it are illustrative of *corrections*, and each traces to versioned code under the firewall. We do *not* claim to have shown the architecture superior to ordinary careful iteration—we cannot, from one program (§6). We conjecture it transfers to other ambiguous, ground-truth-free corpora, and offer it as a discipline worth adopting and testing, not a result.

2 The architecture

Three disciplines, each necessary, jointly enough to make refutation bite.

2.1 The harness: validate on knowns before the unknown

No method is trusted on the real object until it behaves correctly on synthetic ground truth. For each method we assemble (i) *real positives* (period- and register-matched natural-language corpora), (ii) *matched negatives*—structured-but-meaningless generators (auto-copying/self-citation processes [Timm and Schinner, 2020], grille-style table generators [Rugg, 2004], parameterised null models) tuned to the object's own nuisance parameters, and (iii) *cipher/transform controls*. A method that cannot separate these knowns is not used on the unknown, and a *replication gate* requires the pipeline to reproduce already-published baseline statistics before any novel experiment. The harness is what makes a negative result meaningful: "the statistic does not distinguish A from B" is only informative once the statistic is shown to distinguish a synthetic A from a synthetic B. (This positive-vs-matched-negative separation is close kin to the practitioner technique known in applied machine learning as "adversarial validation"; we use it here as a gate on *method admissibility*, not on data drift.)

2.2 The firewall: numbers come only from code

Every statistic in every report is produced by deterministic, versioned code with fixed seeds, written to machine-readable output that records its producing script, commit hash, input-dataset version, and parameters. No number is estimated, remembered, or carried in prose. This is standard reproducible-research hygiene [Sandve et al., 2013], but in this architecture it is also load-bearing for *refutation*: because every claim dereferences to an exact computation, the adversary attacks the computation itself—re-running it, re-scoring it, probing its bands and nulls—rather than a summary sentence. A firewall turns "I don't find your argument convincing" into "your config `X` has its median out of band and you counted it as a match," which is checkable.

2.3 The standing clean-context refutation pass

Every finding is submitted to a skeptic instructed to build the *strongest* case against it, with positive and associational claims held to the harshest scrutiny. The pass is *clean-context*: the reviewer is given the claim, the code, and the data, but not the original analyst's narrative, hopes, or sunk cost. In our instantiation the refuter is a fresh-context large-language-model agent with read access to the repository and no memory of how the finding was reached; it returns a written brief—often several thousand words— quoting the offending numbers and recommending a grade, and **that brief is preserved in a versioned archive** (`docs/refutations/`) so that "the refuter caught it" is itself dereferenceable rather than operator recollection. A claim is graded **A** (replicated/robust) or **B** (candidate) only after it survives; otherwise it is narrowed, downgraded to **C** (directional) or **D** (speculative), or withdrawn. This is the operational descendant of red-teaming, now automated for language models [Perez et al., 2022], and of adversarial collaboration [Mellers et al., 2001], made cheap and repeatable enough to run on *every* finding rather than a chosen few.

We are deliberate about what "independent" does and does not mean here. The refuter is *context*-independent—it has no access to the analyst's narrative or investment—but it is not model- or operator-independent: it is an agent of the same model family, spawned by the same operator, over the same repository, and the operator chooses what to submit. The premise that a finding's author is poorly placed to destroy it is a *motivating hypothesis*, not something we establish. Where stakes are high, a cross-vendor or human refuter is the stronger check; the one genuinely cross-vendor model used in this program served as a data *rater*, not as a refuter, so that stronger check remains future work.

2.4 Connective tissue: grades and a decision ledger

Two lighter rules bind the disciplines. *Evidence grading* (A–D) is attached to every claim, so downgrades are first-class outcomes rather than failures. *"Flag, don't resolve"*: any decision not covered by a locked ledger becomes a surfaced open item on the least-committal path—so the forking paths are logged, not walked in secret.

3 Why it catches things

Each discipline structurally removes a distinct source of error. **The harness removes non-discriminating methods**—a statistic that cannot separate the synthetic knowns is discarded before it can produce a spurious verdict on the unknown, exactly the failure the Indus entropy debate diagnosed [Rao et al., 2009, Sproat, 2014]. **The firewall removes recall drift and gives refutation teeth**: numbers that dereference to code can be attacked; numbers that live only in prose cannot. **The clean-context refuter targets confirmation bias and sunk cost**: an agent

with no investment and no narrative is positioned to press where the analyst will not, and running it on *every* claim— feasible only because it is automated—turns adversarial review from an occasional event into a gate.

The property we most want to emphasise is that the discipline resists errors in *both* directions. The replication-crisis literature concentrates on false positives [Ioannidis, 2005, Simmons et al., 2011], for good reason. But in undeciphered-corpus work an over-strong *negative*—"no cipher of this class fits," "no generative family reproduces the signature"—is equally damaging and more insidious, because a negative silently closes a hypothesis and invites no replication. An architecture that only guards positives will let over-read negatives ship unchallenged. The record below shows the discipline declining to over-read its own negatives—softening them to "unresolved" or "under-determined" rather than asserting them—and once curbing its own *correction* of a negative. We are careful, in §4, not to inflate this into "overturns negatives as readily as positives": the negative episodes ended in non-findings, not in restored positives, which is a milder operation than the outright kills the pass delivered on over-read positives.

4 The record

The empirical core of this paper is the catalogue of the program's own conclusions that the discipline overturned or narrowed before the final synthesis. Table 1 groups the salient cases; the text then expands the most instructive, including the ones where the honest reading is less flattering than the headline. Every entry is a genuine first-pass verdict recorded in the program's versioned results. The corrected *numbers* are firewall-checkable in `results/`; the adversary's *reasoning* is archived in `docs/refutations/` for the iterations from i06 onward, and—candidly—survives only as the synthesis record for the earlier iterations, whose briefs predate the archive (§6).

Circular positives and fitted points. The two most seductive errors were self-referential. A joint-signature test first reported the manuscript "favoured as a generation process"; the only generator that matched it was the one built to imitate it, so the match was true by construction and the claim was narrowed to the robust *negative* it actually supported. One iteration later the same trap reappeared: a generator was graded a "class sufficiency" **[B]**, until it emerged that its constants had been grid-selected against the target's own bands—a fitted point dressed as an a-priori draw—and that its weak-syntax criterion used a threshold a fully order-shuffled control also passes. Recognising the *same* circularity twice is itself a product of the standing pass: the pattern was in the reviewer's checklist because it had been named once.

Mandatory nulls catch false positives that multiple-testing control does not. An anchor hunt produced eighteen "hits" that survived Benjamini–Hochberg control yet were all false: a mandatory permutation null—shuffling the feature labels and re-running the entire pipeline—returned them to chance ($p \approx 0.48$, net discoveries zero). A null that re-executes the whole analysis is a stronger instrument than a p-value correction on its output.

Negatives, honestly. Two cases show the bidirectional property—and its limits. A candidate visual regularity was first over-claimed, then *over-killed* ("rater-idiosyncratic"); the pass showed the kill's central argument was a red herring and the honest status was *unresolved*. And a whole iteration concluded "no tested generative family reproduces the joint signature"; a refutation-directed follow-up showed the obstruction was largely an artifact of a rigid morphology model, and the negative was softened. Both negatives resolved to *non-findings* ("unresolved," "under-determined"), not to restored positives—a real but milder correction than the kills delivered on over-read positives.

The recursive episode, with its skeptical reading stated. The softening above was then *itself* curbed: a preprint headline claimed the generator "reproduces all hard axes," but the program's own result file showed no single configuration matched the full set (0/48), and the headline was

walked back to "a retraction of the hard constraint, not a promotion to no constraint." The flattering reading is reflexive control. The honest reading must also admit the unflattering one: on a *single* question the program shipped three successive headlines (a strong negative, a strong dissolution, a retraction) across consecutive preprint snapshots, which a skeptic may read as instability rather than virtuosity. Two facts discipline that reading. First, what actually *adjudicated* the question was new computation—decoupling the connectivity control, then adding length variance—i.e. ordinary better science; the refutation pass's genuine contribution was narrower: it *proposed* the artifact hypothesis and it *corrected the prose* back to a stable underlying number (the 0/48 was in the result file the whole time). Second, each correction moved the *prose* toward the *firewall*, never away from it—which is the firewall doing its job against analyst headline-writing, and is the part of this episode we would actually defend. We do not claim the standing pass, by itself, resolved the question.

Corrections can also strengthen. Not every refutation downgrades: a cipher-class exclusion first shown over one plaintext language was, at the reviewer's demand for typologically diverse controls, *upgraded* to universal—though the same experiment simultaneously opened a transposition loophole that a later experiment had to close. Even the strengthening was not free of a new obligation.

The sharpest case is external-facing. Late in the program a hand-constructed, decipherable cipher designed to imitate the manuscript was published independently [Greshko, 2025]. We ran its ciphertext through our own discriminators; the first pass scored it as *confirmation* of our standing "cipher-of-real-prose is excluded" result. The refutation pass overturned that: the confirming statistic (a word-order-information measure) had collapsed not because the text was enciphered but because the cipher's authors respaced the plaintext and used heavy homophony—both of which flatten that statistic with word order intact—so the "confirmation" was an artifact, and the exclusion itself was over-broad and had to be retracted to a narrower claim. This is the pattern the architecture most wants to catch: a result that flattered the program's prior conclusion, on a live question with an external interlocutor, reversed before it was acted on. It also exercised the firewall specifically: the write-up of the retraction contained a recalled, unsourced separability figure ("$\sim 30\sigma$") that a further pass traced back to code and corrected to the actual 6.8–8.0σ—a number that existed only in prose is exactly what the firewall exists to forbid, and the discipline caught its own violation of it.

Across eleven iterations the discipline produced, alongside the graded findings, a versioned log of these overturns—conclusions asserted by the analysis, in several cases by the analysis *code*, that were narrowed or withdrawn. We regard that log as the program's most robust output; we do *not* claim it proves the discipline necessary, only that it happened and is checkable.

5 Relation to existing practice

The components are individually familiar; the contribution is their integration and the gate they jointly enforce. *Pre-registration* [Nosek et al., 2018] curbs forking paths by fixing analyses in advance, but cannot help an exploratory program on an object whose right analysis is unknown; the decision ledger and standing refutation are the exploratory analogue. *Multiverse* [Steegen et al., 2016] and *specification-curve* [Simonsohn et al., 2020] analyses report a claim across many defensible specifications; the refuter's job is precisely to find the specification that breaks the headline. *Severe testing* [Mayo, 2018] is the closest philosophical precedent: our gate is an operational attempt to admit only claims that have survived a genuine chance of refutation. *Adversarial collaboration* [Mellers et al., 2001] pairs opposed experts; the automated clean-context refuter is a scalable, sunk-cost-free instantiation, and *red-teaming with language models* [Perez et al., 2022] is its direct technical ancestor. *Reproducible-research* practice [Sandve et al., 2013] supplies the firewall; we add that reproducibility is not only an end (others can re-run) but a *means* (the adversary can attack

the computation). Finally, the harness's positive-vs-matched-negative separation echoes applied ML's "adversarial validation" (training a classifier to tell real from synthetic), repurposed as a method-admissibility gate. The novelty we claim is thus modest and specific: the *combination*—a harness-gated method set, a firewall that gives refutation teeth, and a standing, archived, clean-context refuter as a publication gate applied to negatives as strictly as to positives—not any single ingredient.

6 Limitations and reflexivity

The evidence is a *single program's record*, and we cannot present the controlling counterfactual—the same program run without the discipline—so we cannot say how many overturned claims would have been published elsewhere; we can only show they were the pipeline's genuine first-pass verdicts, corrected across iterations of a *working preprint* (two of them, indeed, were headlines of released snapshots corrected in later ones, not silent pre-report catches). We also did not track the denominator—the total number of first-pass claims, or the claims the refuter *passed* that were nonetheless wrong (the method's own false negatives); a candid "we did not measure this" is more honest than a rate we cannot support. The architecture reduces but does not eliminate error: the refuter can miss things, grades are judgment calls, and there is a real risk of *over*-refutation— killing true findings—for which the double-corrected visual-regularity case is both a cautionary instance and a demonstration that the tension is detectable. The automated refuter shares its substrate's blind spots; an adversary and analyst from the same model family may fail together, so cross-vendor or human refutation is the stronger check we have not yet run. The refutation archive itself is only partial—briefs before iteration i06 were not captured and survive only as the synthesis record, which is precisely the prose-only weakness this paper otherwise warns against. And, in keeping with the method it advocates, this paper was itself put through the pass; its brief is archived, several of its first-draft over-claims (including an over-strong abstract and an inflated "bidirectional" framing) were corrected in consequence, and we regard *that* as the appropriate demonstration, not a proof.

7 Adoption checklist

For a computational program on an ambiguous or ground-truth-free corpus:

1. **Build the harness first.** Real positives, matched structured-meaningless negatives, transform/cipher controls. Do not run a method on the object until it separates the knowns; gate novel work behind reproducing published baselines.
2. **Erect the firewall.** Every reported number from deterministic, versioned code that stamps its script, commit, inputs, and parameters. No number in prose that is not in a results file.
3. **Grade everything A–D**; make downgrades and "not established" first-class outcomes.
4. **Run a clean-context refutation pass on every A/B claim**—an independent reviewer, without the analyst's context, tasked to break it, with authority to downgrade. Hold negatives to the same standard as positives.
5. **Preserve the briefs.** Version the refutation outputs so "the refuter caught it" is checkable, not recollected.
6. **Log decisions; don't resolve silently.** Surface every uncovered choice on the least-committal path.
7. **Refute your corrections.** A walk-back is a claim; subject it to the same pass.

8. **Diversify the adversary** (cross-vendor model, or human) where stakes justify it, and don't mistake context-independence for model-independence.

8 Conclusion

Where the answer cannot be looked up, the discipline of *how* a claim is made must carry the weight that ground truth carries elsewhere. Coupling a validating harness, a firewall that makes every number attackable, and a standing, archived, clean-context adversarial pass produces a pipeline whose default output is a graded claim that has survived an attempt to destroy it—and, as the record shows, a pipeline willing to soften its own negatives to "unresolved" rather than over-read them. We do not claim to have proven this discipline necessary or sufficient; we report that it happened, that it is checkable, and that we found it productive enough to recommend adopting and testing. The transferable object is the architecture, not any reading of any manuscript.

Reproducibility

The case-study program's statistics are produced by deterministic, versioned code (fixed seeds) writing machine-readable results, each recording its producing script, commit, dataset version, and parameters; the refutation briefs from iteration i06 onward, including the one on this paper, are archived alongside. Numbers quoted here are illustrative of corrections and trace to that code.

References

Claire L. Bowern and Luke Lindemann. The linguistics of the voynich manuscript. *Annual Review of Linguistics*, 7:285–308, 2021. doi: 10.1146/annurev-linguistics-011619-030613.

Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking". *Department of Statistics, Columbia University (working paper)*, 2013.

Michael Greshko. The Naibbe cipher: a substitution cipher that encrypts Latin and Italian as Voynich manuscript-like ciphertext. *Cryptologia*, 2025. doi:10.1080/01611194.2025.2566408.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

Deborah G. Mayo. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 2018.

Barbara Mellers, Ralph Hertwig, and Daniel Kahneman. Do frequency representations eliminate conjunction effects? an exercise in adversarial collaboration. *Psychological Science*, 12(4):269–275, 2001.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of EMNLP 2022*, 2022.

Rajesh P. N. Rao, Nisha Yadav, Mayank N. Vahia, Hrishikesh Joglekar, R. Adhikari, and Iravatham Mahadevan. Entropic evidence for linguistic structure in the indus script. *Science*, 324(5931): 1165, 2009.

Sravana Reddy and Kevin Knight. What we know about the voynich manuscript. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 78–86, 2011.

Gordon Rugg. An elegant hoax? a possible solution to the voynich manuscript. *Cryptologia*, 28(1): 31–46, 2004.

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10):e1003285, 2013.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020.

Richard Sproat. A statistical comparison of written language and nonlinguistic symbol systems. *Language*, 90(2):457–481, 2014.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

Torsten Timm and Andreas Schinner. The voynich manuscript: Symbol roots, features, and a self-citation generative method, 2020. Self-citation / auto-copying generation hypothesis. details.

Failure mode	**First-pass claim**	**What caught it**	**Direction**
Circular positive	"favoured as a *generation* process" (E19)	The only matching generator was tuned to the target; match is by construction	false positive
Fitted-to-target	"the generative class is *sufficient*" [B] (E21)	Constants grid-selected to hit the target bands; weak-syntax test passed by a full shuffle	false positive
Missing null / control	18 word→feature "anchors" (E7); "needs constructed morphology" (E6); "whitening confirms" (E8)	A mandatory permutation null returned the 18 to chance ($p \approx 0.48$); a demanded abjad control (E6); Σ ill-conditioned (E8)	false positive
Broken metric / artifact	A "meaning certificate" from a word-order statistic; a trivial-endpoint effect (E9)	Two axes shown broken; a pre-registered commitment retracted	false positive
Confound	Currier A/B "differ in grammar" (E14)	Content-controlled study: the difference reverses within-section (E17)	false positive
Over-read negative	"root↔leaf KILLED" (E12); "no generative family reproduces the signature" (i08)	The kill's key argument was a red herring → *unresolved* (E12); the coupling was a morphology artifact → *softened* (i09)	negative → non-finding
Recursive (self-check)	The i09 walk-back itself: "reproduces *all* axes"	No single config matches all axes (0/48) → "retraction, not reversal" (though new experiments, not the pass alone, adjudicated)	over-correction
Demanded-control upgrade	A Latin-only cipher exclusion (E19)	Diverse-language controls *strengthened* it to universal (E19b)—while opening a transposition gap later closed by E20	strengthening
External-facing + recall error	"the cipher-of-real-prose class is excluded", then a first-pass "confirmed against a concurrently-published cipher" [Greshko, 2025]	The pass showed the confirming statistic was confounded by the cipher's respacing/homophony → retracted the whole exclusion; a further pass on the retraction write-up caught a recalled, unsourced "$\sim 30\sigma$" (true value 6.8–8.0σ)	over-read positive + firewall catch

Table 1: Selected entries from the program's self-correction record (i03–i11). Corrected numbers are firewall-checkable; refutation briefs from i06 onward are archived (`docs/refutations/`). Note the *negative* and *recursive* rows—the direction the replication-crisis toolkit usually ignores—and their honest, milder outcomes.