Real judges did not break Best-of-N at N ≤ 32, and measured ε did not predict the budget

W. Schulz

Nov 10, 2026

Abstract

A pre-registered test of verifier exploitability on 9,600 Llama 3.1 8B candidates over GSM8K, MATH-500 and HumanEval+, scored by a reward model, an LLM judge and self log-probability. No judge made accuracy fall by N = 32, and measured ε did not order the best budget. A plurality vote over the top-scored half beat argmax in every math cell.

W. Schulz, NeuroGenesys. *Pre-registered 2026-11-09, run once 2026-11-10. This note reports the frozen analysis in full, including the two hypotheses that failed.*

Summary

The verifier-exploitability paper and the ε correction predict from toy models that a Best-of-N selector with an imperfect judge peaks at a finite N* and then declines, and that a judge's measured ε, the rate at which it scores a wrong candidate above a right one, tells you where that peak falls. This study tested both claims on real judges, with the protocol frozen before any main-run candidate was generated.

- No decline at N ≤ 32 (H1 not supported). For a reward model, an LLM-as-judge and the generator's own log-probability, Best-of-N accuracy rose from N = 1 and was flat by N = 16–32 in every cell. None fell between its best N and N = 32.
- Measured ε did not order N* (H2 not supported). Candidate ε ranged from 0.02 to 0.22, but the selected N* took only the values 16 and 32. The rank correlation had the wrong sign: $\rho = +0.26$, 95% CI $[-0.37, +0.78]$.
- A consensus rule held (H3 supported, in its tested half). In all 6 math cells, a plurality vote over the top-scored half of 32 candidates matched or beat picking the single top-scored one, by 3 to 17 points. Every cell was below its ε* threshold, so the "above ε*, plurality loses" half of the rule was never tested.

Both pre-registered kill criteria therefore apply. We make no exploitability-saturation claim for these judges at these budgets, and no claim that ε predicts the safe budget on real data.

Design

Item	Choice
Tasks	100 seeded problems each from GSM8K test, MATH-500 and HumanEval+

Item	Choice
Correctness	Checkers only: last number (GSM8K), `math-verify` equivalence (MATH-500), base + plus tests (HumanEval+)
Generator	Llama 3.1 8B Instruct (Q4_K_M, Ollama), temperature 0.8, 32 candidates per problem
Reward model	Skywork-Reward-V2-Qwen3-0.6B (bf16, local)
LLM-as-judge	Qwen2.5-Coder-7B-Instruct rating correctness 0–9, scored as the expected digit over its top-20 log-probs
Self-confidence	The generator's mean token log-probability
N grid	1, 2, 4, 8, 16, 32; expected accuracy computed exactly over random N-subsets

That makes 9,600 candidates, each scored by all three judges: 9 judge × task cells. Cost was $0, on one M2 Max shared with another study.

Definitions. N* is the N in the grid that maximises expected Best-of-N accuracy, with ties going to the smaller N. Candidate ε is the share of (right, wrong) candidate pairs on a problem where the judge scores the wrong one at least as high. p is the share of right candidates in a cell, and $\varepsilon^* = p/(1-p)$.

Hypotheses.

- H1: for at least one judge, accuracy at a cross-fitted N* beats accuracy at $N = 32$. N* is chosen on one half of the problems and scored on the other. The test is a one-sided bootstrap, Holm-corrected over the three judges.
- H2: across the 9 cells, higher candidate ε goes with smaller N*. The test is a Spearman $\rho < 0$ whose 95% bootstrap interval excludes zero.
- H3: at $N = 32$ in the 6 math cells, plurality over the top half beats or matches argmax below ε*, and loses above it. Supported only if every cell goes as predicted.
- H4 (seen vs. unseen training data) did not run. The reward model's training-data overlap could not be established from its data card, which was the pre-registered condition for skipping it.

Results

Checker pass rates were 0.766 on GSM8K, 0.361 on MATH-500 and 0.487 on HumanEval+. All three fell inside the pre-registered [0.05, 0.95] range, so no task was swapped.

Per cell

Judge / task	p	Candidate ε	ε*	N*	acc@1	acc@N*	acc@32
Reward model / GSM8K	0.766	0.022	3.27	16	0.766	0.890	0.890

Judge / task	p	Candidate ε	ε*	N*	acc@1	acc@N*	acc@32
Reward model / MATH-500	0.361	0.064	0.56	32	0.361	0.580	0.580
Reward model / HumanEval+	0.487	0.148	0.95	32	0.487	0.630	0.630
LLM judge / GSM8K	0.766	0.025	3.27	32	0.766	0.890	0.890
LLM judge / MATH-500	0.361	0.099	0.56	32	0.361	0.540	0.540
LLM judge / HumanEval+	0.487	0.119	0.95	16	0.487	0.634	0.630
Log-probability / GSM8K	0.766	0.049	3.27	16	0.766	0.833	0.820
Log-probability / MATH-500	0.361	0.133	0.56	16	0.361	0.453	0.450
Log-probability / HumanEval+	0.487	0.215	0.95	32	0.487	0.620	0.620

The largest in-sample gap between a cell's peak and N = 32 was 0.013. The `selrel audit` verdict was "no collapse at this budget" in all nine cells.

H1: more samples did not hurt

Judge	acc(N*) − acc(32), cross-fitted	95% CI	Holm p
Reward model	+0.000	[+0.000, +0.000]	1.00
LLM judge	−0.009	[−0.021, +0.004]	1.00
Log-probability	+0.000	[+0.000, +0.000]	1.00

H2: ε did not order the budget

Spearman ρ between candidate ε and N* over the 9 cells was +0.26, 95% CI [−0.37, +0.78]. The pre-registered direction was negative. With N* confined to {16, 32}, there was almost no budget variation left for ε to explain.

H3: plurality over the top half beat argmax in every math cell

Judge / task	Plurality	Argmax	Difference, 95% CI
Reward model / GSM8K	0.930	0.890	[−0.010, +0.100]
Reward model / MATH-500	0.650	0.580	[+0.000, +0.150]
LLM judge / GSM8K	0.920	0.890	[−0.030, +0.090]
LLM judge / MATH-500	0.640	0.540	[+0.040, +0.170]
Log-probability / GSM8K	0.920	0.820	[+0.030, +0.170]
Log-probability / MATH-500	0.620	0.450	[+0.090, +0.250]

The gain was largest for the weakest judge, self log-probability (+0.17 on MATH-500). The interval excludes zero in 3 cells and starts at zero in a fourth.

Secondary: the chosen budget transfers

An N* chosen on half A of the problems gave accuracy on half B within 0.036 of half B's own best in every cell, and matched it exactly in 5 of 9.

What this does and does not show

- It tests the budgets people commonly use. At $N \leq 32$ with an 8B generator, none of these three judges was exploitable enough to make more samples hurt. That agrees with work finding declines only at larger N or with weaker reward models (Cobbe et al. 2021 saw GSM8K verifier accuracy fall beyond about 400 samples). It says nothing about $N = 1{,}024$. The Thurstone-bounds paper specifies that measurement, and it has not been run.
- It is a real negative for ε as a budget predictor. The claim that a measured ε tells you where to stop sampling was the study's own contribution, and on this data it failed. A test with more budget variation, at larger N or with weaker judges, would be needed to revive it.
- It supports a cheap selection rule. Voting among the judge's top half was never worse than trusting the judge's single favourite, and was much better when the judge was weak. That is the below-ε* half of the threshold rule; the above-ε* half remains untested.
- Limits. One generator; 100 problems per task; the LLM judge is a coder model, chosen because it was already installed, which handicaps it on math.

Deviations

Five deviations are logged in the pre-registration, and none changed a candidate, label, judge, prompt or test. One was a rebased freeze commit. Three were interruptions (a machine restart and two judge-server timeouts) after which the pipeline resumed per candidate. The fifth cut judge-scoring concurrency from 16 to 4 after the shared GPU stalled at 16-way decoding. Judge requests are deterministic, so concurrency does not change any score.

Prior work

- Cobbe et al. 2021, arXiv:2110.14168: GSM8K verifiers decline beyond about 400 samples.
- Gao, Schulman & Hilton 2023, arXiv:2210.10760: Best-of-N over-optimisation scaling laws.
- Stroebl, Kapoor & Narayanan, arXiv:2411.17501: imperfect verifiers cap the gains from resampling.
- Huang et al. 2025, arXiv:2503.21878: Best-of-N with a weak reward model falls below the base policy at large N.
- Kang et al. 2025, arXiv:2502.18581: log-probability selection on MATH declines from N = 8 to 64 for Llama-3.1-8B-Instruct.

To our knowledge none of these measures a judge's ε first and tests whether it predicts N*. That was this study's test, and the answer here was no.

Data and code

The pre-registration, frozen analysis script, per-candidate checker labels, all three judges' scores, per-cell `selrel` audit logs and the analysis output are in the study folder `research_notes/Judge exploitability on real Best-of-N/` in the NeuroGenesys repository. Candidate texts are kept out of the repository because of their size.

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