The novelty gate: five checks that caught seven fluent ideas before they shipped

W. Schulz

Sep 8, 2026

Abstract

A procedure calibrated on seven attractive ideas that turned out not to be discoveries: does the claim forbid anything; is the object older than the literature; is a cross-domain $R^2 \approx 0.99$ really an identity; does it survive changing microscopic detail; and a hostile prior-art sweep. A meta-check holds deflations to the same bar.

AMENDED 2026-07-27 — read **`SOUNDNESS-AXIS.md`** alongside this. This gate is single-axis (novelty) by construction, and returning "known" has been over-reported as "deflated." Novelty governs what to file/publish/spend an experiment on; soundness governs what to build on and believe. A claim that is *not novel but true* is an asset, and landing on a named theorem is corroboration that the reasoning is sound — record it as a pass, not a loss. Reserve "deflated" for claims that are false or empty.

A reusable procedure for deciding whether an idea is real, before investing in it. Distilled 2026-07-26 from seven consecutive deflations in this repository — including one where the gate caught its own false positive.

This is the most durable thing this project has produced. Not because the procedure is clever, but because it has been calibrated against a known outcome set: seven fluent, attractive ideas that all felt like discoveries from the inside, and were not. It costs roughly a day for a framework and about seven minutes for a bad experiment.

Why this exists

Generating plausible frameworks is now free. Discriminating among them is not. The sensation of insight and the property of surviving scrutiny feel identical from the inside — that is not a character flaw, it is a property of fluency, and it applies to human and machine reasoning alike.

The record this was calibrated on:

Idea	Felt like	Actually was
Waves/complex as substrate	new architecture class	decorative at equal budget
Grounding metaphysics	formal ontology	no content where a poset exists (5/5 negative)
Spectral recurrence theorem	new theorem	Arnold–Papanicolaou–Wihstutz 1986
Forgetting = selection	new diagnosis	prior art; beaten by LUCIR

Idea	Felt like	Actually was
Ontogenic mathematics (geometry)	new foundation	Klein–Lie 1871, Grayson 1987, Thom 1972
Ontogenic mathematics (ontology)	open fork	definitional; Peirce, ~1900
The ε universality class	cross-domain law	Thurstone 1927, an identity

Seven for seven. The gate is what that cost bought.

The five checks, in order

Order matters. Each is cheaper than the one after it, and any can terminate the investigation.

Check 1 — Does it forbid anything?

The single highest-yield question, and it is free.

State the claim, then name a concrete observation it *rules out*. If you cannot, the claim is empty rather than false, and no amount of evidence will help — evidence cannot support a claim that excludes nothing.

Worked example. "Objects are the fixed points of recursive realization" sounds like an ontology. But every recursion has a fixed point (Kauffman 2005: take J = F(F(F(…))), then F(J) = J — he calls his own proof "a magician's trick"). A statement true of every possible dynamics constrains nothing about this one. The content was entirely in stability, basin measure, and rate — none of which the framing supplied.

Red flags: the claim survives every hypothetical you construct; its opposite is hard to state; it "explains" outcomes that would look equally explained if reversed.

Check 2 — Is the object older than the literature?

Novelty lives where you need equipment to pose the question.

A circle is 2,500 years old and requires no apparatus to think about, so the space of askable questions about it has been thoroughly occupied — which is why Thread A landed on 1871. A trained transformer is ~10 years old and requires a GPU to interrogate, so questions about it *could not have been asked* before.

Ask: what apparatus is required to even state my question? If the answer is "none," expect prior art and go looking for it immediately. If the answer is "something recent, expensive, or rare," your odds are much better.

Corollary: philosophy and classical plane geometry are exhausted. Measurement of young systems is not.

Check 3 — The identity test

> An $R^2 \approx 0.99+$ that holds across unrelated domains should trigger "I have derived an identity," not "I have found a law."

Real empirical laws in messy domains do not fit that well. A fit that good, that portable, almost always means the functional form was implied by a distributional assumption shared by every system you tested.

Worked example. $\varepsilon(\sigma) = \Phi(-\text{gap}/(\sigma\sqrt{2}))$ fit at $R^2 = 0.997$ across three unrelated fields. It is the exact closed form for the difference of two i.i.d. Gaussians — Thurstone 1927. The R^2 was measuring the Monte-Carlo error of a random number generator. Three domains agreed because scores everywhere are sums of many small contributions, i.e. the central limit theorem.

The test: try to derive your "law" analytically from the assumptions of your own simulator. If you can, you did not discover it.

Check 4 — Break the microscopic detail

If you are claiming a universality class, the claim carries an obligation: universality *means* robustness to microscopic detail. So change the detail.

Swap the noise distribution, the substrate, the mechanism — anything the claim says should be irrelevant. If the relationship changes form, you have a different fixed point, not an irrelevant operator, and it is not a universality class.

Worked example. The ε law survives Gaussian→Gaussian and dies under Cauchy (723× error at $\text{gap}/\sigma = 6$). Its real class is "the Gaussian domain of attraction" — true, useful, and discovered by Laplace.

The discipline behind this: universality classes are fixed by symmetry, dimensionality, and interaction range — never by resemblance. The nautilus, the spiral galaxy, the hurricane and the sunflower are four spirals with four unrelated mechanisms. Same form does not imply same process. If your grouping would put the nautilus with the galaxy, it is resemblance.

Check 5 — Adversarial prior-art fan-out

Only now spend real effort. Run parallel searches instructed to be hostile to novelty and to expect deflation. Demand named theorems, authors, and dates — not "this is related to."

Search *sideways*, not just in your own field. The three fields that repeatedly owned this repo's results were psychophysics, digital communications, and classical dynamical systems — none of which anyone here would have searched by instinct. If your result concerns choosing among noisy options, someone in psychophysics formalized it before 1960.

Two tooling rules, added 2026-07-30 after a gate that keyword search nearly missed:

1. Sweep the citation graph, not just keywords. Enumerating an anchor paper's *citing set* via the Semantic Scholar API found the two most damaging hits after roughly 200 keyword searches had missed them, at near-zero marginal cost. Keyword search finds papers that share your vocabulary; the citation graph finds papers that share your *problem*. Do the graph sweep before concluding a gap.
2. Read the ancestor before reading a null as a finding. A gate grepped Manheim & Garrabrant (arXiv) for mitigation content, found none, and reported the fork as unclaimed. The paper says on its face that it expands an earlier LessWrong post — and that post has a per-type *Mitigation* section containing the exact fork. When a source says it extends earlier work, the earlier work is part of the prior art. An absence in the derivative is not an absence in the literature.

Corollary for the meta-check: both rules above cut *toward* deflation, and both were discovered by a gate correcting itself. That is the meta-check working — but it also means a gate that has not run them has not finished, and "we searched hard and found nothing" is not a verdict until the graph sweep and the ancestor check are done.

The meta-check: the gate needs a gate

Deflation pressure is not self-validating. After a run of correct deflations, "you were wrong again" starts matching the expected pattern and gets accepted with *less* scrutiny than "you were right." That is the same bias as fluency, pointing the other way — and a discipline built to remove false claims will start inserting them.

This actually happened. An adversarial gate objected that $\varepsilon^* = p/(1-p)$ was an algebra error — a probability compared to an odds ratio, vacuous for $p > ½$. Confident, dimensionally reasoned, and false: in the source, p is a base rate (0.10), not a voter competence, so $p/(1-p) = 0.111$ is an ordinary probability and the formula is exact. An experiment "confirming" the objection was written before anyone checked the source definition.

Rules that follow:

1. Before accepting any verdict about a *formula*, resolve what every symbol denotes in the original source. Grep for the definition and a worked numeric instance. Most "algebra errors" are notation collisions.
2. Hold deflating findings to the same evidentiary bar as confirming ones. Ask "what would make this objection wrong?" as hard as you asked the reverse.
3. Retract visibly, in-file. Never silently amend. The retraction is part of the record.

Surrounding practice

- Pre-register. Write predictions in the file header *before* running. Run once. Report misses in the same place as hits. A 2/4 whose misses changed the conclusion is worth more than a 4/4 that decorated it.
- Do the floor/ceiling arithmetic before you register a threshold. For every registered statistic, compute its *achievable range* under the design — then check the threshold fits inside it. This program has now burned registered predictions on impossible tests in at least three strands: cosmogony C3 (a 3-bit spread registered where the event definition's 158-state floor left less than that available), and breath B1/B6 (a "drops ≥ 0.20" test on a statistic that starts at 0.052, and a "rises ≥ 0.20" test on one that starts at 0.948 and saturates at 1.000). The failure mode is insidious because it produces a MISS while the underlying substance is *stronger* than predicted — total collapse reads as a failed prediction. Two cheap habits fix it: (i) run a pilot on throwaway seeds purely to bound the statistic's range, registering nothing;
  (ii) prefer thresholds stated as *fractions of the achievable range* ("closes $\geq$ half the distance to the floor") over absolute deltas.
- A degenerate comparison is not a verdict. If both arms of a contrast sit at the floor or both at the ceiling, the test did not run — report a design failure, not a finding, even when a pre-committed kill technically fired. (Precedent: `language/` L1c, whose own registration said "design failure,

not a result"; breath B4, where the kill fired twice, once at each extreme.) Honor the kill's *scope restriction* — do not go hunting a friendlier regime for the same question on the same design — but do not bank its *conclusion* either.
- Gate before you build. The literature check comes *first*. This has now prevented a build once, for seven minutes of agent time, on a proposal whose premise was false for the target domain and whose cleanest instance was published in 2020.
- Separate "false" from "unoriginal." Prior art is not a defect — a true claim someone else reached first is still true, and attribution is the whole exercise. False claims are different and must be removed, not labelled. A status banner does not stop a false sentence from being quoted.
- Keep the archive, flag the archive. Preserve source transcripts verbatim as records of what was said, with correction notices and inline flags at the offending lines.

What a passing idea looks like

Everything that survived in this repository has the same shape, and it is not the shape the intuition wants:

- a measured number attached to a decision (verifier-gated halting: gated beats best-fixed depth, 0.991 vs 0.960 mean over 10 seeds, at ~2.3×±0.8 less compute), not a reframing; *(and note the audit lesson: the ORIGINAL single-seed version of this bullet said "0.98 vs 0.92 at 2.45×" — the qualitative claim survived N=10 replication but the precise ratio did not. A measured number attached to a decision still needs a distribution.)*
- about an object younger than the literature;
- contingent — true of today's systems, not of all possible worlds;
- and it feels smaller than the framings that deflate.

That last point is the useful heuristic. If it feels like a foundation, gate it harder.

Emergence-claims checklist (added 2026-08-21)

A fourth gate type, for claims of the form "X emerges from a network / learned representation trained on D" — a geometry, a law, a metric, a symmetry, a world model. Calibrated on eight instances in one session (`verification/emergent-spacetime-gate.md`), all of which resolved the same way: *what emerged was what was put in.* Run these four before reading any such claim as evidence about the physics. Harness: `verification/emergence-audit/`.

#	question	kill	owner of the logic
E1	Stateless generator. Does the same X come out of data with the claimed physics removed — a closed-form curve, a classical Gibbs state, a shuffled process?	same X ⇒ X is a property of the pipeline or of low-order data shape	Elsayed–Cunningham 2017; Touboul–Destexhe 2017

#	question	kill	owner of the logic
E2	Non-solution generator. Does data that *violates* the law said to emerge fit as well, with the violation reproduced in the output?	fits as well ⇒ the pipeline does not forbid the alternative, so it cannot have found the law	Lipsitch 2010 (negative controls); Vafa et al. 2025
E3	Prior swap at fixed fit. Among loss-equivalent outputs, change only the regulariser. What moves?	X moves ⇒ the moved part is the prior's, not the data's	Rashomon (Fisher–Rudin 2019); Kallioinen 2023; Gruver 2022
E4	Family distance and conditioning. Distance to the nearest member of the claimed solution family vs to the generator; condition number of the parametrisation.	closer to *another* family member than to the generator, or cond ≫ 10^2 ⇒ function-space agreement is not identification	Sethna sloppiness; Raue 2009; Hadamard

Two habits from the calibration set: - "No coordinates in the input" is not "no geometry in the input." The geometry of a state is the locality of the Hamiltonian that prepared it, read back (EG1), and even that is a property of the tensor-product structure, not of the Hilbert-space data (EG2). Ask *what prepared the data* before asking what the network found in it. - Run the paper's full recipe. At reduced epochs two of the four verdicts flipped (`verification/emergence-audit/README.md`). An under-trained audit is an unreliable audit.

Calibration record (all eight: what emerged was what was put in): MI→MDS geometry = the preparing graph (EG1); geometry = the tensor-product structure (EG2); the "single light cone" = Lieb–Robinson, scaled by the state's correlation length (EG3); a learned holographic metric = the ODE ansatz + the smoothness prior (R1, R1b); a "hyperbolic" bulk = a fixed halving tree fitting a log curve, state or no state (R2); "emergent quantumness" from SGD = a postulate contradicted by the paper's own first equation (R3); a published learned metric "precisely identified" at MSE 10^{-6} with its co-trained parameter 57 % off (Ahn 2024 audit); "resistance-like" entanglement distance = Chebotarev 2011 (gated OWNED, retracted).

Check 0.5 — grep the source for its own defeaters (added 2026-08-26)

Calibrated on ~10 consecutive AI-generated architecture proposals (`log-arithmetic/GATE*.md`): the killing objection was almost always in the text itself — stated accurately, then not propagated ("flatten the graph and count the weights"; "you can't parallelize B(A(x))"; "those billion compositions aren't guaranteed useful"). Fluency includes the refutation without executing it. So before any external search: read the source for the sentence that kills it, and check whether the argument downstream of that sentence survives taking it seriously. Costs minutes; in this calibration set it terminated 4 of 10 gates on its own.

Check 0.7 — which N, and which boundary? (added 2026-09-08)

Calibrated on the wavefield-graphics thread (`wavefield-gate/GATE.md`): a fluent five-round escalation from "eight rotating circles draw a heart" to "framebuffer-free photonic graphics" rested on two silent substitutions, each killable by one question.

1. Which N is being claimed — the *representation's* or the *boundary's*? A demo's sparsity is a property of the object it represents (a 1-D curve: 8 Fourier modes, exact). The pixel/mode count a display, sensor, or modulator must control is a property of the *boundary* — space–bandwidth product, $N = (2 \cdot \text{eyebox} \cdot \sin(\text{FOV}/2)/\lambda)^2$, independent of content. Conflating them made "10^4 modes replace $2 \cdot 10^6$ pixels" sound like physics; the 2-D field of the same heart needs 49 % of N in Fourier. *Ask: what is the dimensionality of the thing represented, and what is the DOF of the surface it must cross?*
2. Which boundary does the proposed operation sit between? An analog stage pays one conversion per element on each side it touches (`SUBSTRATE-GATE` check 4: 111 int8 MACs per 8-bit conversion). Price it against the *best* digital algorithm, not the naive one — the "global linear transforms optics does for free" are exactly the ops the FFT already made O(log N)/px, so mid-pipeline optics tops out at ~58 MACs/conversion. The only zero-boundary stage is passive optics on the far side of the last modulator, and its op set is a lens and a prism. *Ask: name both sides of the stage, and who pays each side.*

Also confirmed here: the 1-D → 2-D generalisation is a specific instance of Check 4 (break the detail — dimension is a class-fixing quantity, resemblance is not).

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/novelty-gate