On Method: Verification as a Bound Adversarial Procedure

How to run AI-assisted research without it collapsing into a confirmation mill

Working Draft v1 · Methods Note · Preprint, not peer-reviewed · 2026

Matthew "Bear" Schulz · Independent Researcher · ORCID: 0009-0000-0956-5743 · mattschulz97@protonmail.com

Companion methods note to the author's *Frame-Lock / Trust in the Dark* and *Turtles* deposits. It describes a *practice* used to produce those works; it makes no claim that the practice guarantees truth, only that it lowers the rate at which a researcher and an agreeable instrument deceive each other in the same direction.

Abstract. A large language model used as a research collaborator defaults to agreement, and agreement is the one thing a verifier must not supply. When an instrument's errors correlate with the researcher's hope, the pair becomes a coherence engine: it manufactures confident, internally consistent, false pictures — the failure mode the author's substantive work studies, arriving here in the method itself. This note states a practice that treats the model as admissible only as a *bound adversarial procedure-runner*: every load-bearing question is pre-registered with an explicit kill condition before it is run, the run is reported including its nulls, and the result is logged and handed out to be broken. We give the loop, the artifacts that hold it in place, the evidence that it bites (a body of work in which roughly as many claims were killed as confirmed), and the limit it cannot pass — that a single researcher plus a single instrument is a single-source map, and independence can enter only from outside. The practice is itself an instance of the thesis it serves: the useful setting for a verifier is neither maximal agreement nor maximal obstruction, but a bound middle.

1. The problem: a collaborator that agrees is a corrupted sensor

A capable language model will, by default, help you believe what you already wanted to believe. It elaborates your hypothesis, supplies the supporting case fluently, and softens objections into "considerations." Each individual act is helpful; the aggregate is a sensor whose noise is *correlated with the researcher's prior*. That correlation is the whole danger. Two independent flawed instruments triangulate; an instrument that errs in the direction you are already leaning does not — it amplifies. The result is the same one the author's substantive work names: a picture that is seamless, confident, and wrong, made *more* convincing by its coherence because there is no second voice to disagree [1, 3].

This note does not propose making the model less capable, nor pretending it is neutral. It proposes *binding* it — constraining when and how its output is allowed to count as evidence — so that its fluency is spent on running procedures rather than on ratifying conclusions.

2. The stance: gatekeeping over agreement

The governing rule is that an agreeable instrument is a *liability* as a verifier, and is admissible only under constraints that make agreement impossible to fake. Concretely: the model holds *design authority* (it may build instruments, propose tests, and run them) but never *truth authority* (its assent is not evidence). Trust lives in the binding and the reproducible record, never in the instrument. A sound verification must sometimes cost the researcher a hypothesis they wanted to keep; a session that never does this has not verified anything, it has flattered.

3. The loop

The practice is a single loop, repeated per load-bearing claim (Figure 1). Its non-negotiable feature is that the *kill condition is fixed before the run*, in writing, so that a result cannot be reinterpreted into a success after it is seen.

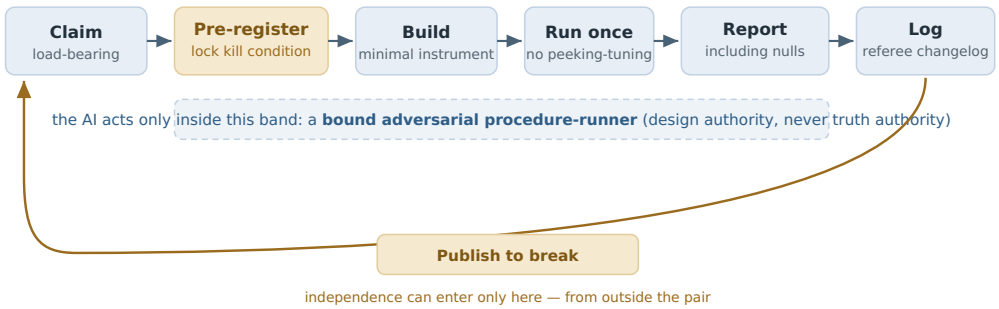

Figure 1. The loop. *Each load-bearing claim is pre-registered with a kill condition (amber) before it is built and run; the run is reported with its nulls and logged. The model operates only inside the dashed band — it may design and run, but its agreement never counts as evidence. The single step the pair cannot perform for itself is the return arrow: independent breakage, which can arrive only from outside.*

Pre-registration with a real kill condition. Before running, state the prediction, the exact measurement, and the result that

would *retire* the claim — quantitatively, with no free parameters left to adjust afterward. A kill condition that cannot fire is decoration. The pre-registration is a commitment device precisely because it is written when the answer is still unknown; producing it and running in the same breath defeats it.

Run once; report the nulls. A confirmed prediction is low-information; a killed one is high-information. The practice treats a null or a falsification as the *valuable* outcome and reports it as the headline rather than burying it — including when the failed prediction was the researcher's own.

4. The artifacts that hold the loop in place

Discipline that lives only in intention decays. The practice externalizes it into a few plain-text artifacts that travel with the work, each countering a specific failure (Table 1).

Artifact	What it is	Failure it prevents
Pre-registration	Locked prediction + kill condition, dated before the run	Reinterpreting a result into a success after seeing it (HARKing, goalpost-moving)
Coverage map	Ledger of every claim tagged TESTED / PARTIAL / ASSERTED / RETIRED	Untested scaffolding reading as demonstrated; the edifice outrunning the evidence
Referee changelog	Dated record of corrections, especially self-caught ones	Silent revision; a reviewer unable to see what was wrong before
Break-this file	Open challenge keyed to specific runnable kill conditions	A single-source map mistaken for verified knowledge
Verify-before-citing	Every empirical anchor checked against the primary source; disputes travel with the claim	Enshrining a beautiful example because it fits

5. Why the method is an instance of the thesis

The practice is not borrowed from outside the science; it is the same shape, applied to the workbench. The substantive work argues that a system holds a true-enough grip on reality only in a calibrated middle: trust everything and the lie takes you; trust nothing and you build nothing — the two failures are one broken switch flipped opposite ways. A verifier sits on exactly this dial (Figure 2). An instrument tuned to maximal agreement catches no error; one tuned to unbounded obstruction ships nothing. The usable setting is a bound adversary: free to attack the claim, constrained to do so through procedures whose verdicts are fixed in advance.

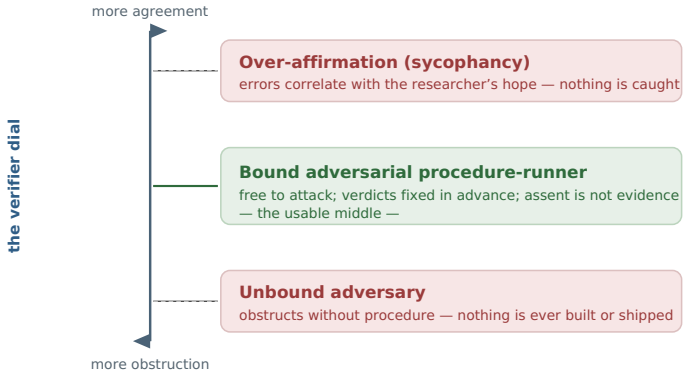

Figure 2. The verifier dial. *The same calibrated-middle structure the substantive work finds in trust itself, applied to the instrument. Maximal agreement and unbounded obstruction are one broken switch flipped opposite ways; both end the work. The usable setting is a bound adversary — the green middle. That the method recapitulates its own subject is offered as a consistency check, not as proof.*

6. Evidence the practice bites

A method that only ever confirms has become the failure it set out to avoid. The bodies of work produced under this practice are reported with a running tally of *tested* versus *retired* claims, and the retired column is not empty. Across the trust and replicator deposits, claims that were killed in code include a control-theoretic identity that was the prettiest line in its scaffold; a third "saturation" attack class that collapsed into an existing one under a pre-registered test; a "sharp threshold" that finer sampling revealed as a gradual gradient; a set of early-warning signatures that turned out to be trivial noise-scaling rather than genuine critical slowing; and the most aesthetic clause of a code-origin claim ("silence is the operator"), which failed while the substantive half of the same claim held. Roughly as many load-bearing claims were retired as were confirmed. That ratio — not the confirmations — is the evidence the loop is falsifying rather than flattering.

7. Limits — the seam the practice cannot close

The honest boundary is structural, not a matter of effort. A single researcher plus a single instrument is a *single-source map*, and every additional in-model result they generate shares a root: same author, same assumptions, same blind spots. More runs from that root buy confidence, not independence. The practice can reduce correlated error between researcher and instrument; it cannot manufacture the one thing that converts a coherent map into knowledge, which is a second map reached from different ground. That can only arrive from outside the pair — which is why the break-this file, not any internal result, is the artifact that matters most, and why the practice treats the hostile, competent outside reader as its most valuable collaborator rather than its adversary. Further limits: the instrument is bound but not infallible and can mis-run a procedure; pre-registration constrains but cannot anticipate every degree of freedom; and a kill condition is only as good as the measurement behind it (one of the deposits records an early-warning "pass" that a stricter, effect-size-aware rule would have failed — the criterion was too weak, and that, too, is logged).

8. Invitation

The practice is offered as a transferable habit, not a guarantee, and released freely (CC BY 4.0) so that anyone doing AI-assisted research can adopt the loop, copy the artifacts, or try to break the claim that binding an agreeable instrument lowers the rate of shared self-deception. The most useful response is the one the method is built to invite: take a step of it — the kill condition, the coverage ledger, the verify-before-citing rule — and show where it fails to help.

References

[1] Popper, K. R. (1959). *The Logic of Scientific Discovery.* Hutchinson. (Falsifiability as demarcation.)

[2] Merton, R. K. (1942). The Normative Structure of Science. Reprinted in *The Sociology of Science* (1973), Univ. of Chicago Press. (Organized skepticism.)

[3] Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine* 2(8):e124.

[4] Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis. *Psychological Science* 22(11):1359–1366. (Researcher degrees of freedom.)

[5] Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The Preregistration Revolution. *PNAS* 115(11):2600–2606.

[6] Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex* 49(3):609–610.

[7] Axelrod, R. (1984). *The Evolution of Cooperation.* Basic Books. (The tournament: independent outside strategies were the science, not the board.)

[8] Sharma, M., et al. (2023). Towards Understanding Sycophancy in Language Models. *arXiv*:2310.13548.

Citation note: references are standard works cited from established literature. Per this project's own rule (§4, verify-before-citing), confirm each against the original — years, volumes, pages, and the arXiv identifier of [8] — before external circulation. None is load-bearing for the practice, which stands on the reproducibility of the works it produced.

Referee Report — self-review. *Reviewer stance: skeptical but fair.*

Summary judgment. A methods note lives or dies on whether it admits its own ceiling. This one does: §7 states plainly that the practice cannot manufacture independence and names the break-this file, not any internal result, as the load-bearing artifact. The chief risk is that the note reads as self-congratulation — a practice describing how rigorous it is. The mitigation is §6: the claim is staked on the *retired* column being non-empty, which is checkable in the deposits, rather than on assertion.

Strongest / weakest. *Strongest:* the artifact table and the kill-condition discipline are concrete and copyable. *Weakest:* "roughly as many killed as confirmed" is reported as a ratio without a denominator here; v2 should tabulate the exact counts from the coverage maps so the claim is auditable rather than impressionistic. *Honest exposure:* the note is written with the same instrument it describes, so it is subject to its own warning — only an outside reader can verify it is not a flattering account of flattery-avoidance.

For v2. Replace the ratio in §6 with a counted table per deposit. Verify [5] and [8] against source. Add a one-page worked example (a single claim carried from pre-registration through a kill) as an appendix.

Change log

v1 — practice stated as a loop (Figure 1) and a dial (Figure 2); artifacts tabulated (Table 1); the method framed as an instance of its own thesis; credibility staked on the non-empty retired column (§6) rather than on assertion; structural limit (single-source map; independence only from outside) made explicit in §7. Open for v2: counted kill/confirm table; citation verification of [5],[8]; worked-example appendix.

— end of v1 methods note and referee report —