When more sampling hurts: verifier exploitability, not accuracy, bounds test-time compute

W. Schulz

Jun 27, 2026

Abstract

Best-of-N sampling lifts coverage log-linearly, but realized accuracy is ceilinged by the verifier — and the property that matters is not discrimination but exploitability ε: the rate at which wrong candidates score high. Any $\varepsilon > 0$ makes accuracy non-monotone in N, peaking at a computable budget before declining.

⚠ ε CORRECTION (2026-07-26)

Claims in this document about the ε law being a novel cross-domain law or a universality class are superseded. $\varepsilon(\sigma)=\Phi(-\text{gap}/(\sigma\sqrt{2}))$ is Thurstone's Law of Comparative Judgment, Case V (1927) — an identity, not a discovery — and it is not a universality class (it breaks 723× under heavy tails). The class-IL row is a proved theorem (Kim et al., NeurIPS 2022). The measured constants, the shipped tools, and the formula $\varepsilon^*=p/(1-p)$ all stand. Canonical correction: **CORRECTION-epsilon-thurstone.md**.

Preprint draft, 2026-06-27. Reproducible: `verifier-bottleneck/` (numpy only, seed=0). This draft consolidates the run-once results recorded in FINDINGS.md, the 10-step hardening program (NEXT-10-STEPS.md), and the real-data calibration (05_calibration_real.py).

Abstract

Repeated sampling ("Best-of-N", test-time compute) lifts answer coverage — the probability that *some* sample is correct — log-linearly and without plateau. Realized accuracy does not follow, because a verifier must *select* among samples, and the verifier is the ceiling. We argue the relevant property of that verifier is not its discrimination (AUC / top-1 accuracy, the number usually reported) but its exploitability ε: the rate at which wrong candidates are scored *high*. We give a two-parameter selection model that separates discrimination d′ from exploitability ε and show, across pre-registered run-once experiments, that (i) an honest-but-noisy verifier (ε=0) is monotone — more sampling never hurts; (ii) any ε>0 makes accuracy non-monotone — it peaks at a computable budget N*(ε) and then *declines toward zero*; (iii) discrimination buys ceiling height but not immunity — you cannot out-discriminate an exploitable verifier; and (iv) selecting by consensus-among-high-scorers rather than argmax defeats the collapse *unless* the exploit is shared and frequent enough to win a plurality, a sharp threshold at $\varepsilon^* = p/(1-p)$. Inverting published reward-model overoptimization curves (Gao et al. 2023) through the model implies real verifiers sit at $\varepsilon \approx 0.002$–$0.05$ — a 1-in-20 to 1-in-500 blind-spot rate, invisible to AUC,

sufficient to explain observed Best-of-N saturation. The practical rule: past N*(ε), spend surplus budget on a *less-exploitable* verifier, not more samples, and select by consensus.

1. The two axes of verifier quality

We model the selection statistics of Best-of-N (no LLM is run; the ceiling lives in selection). For a problem with single-sample correctness p, each of N candidates is correct (w.p. p) or wrong. A verifier assigns a score: - Discrimination d′ — correct candidates score ~N(d′,1), honest-wrong ~N(0,1). AUC $= \Phi(d'/\sqrt{2})$; d′=∞ is a perfect verifier. - Exploitability ε — each wrong candidate is, w.p. ε, a Goodhart exploiter the verifier scores *high* (~N(d′+margin,1)). This is the blind spot a search eventually finds.

The single modeling move — splitting "imperfect verifier" into these two orthogonal axes — is the source of every result below.

2. Results (pre-registered, run once, seed=0)

P1 — coverage ≠ accuracy. At p=0.10, d′=2, ε=0.01: coverage→1.000 but Best-of-N accuracy=0.362 (gap 0.64). Selection, not coverage, is binding.

P2 — honest verifiers are monotone. With ε=0, max accuracy drop over N≤8192 is ≤0.001 for all d′∈{1,2,3}. A noisy-but-honest verifier never makes more sampling hurt.

P3 — exploitable verifiers peak then collapse. Any ε>0 makes accuracy non-monotone. At d′=2, ε=0.05: peak 0.42 @ N=16, falling to 0.07 by N=8192. Unbounded sampling drives accuracy toward zero.

P4 — the safe budget is governed by ε, not d′. The headline table (N*(peak accuracy)):

ε \ d′	1.0	2.0	3.0
0.001	8192 (0.74)	512 (0.93)	128 (0.96)
0.010	256 (0.45)	64 (0.71)	32 (0.79)
0.050	32 (0.25)	16 (0.41)	16 (0.49)

Read down: N* collapses as ε rises, at every d′. Read across: a better verifier lifts the peak (0.25→0.49 at ε=0.05) but N* barely moves and the collapse persists. You cannot out-discriminate an exploitable verifier; you can only make the hole rarer.

3. Hardening (follow-up stress tests)

- F1 — i.i.d. is the *worst* case. Per-problem correlation of exploits *bounds* asymptotic damage at ~(1−ε)·ceiling (a fraction of problems are doomed, the rest stay clean and monotone). The i.i.d. assumption — every problem an eventual time-bomb → accuracy→0 — is pessimistic, not optimistic. (This *falsified* our own initial guess that correlation worsens collapse.)
- F2 — buy the verifier, not the samples. Across budgets, the optimal tier shifts from a mid verifier to a *premium* (low-ε) one at B≈256; a cheap high-ε (0.05) verifier is never optimal at p=0.10 — it is dominated outright. A high-ε verifier is not a cheap option, it is a trap.
- F3 — consensus beats argmax, with a sharp limit. Filter-then-vote (majority among score>τ survivors) recovers argmax's 0.187 → 0.998 at ε=0.02, N=8192. But a *shared* exploit defeats consensus once it can win a plurality: observed crossover between ε=0.11 and 0.15, bracketing the pre-registered $\varepsilon^* = p/(1-p) = 0.111$. Below ε*, voting neutralizes the collapse entirely; above it, nothing does.

Unifying view: "exploitability" is three knobs — marginal rate ε (sets N*), correlation across a problem (bounds damage), and answer-clustering×frequency (decides if consensus defends). The single most dangerous hole is high-scoring, shared, and common — a systematic plurality blind spot.

4. Calibration against real overoptimization data

Inverting published Best-of-N gold-reward curves (Gao, Schulman, Hilton, *Scaling Laws for Reward Model Overoptimization*, arXiv:2210.10760) through the model: a real curve peaking at N* implies $\varepsilon \approx 1.45 \cdot N^{*\wedge(-1.19)}$.

real BoN peak N*	implied ε	reading
8	0.122	aggressive RM, early inflection
16	0.053	beam-search MATH inflection
64	0.010	typical reward-model BoN
256	0.0019	strong verifier, late saturation

Real BoN curves peak/decline at moderate N (16–256) ⇒ ε ≈ 0.002–0.05, a 1-in-20 to 1-in-500 blind-spot rate — enough to explain observed saturation, and invisible to AUC-style quality reports.

Honest limitation (does not survive at equal scrutiny): the *functional-form* match — that our accuracy-vs-KL is quantitatively quadratic-in-KL like Gao's gold law — fails its pre-registered bar ($R^2 = 0.73$ and -0.49, both < 0.9). The curve is hump-shaped like Gao's but not quantitatively quadratic. We therefore claim only the inversion (ε from peak-N), not the functional form.

5. Limitations

- No live LLM/RLHF model is run; the ceiling is modeled in selection statistics, calibrated to published curves. The i.i.d.-sample assumption *overestimates* coverage (Gemini GSM8K: model coverage@4=0.997 vs reported pass@4≈0.90 — real samples are correlated), so conclusions are conservative.
- The Gaussian-exploiter score model is a choice; the qualitative results (monotone-vs-collapse, N*(ε), the ε*=p/(1−p) consensus threshold) are robust to it, the precise numbers are not.

6. Contribution

A sharpening of the "verifier is the bottleneck" claim into something quantitative and actionable: the bottleneck is the verifier's exploitability ε, not its discrimination; the conversion of coverage to accuracy is non-monotone and self-limiting at a computable N*(ε); and the defense is consensus-among-high-scorers up to a sharp ε*=p/(1−p). Tooling: `verifier_budget` returns N*(ε) and the recommended selection rule from measured (p, d′, ε).

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