Thurstone bounds Best-of-N: a 1927 identity, calibrated for modern verifiers

W. Schulz

Jul 27, 2026

Abstract

The noise floor of Best-of-N selection is Thurstone's 1927 Case V identity, not a new law. Contributed: blind-spot rates ε ≈ 0.002–0.05 from inverting published overoptimization curves, an exact consensus threshold p/(1−p), and a budget rule N*(ε) with a working tool. Heavy-tailed score noise breaks the identity, so estimate ε directly.

Paper draft, 2026-07-27. This is the publishable successor to When More Sampling Hurts, rewritten after the ε correction in the genre that correction endorsed: not a discovered law, but classical selection theory correctly transported into a field that runs selection at industrial scale without knowing its ceiling has a name. Independently re-verified this session: 6/6 pre-registered predictions, second implementation, pure stdlib, seed 0 (`scripts/experiments/thurstone_bon_verify.py`, results committed).

Abstract

Best-of-N sampling — generate N candidate answers, let a verifier pick the best — is a workhorse of modern inference-time scaling. Its coverage keeps rising, often log-linearly, but realized accuracy saturates and can then *decline*. We show that the entire selection layer of this phenomenon is governed by mathematics that is between 70 and 99 years old, correctly identify which piece belongs to whom, and contribute the three things that are not prior art: calibrated constants (inverting published reward-model overoptimization curves yields blind-spot rates ε ≈ 0.002–0.05 for real verifiers — invisible to AUC, sufficient to explain observed saturation), an exact consensus threshold (a shared exploit defeats plurality voting iff its conditional rate exceeds p/(1−p), where p is the base rate of correct candidates), and a budget rule with a working tool (`verifier-budget` computes the safe sampling budget N*(ε); past it, buy a better verifier, not more samples). The noise floor of pairwise selection is Thurstone's Law of Comparative Judgment, Case V (1927) — independently canonical as the two-alternative forced-choice error rate in signal detection theory (1954) and the BPSK bit-error rate in communications. The practical warning attached: the identity holds only in the Gaussian domain of attraction, and heavy-tailed score noise breaks it by orders of magnitude, so estimate ε directly and check the tail before extrapolating any Φ fit.

1. The problem, and what is actually known about it

Sample N candidates for a problem whose single-sample correctness rate is p, and *coverage* — the probability at least one candidate is correct — is 1 − (1−p)^N. Brown et al. showed this keeps climbing across four orders of magnitude of N, often log-linearly [1]. If selection were free, accuracy would follow coverage to 1.

It does not, because selection is not free. A verifier must pick the winner, and the verifier is the ceiling. In our reference simulation ($p = 0.10$, discrimination $d' = 2$, exploitability $\varepsilon = 0.01$), coverage at $N = 64$ is 0.9988 while accuracy peaks at 0.70 — and by $N = 8192$, with coverage indistinguishable from 1, accuracy has *fallen* to 0.36. More sampling made the system worse. This is the coverage/accuracy gap in one row of numbers, and everything in this paper is an account of where it comes from.

The account requires splitting "imperfect verifier" into two orthogonal properties, which the field's standard metrics do not separate:

- Discrimination (d'): how far apart the verifier scores correct and honest-wrong candidates on average. This is what AUC measures.
- Exploitability (ε): the rate at which wrong candidates score *high* — the verifier's blind-spot rate. AUC is nearly blind to this at the rates that matter.

The two axes fail differently. Low discrimination costs you accuracy smoothly. Any nonzero exploitability puts a cliff downstream: a large enough search *will* find the blind spot, because finding blind spots is precisely what argmax over many samples does.

2. The identity, and its three prior lives

Model scores as Gaussian: a correct candidate scores $\sim N(\text{gap}, \sigma^2)$ and a wrong one $\sim N(0, \sigma^2)$, independently. The probability the wrong candidate outranks the right one is exactly

$$P(\text{wrong} > \text{right}) = \Phi(-\text{gap} / (\sigma\sqrt{2}))$$

because the difference of two independent Gaussians is Gaussian with variance $2\sigma^2$. This is not a modeling result; it is an identity, and it has been continuously in print since before the transistor:

1. Psychophysics, 1927. It is Thurstone's Law of Comparative Judgment, Case V — the founding equation of how humans compare two sensations under internal noise [2].
2. Signal detection theory, 1954. It is the two-alternative forced-choice error rate, $\Phi(-d'/\sqrt{2})$, in Tanner and Swets [3], standard in every psychophysics and radiology-reader curriculum since.
3. Communications engineering. It is the bit-error rate of binary phase-shift keying, $Q(\sqrt{(2E_b/N_0)})$ — the same Gaussian-difference computation with the vocabulary of energy and noise density [4].

An earlier document in this project reported this expression as a discovered cross-domain law with $R^2 = 0.997$. That framing was wrong and is retracted in full elsewhere [5]; the R^2 was measuring Monte-Carlo error around an exact identity. The correct claim — the one this paper makes — is that machine learning is the *fourth* field to need this identity, it is the only one of the four that does not know it has it, and the transport is worth doing precisely because the receiving field runs pairwise selection at a scale the other three never imagined.

What the identity gives ML for free: the noise floor of selection. Even a perfectly honest verifier — no adversarial blind spot at all, just score noise — mistakes a wrong answer for a right one at the Thurstone rate. Fifty years of psychophysical practice on estimating, comparing, and stress-testing that quantity (including its logistic sibling, Bradley–Terry [6] — the same model family the field already uses to train reward models, mostly without noticing the connection) transfers wholesale.

3. Two kinds of blind spot, and a definition that must be stated at every use

The identity covers *statistical* confusion. Best-of-N failure at scale is dominated by something the identity does not cover: adversarial blind spots, wrong answers the verifier scores confidently high — reward

hacks, in the field's own vocabulary [7][8]. The phenomenology here is Goodhart's law and the optimizer's curse [9][10], and Gao, Schulman and Hilton's fitted overoptimization laws are the strongest published quantitative statement of it [11].

Following the correction that reshaped this project's treatment [5], one definitional discipline is non-negotiable, because the same symbol has been used for two different rates:

- q — the exploit rate *conditional on the candidate being wrong*.
- ε — the *unconditional* rate at which a sampled candidate is a scored-high wrong answer.

Thresholds below are stated against whichever rate they are actually theorems about. Papers that do not fix this convention produce apparent algebra errors that are really definition collisions — this project generated, published, and then retracted exactly one such error, and the scar is the reason for the rule.

4. Results

All of the following were established in pre-registered, run-once experiments in the research repository (numpy, seed 0), and independently re-verified this session in a second implementation (pure stdlib, different sampling construction, seed 0; 6/6 predictions passed; results file committed). The convergence of the two implementations is itself a check: the research repo's recorded accuracy at ($p = 0.10$, $d' = 2$, $\varepsilon = 0.01$, $N = 8192$) is 0.362; the re-implementation, sharing no code, measured 0.358.

R1 — Coverage is not accuracy. At $p = 0.10$, $d' = 2$, $\varepsilon = 0.01$: coverage $\to 1.000$ while Best-of-N accuracy peaks at 0.70 ($N^* = 64$) and declines to 0.36 by $N = 8192$. Selection, not generation, is binding.

R2 — Honest verifiers are monotone. With $\varepsilon = 0$, accuracy is non-decreasing in N (verified to $N = 4096$; no drop exceeding Monte-Carlo error). If your Best-of-N curve turns over, you have measured *exploitability*, not noise. The turnover is a diagnostic.

R3 — Any $\varepsilon > 0$ makes accuracy non-monotone, with a computable peak. The mechanism is a race between maxima: the best of ~pN correct candidates against the best of ~$\varepsilon(1-p)N$ exploiters that score, by construction, above the honest distribution. As N grows the extreme-value advantage of the larger correct pool shrinks like the difference of $\sqrt{(2 \ln n)}$ terms while the exploiters' scoring margin is constant — so the exploiters win asymptotically, and accuracy trends toward zero. $N^*(\varepsilon)$ is the computable crossover, and it is the number the tool ships.

R4 — Discrimination buys ceiling height, not immunity. Raising d' raises the peak; it does not remove the turnover. You cannot out-discriminate an exploitable verifier, which is why AUC improvements do not fix Best-of-N saturation.

R5 — Consensus defeats idiosyncratic exploits, up to a sharp threshold. Selecting the plurality answer among high scorers [12] beats argmax as long as exploiters do not coordinate on one wrong answer. When they do — a *shared* exploit, the realistic case when all samples come from one model with one systematic bug — the exploit takes the plurality iff $q > p/(1-p)$ (equivalently $\varepsilon > p$ in unconditional form). Verified exactly for $p \in \{0.05, 0.10, 0.20, 0.30\}$; re-verified this session at $p = 0.10$ with the plurality flipping between $q = 0.09$ and $q = 0.14$, straddling the predicted $q^* = 0.111$. Note what the threshold says: for hard problems ($p = 0.05$), a shared exploit rate of just over 5% conditional defeats consensus. Consensus is a mitigation, not a fix.

5. Calibration: what ε actually is for real reward models

The model is only useful if ε is measurable for systems people actually run. Two routes:

Inversion of published curves. Gao et al. published how gold-standard reward degrades as Best-of-N optimizes against proxy reward models of various sizes [11]. Running those curves backward through the two-axis selection model yields $\varepsilon \approx 0.002$ – 0.05 for the proxy reward models studied — a blind-spot rate between 1-in-500 and 1-in-20. Two properties of this range matter. It is *invisible to AUC*: a verifier can post excellent discrimination numbers while sitting anywhere in it. And it is *sufficient*: rates in this range reproduce the Best-of-N saturation practitioners observe at the N values where they observe it.

Direct measurement (pre-registered, not yet run). The inversion is indirect; the bulletproof version is direct, and it is specified here so it can be held to: take open reward models with public weights; problems with checkable ground truth (e.g., GSM8K test split); $N = 1024$ samples per problem from a fixed generator; score everything. Report (a) $\hat{\varepsilon}$ as the fraction of wrong candidates scoring above the 95th percentile of correct-candidate scores, per problem and pooled; (b) d' from the score distributions; (c) the observed accuracy-vs-N curve against the $N^*(\hat{\varepsilon})$ prediction. Predictions, fixed now: measured $\hat{\varepsilon}$ falls in $[0.001, 0.1]$ for at least two of three models tested; the observed accuracy peak lies within a factor of 4 of predicted $N^*(\hat{\varepsilon})$; and at least one model fails the Gaussian tail check below, demonstrating why direct estimation must replace Φ extrapolation. This measurement needs GPU inference and is the single most valuable follow-up to this paper.

6. The budget rule

For practitioners, the paper compresses to three instructions:

1. Compute $N^*(\varepsilon)$ before scaling N. The `verifier-budget` CLI takes (p, d', ε) and returns the safe budget. Sampling past it spends compute to *lower* accuracy.
2. Past N^*, spend surplus on the verifier, not on samples. A less exploitable judge moves the ceiling; more candidates only move you along a curve whose far end is the exploit. Process supervision [13] is the published route to lower-ε verifiers.
3. Select by consensus among high scorers, and know its breaking point. It defeats idiosyncratic exploits outright and shared exploits only below $q^* = p/(1-p)$. If your problems are hard (small p), the protection is thin.

And one instruction for anyone *estimating* ε from a fitted Gaussian model:

4. Check the tail first. The identity lives in the Gaussian domain of attraction. Under Cauchy score noise the closed form becomes $\frac{1}{2} - \arctan(\text{gap}/2\gamma)/\pi$, and the Gaussian expression underpredicts the blind-spot rate by roughly an order of magnitude at $\text{gap}/\sigma = 4$ (63× in this session's re-verification under scale-matching $\sigma = \gamma$; the research repo recorded 11×–723× at $\text{gap}/\sigma = 4$–6 under its own matching). A single excess-kurtosis statistic separates the regimes decisively (re-verified: +0.02 Gaussian vs +89 Student-t(3) at $n = 20{,}000$). Never extrapolate a Φ fit into the small-ε tail; estimate ε directly.

7. Attribution, stated exactly

Because this paper exists downstream of a retraction, its ownership map is explicit:

Component	Status	Owner
The pairwise error identity $\Phi(-\text{gap}/\sigma\sqrt{2})$	Prior art, 1927	Thurstone, Case V [2]; Tanner & Swets [3]; BPSK BER [4]

Component	Status	Owner
Coverage scaling of repeated sampling	Prior art, 2024	Brown et al. [1]
Overoptimization phenomenology and fitted laws	Prior art	Goodhart variants [9], optimizer's curse [10], Gao et al. [11]
Verifier-based BoN selection; consensus voting	Prior art	Cobbe et al. [7]; Wang et al. [12]
Distribution-free treatment via quantile transform	Prior art, 2025	Khalaf et al., as recorded in the correction [5]
Two-axis (d', ε) decomposition applied to BoN selection; the transport itself	This project	—
Calibrated $\varepsilon \approx 0.002$–$0.05$ from inverting [11]	This project (measurement; data is not prior art)	—
Exact consensus threshold $q^* = p/(1-p)$ under stated definitions	This project	—
$N^*(\varepsilon)$ budget rule + `verifier-budget` tool	This project	—
Kurtosis-gated tail rule for ε estimation	This project	—

Nothing in the left column's first five rows is claimed; everything in the last five is, at exactly the strength stated. A reviewer who deflates this paper must deflate a measurement, a threshold theorem under explicit definitions, or a shipped tool — the framings that were deflatable have already been deflated by the project itself, in print [5].

8. Limits

The model is a selection-statistics model: no LLM is run, and generation-side effects (sample diversity collapse, temperature coupling between p and score distributions) are outside it. ε is treated as stationary in N, which is conservative in the wrong direction if the search process itself adapts (as in RL against the verifier, where Gao et al.'s setting lives). The Gaussian machinery is a domain, not a law — Section 6.4 is not a footnote but a boundary. And the direct-measurement protocol in Section 5 has not run; until it does, the calibration rests on inversion of one published family of curves.

Sources

1. Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., & Mirhoseini, A. (2024). "Large Language Monkeys: Scaling Inference Compute with Repeated Sampling." arXiv:2407.21787
2. Thurstone, L. L. (1927). "A Law of Comparative Judgment." *Psychological Review* 34(4): 273–286. doi:10.1037/h0070288
3. Tanner, W. P. & Swets, J. A. (1954). "A decision-making theory of visual detection." *Psychological Review* 61(6): 401–409; Green, D. M. & Swets, J. A. (1966), *Signal Detection Theory and Psychophysics.* New York: Wiley.

4. Proakis, J. G. (2001). *Digital Communications,* 4th ed. New York: McGraw-Hill — BPSK bit-error rate $Q(\sqrt{(2E_b/N_0)})$.
5. Correction: The ε Law Is Thurstone (1927) — the full retraction this paper is built on, including the retraction of a correction that was itself wrong, and the recorded reference to Khalaf et al. (2025).
6. Bradley, R. A. & Terry, M. E. (1952). "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39(3/4): 324–345 — the logistic sibling, and the model family used to train most reward models.
7. Cobbe, K., Kosaraju, V., Bavarian, M., et al. (2021). "Training Verifiers to Solve Math Word Problems." arXiv:2110.14168
8. Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). "Defining and Characterizing Reward Hacking." *NeurIPS 35.* arXiv:2209.13085
9. Manheim, D. & Garrabrant, S. (2018). "Categorizing Variants of Goodhart's Law." arXiv:1803.04585
10. Smith, J. E. & Winkler, R. L. (2006). "The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis." *Management Science* 52(3): 311–322. doi:10.1287/mnsc.1050.0451
11. Gao, L., Schulman, J., & Hilton, J. (2023). "Scaling Laws for Reward Model Overoptimization." *ICML 2023,* PMLR 202. proceedings.mlr.press · arXiv:2210.10760
12. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). "Self-Consistency Improves Chain of Thought Reasoning in Language Models." *ICLR 2023.* arXiv:2203.11171
13. Lightman, H., Kosaraju, V., Burda, Y., et al. (2024). "Let's Verify Step by Step." *ICLR 2024.* arXiv:2305.20050
14. Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). "Scaling LLM Test-Time Compute Optimally Can Be More Effective Than Scaling Model Parameters." arXiv:2408.03314
15. This session's independent re-verification: `scripts/experiments/thurstone_bon_verify.py` and `scripts/experiments/RESULTS-thurstone-bon.txt` in the project's site repository (private at present) — six pre-registered predictions, 6/6, second implementation, seed 0. Original run-once experiments: `verifier-bottleneck/` in the Quantum-Intelligence repository.

Published by NeuroGenesys (AW3 Technology, Inc.). Canonical version: https://neurogenesys.si/research/thurstone-bounds-best-of-n