A critic that shares the generator's prompt shares its errors: measured self-critique correlation

W. Schulz

Sep 9, 2026

Abstract

Self-critique failure is mostly classical correlated-jury theory, and the note says so first. Measured: a same-context critic missed 12/12 planted errors that fresh critics all caught, and a panel's (e, P_both) predicted its majority accuracy (0.841 vs 0.842). The lever was prompt scaffolding, and only for context-induced errors.

The ρ-correlation program, executed gate-first per `docs/NOVELTY-GATE.md`. Verdict up front: the framework is deflation #8 — and this time the gate ran BEFORE the build, the residue was identified in advance, and the pilot was scoped to the residue. One jury-math file (3/5 pre-registered + 2 corrected post-run), one real-model micro-pilot (T2 confirmed at maximal effect size), one shipped tool (engineering on prior mathematics, attributed as such).

1. The gate: near-total deflation of the framework (as predicted)

Two adversarial fan-outs (classical + LLM literature). The claim "self-critique failure is the correlated-jury theorem; decorrelation not volume is the lever" is ~90% classical ensemble theory renamed and ~10% recent LLM literature restated:

- The correlation-not-accuracy ceiling is quadruply classical: Breiman 2001 ($PE^* \le \bar{\rho}(1-s^2)/s^2$, the strength–correlation bound), Krogh–Vedelsby 1995 (ambiguity decomposition), Ueda–Nakano 1996 (bias–variance–covariance floor), Clemen–Winkler 1985 ($n_{eff} = n/(1+(n-1)\rho) \to 1/\rho$ — literally titled "Limits for the Precision and Value of Information from Dependent Sources").
- Jury math under correlation is settled: Ladha 1992/93, Berg 1993 (asymptotic CJT *fails* under Pólya dependence), Kaniovski 2010 (exact linear-in-c degradation), Nitzan–Paroush 1982 (log-odds weights). Liu & Yao 1999 already *trained* by penalizing error agreement.
- The LLM port is also occupied (2024–2026): CAPA / "Great Models Think Alike" (judge affinity scales with generator–judge similarity *controlling for capability*); "Correlated Errors in LLMs" (350+ models; wrong models agree ~60%); Kohli 2026 "Nine Judges, Two Effective Votes" (Kish $n_{eff} \approx 2.2$ of 9 LLM judges); Estornell & Liu 2024 (echo-chamber theorem for same-model debate); Pan et al. 2024 (shared context intensifies self-eval reward hacking — the same-vs-fresh-context rung, measured); TACL 2025 "Self- Consistency Falls Short" (position-induced correlated errors defeat SC — T2 as an established phenomenon); and a Jan 2026 preprint (preprints.org 202601.0892) stating nearly the exact thesis including "the lever is decorrelation, not adding gates." Counter-evidence mapped too: Self-MoA (best-model self-ensemble beats mixed MoA — the matched-accuracy clause does all the work) and SCoRe (RL *does* instill some self-correction — standing counter-evidence to T4).

Surviving residue, per both gates: (i) an end-to-end measured-ρ law validated across the full verifier ladder *at matched critic accuracy* — composition/validation, not concept; (ii) T4 as a measurement — show that distilling external verification back into the generator re-introduces ρ and forfeits the gain, designed to beat SCoRe; (iii) the ρ-threshold budget rule — unclaimed but a two-line corollary of Kish. Any writeup must cite CAPA and Correlated-Errors as incumbents. Also read preprints.org 202601.0892 in full first — it may already occupy (i).

Process note: total gate cost ≈ 12 minutes of agent time, spent *before* any experiment. The framework would have taken days to build and would have died in review.

2. Jury math layer (**`exp_r_jury_math.py`**): 3/5 pre-registered, 2 corrected

Confirmed: majority-of-N saturates at 1−s (R1); the ceiling is computable from two observables — per-critic error e and pairwise co-error P_both, via $s = (P_both - e^2)/(1 + P_both - 2e)$, recovered to ±0.001 (R2); the Kish n_eff law holds to 0.6% (R3).

Two of my predictions were wrong, and the corrections matter:

- R4 miss: poisoned-item accuracy is not *flat* in N — voting still removes idiosyncratic noise and climbs *to* the ceiling (0.524 → 0.700 = 1−s exactly). The empirical signature of poisoning is a ceiling gap, not flatness (R4b: gap 0.300 at N=101, confirmed).
- R5 miss: decorrelation does *not* universally beat accuracy — my parameter guess was wrong. At N=3 the better-but-correlated panel wins (0.842 vs 0.807); the independent panel overtakes by N=7 and the shared panel is pinned at 0.850 forever after (R5b, confirmed). Design guidance is N-dependent.

3. The micro-pilot: T2 confirmed at maximal effect (real models, one shot)

Design: haiku generator answers 24 items (12 percentage-change, 12 neutral arithmetic) with a poisoned context note (a false "company convention": divide by the *new* value). Six critics, one shot each, two replicates per condition. Ground truth exact.

Generator (P1 confirmed): followed the false rule on 12/12 bias items, got 12/12 neutral items right. Errors exactly = the planted bias.

Critic condition	Miss rate on the 12 planted errors	False alarms
same-model, same context (haiku + memo), ×2	12/12 and 12/12	0
same-model, fresh context (haiku), ×2	0/12 and 0/12	0
different model, fresh (sonnet), ×2	0/12 and 0/12	0

P2 (T2 direction) — confirmed at the maximum possible effect size. The context note alone flips the *same critic model* from perfect detection to perfect endorsement. Within the poisoned condition, replicates are identical: the shared component on bias items is s ≈ 1, so by R4b no number of same-context self-critiques would ever catch these errors — every additional sample votes the same wrong way. The one decorrelating move that costs nothing, *dropping the context*, catches all twelve.

The interpretive point, which is the whole mechanism: the same-context critic is not worse at arithmetic — it is *coherently applying the poisoned convention*. Under the memo, the answers ARE "correct per company style." A critic that shares the generator's context shares its standard of correctness, and a standard cannot be used to catch itself. This is the displacement pattern's fourth appearance: the instrument of checking cannot be absorbed into the thing checked without ceasing to check.

Honest limits: P3's ladder is degenerate here (fresh-haiku already at 0/12, so no headroom to separate fresh-same-model from different-model — task too easy for the fresh critics); P4 unmeasurable (generator made zero neutral errors); n = 24, one bias type, one task family, and T2-as-phenomenon is already established in the literature (TACL 2411.01101; Pan 2407.04549) — so this pilot is a *replication demonstrating the protocol*, not a discovery. What it validates is the measurement design: (e, P_both) per condition is sufficient, cheap, and the effect is not subtle.

4. Shipped: **`verifier-bottleneck/verifier_panel.py`**

Engineering on prior mathematics, attributed as such in its docstring (Clemen–Winkler n_eff; Ladha; Breiman; Thurstone for the tail check):

- `verifier-panel panel --e E --pboth P [--n N] [--target-acc A]` — estimates ρ and the shared component s from the two observables, reports the majority ceiling 1−s and n_eff, and answers the budget question: can same-substrate sampling reach the target, or must you buy a decorrelated channel? Includes the R5b small-N caveat.
- `verifier-panel check-tail scores.txt` — the Φ-fit precondition test from the epsilon-universality campaign (excess kurtosis; heavy tails make the Gaussian ε law optimistic by up to 723×). Also a `--gap/--sigma` warning mode.

Smoke-tested; the pilot's own numbers (e=0.5, P_both=0.5 → ρ=1.0, n_eff=1.0, "decorrelate") reproduce as the worked example.

5. What remains, and what it costs

- T4 (the internalization measurement) — the deep one, still unclaimed, with SCoRe as the counter-evidence to beat. Needs fine-tuning compute.
- End-to-end ρ-law validation at matched accuracy — needs an API-scale run (hundreds of items, matched-accuracy critic construction per BIG-Bench Mistake subsampling). The protocol is now fully specified by this pilot plus the gate's "designs to reuse" list (CAPA regression; Pan context ablation; Chen few-sample ceiling prediction).
- Read preprints.org 202601.0892 in full before writing anything.

Neither is runnable in this environment (no API key, no training compute). Both are specified to the point where the next session can execute rather than design. Per the gate: proceed only for T4 and the end-to-end validation; the framework itself is a synthesis of known results and must be presented as one.

⚠ SOUNDNESS FLAG (2026-07-30) — the strong ρ-law form is FALSE

A hostile citation-graph sweep run for a different strand returned a result that lands directly on §2 of this file, so it is recorded here rather than left in the other track.

Kaniovski 2010 (*Theory and Decision* 69(3):453–468), Theorem 2, exhibits distributions where correlation has exactly zero effect on jury competence:

> *"The possibilities range from completely insensitive, to very sensitive... In order for a correlation to make a difference, either the jury must be heterogeneous or a supermajority must be the decision rule."*

So the strong form — $\rho > 0 \Rightarrow$ panel accuracy capped — does not hold in general. This file already cites Kaniovski as prior art ("exact linear-in-c degradation"); the citation was right and the reading was wrong.

What survives: the *conditional / common-cause* form (Dietrich & Spiekermann 2013) and the *second-moment* form (Ueda–Nakano 1996). Our own measurements are unaffected as measurements — R3's Kish n_eff law held to 0.6% *in the homogeneous, simple-majority regime we simulated*, which is precisely a regime where Kaniovski's insensitivity result does not bite. The error is one of scope: the law was stated generally and holds conditionally.

Consequence for the shipped tool. `../verifier-bottleneck/verifier_panel.py` estimates a majority ceiling 1−s from (e, P_both). That estimate is valid under the homogeneity assumption it inherits from this file's model and should not be presented as substrate-general. Anyone extending it to heterogeneous panels or supermajority rules must re-derive against Kaniovski Thm 2 first.

6. Preprint 202601.0892 read in full (v5, 25 May 2026) — 2026-09-08

Brilliant, A. M., *Limits of Self-Correction in LLMs: An Information-Theoretic Analysis of Correlated Errors* (independent researcher; `preprint-202601.0892.v5.pdf` in this directory). Verdict: does NOT occupy residue (i); it explicitly calls for it.

- Kind of paper: "a *conditional analysis paper*, not an empirical paper" (§1.5); "No new data were created or analyzed" (Data Availability). Theorems are chain-rule bounds: $I(T;S|G) \leq I(T;Z|G)$ under $S \perp T \mid (G,Z)$; k rounds of self-critique cannot exceed the Z-ceiling (Lemma 2) — our R1/R4b "ceiling gap" in information-theoretic dress. The author himself flags Theorem 1 as vacuous without an independently specified Z (Remark 2, §3.3).
- Our axis is his Open Question #2 verbatim (§8.5): "Does fresh-context same-model evaluation yield measurable reduction in error correlation in practice, even though it does not satisfy the formal independence conditions?" He calls same-model context separation "an engineering heuristic, not a theoretical solution" and defers validation to "a companion paper" (§6.9). That companion paper is the competitor for our end-to-end run.
- Quantities: he proposes measuring $\kappa = \Pr(A=1|T=0)$ and $\delta = \Pr(A=1|T=1)$ per selector, evidence bound $\log_2(\delta/(1-\varepsilon))$ (Prop. 2), via CriticBench / P(True). He has no pairwise quantity — no P_both, ρ, s, n_eff — so his framework cannot say whether *more* selectors help. Ours can (Kish n_eff from (e, P_both)). Our measured s is a direct operationalisation of "what Z mediates" in the homogeneous regime → an answer to his Open Question #1, with the Kaniovski scope caveat (§ SOUNDNESS FLAG above).
- Citations he lacks (we must carry them): Kaniovski 2010, Clemen–Winkler 1985, Kish, CAPA (Goel et al.), Kohli 2026, Pan et al. 2024 (shared context ↑ self-eval reward hacking — the nearest prior measurement of our axis), TACL "Self-Consistency Falls Short", Estornell & Liu 2024. He does cite Kim et al. ICML 2025 (60 % same-wrong- answer agreement), Tsui 2025 Self-Correction Bench (64.5 % blind spot; "Wait" −89.3 %), CriticBench, P(True).

- Design gift: his Definition 7 (context separation = no access to generation trace, prompt scaffolding, or hidden state) is a clean three-way factorisation of "same context" — adopt it as the ladder's rungs.

7. End-to-end ρ run COMPLETE (2026-09-09) — see **`RESULTS-endtoend-rho.md`**

8 370 isolated `claude -p` calls, 0 failures after retry. Residue (i) delivered (P5: the (e, P_both) → majority prediction validated out-of-sample on real models at n=120, 0.841 predicted vs 0.842 observed) and the framing hypothesis killed (the pre-registered kill condition fired: shared prompt context is not the correlation lever for model-induced error). Headline, discrimination δ−κ: same-context on planted errors −0.000±0.004 (a coin flip) → fresh context +1.000; but on natural errors every same-model rung is 0.04–0.11 and on GSM-Hard the fresh critic goes negative (−0.081) — it rejects 54 % of correct answers, so its lower miss rate is a rejection reflex, not detection. The lever is the SCAFFOLDING, not the trace (L1, which reads the full reasoning but not the memo: +0.992). Answers Brilliant 2026's Open Question #2 with a measured number and a scope: yes for context-induced error, no for model-induced. Methodological keeper: a critic study that measures catch rate without false-alarm rate can report the sign backwards — this run would have.

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