WHETHER OR WHERE? TAIL CALIBRATION SIGNALS FINITE-BUDGET REWARD HACKING IN BEST-OF-N SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Best-of-N (BoN) inference returns the candidate with the highest proxy reward, so a larger N tests the proxy's extreme tail more than its average ranking quality. We show that the exact BoN value is a Beta-weighted integral of conditional true reward over within-prompt score quantiles. It follows that BoN depends on scores only through their ranks. It also follows that BoN can decline only if the calibration curve in the tail turns downward, and the budget where the decline appears depends on where that turn lies. Under a bounded additive model, the error tail sets the limit: Gaussian errors keep improving, Laplace errors reach a plateau, and heavy-tailed errors return to the prior mean. When error variance is largest at low true reward, even Gaussian noise can push BoN below random selection. Synthetic phase maps agree with this taxonomy. On real candidate pools spanning several verifiable domains, base policies, and learned reward models, tail failures are rare and all occur in code generation. A held-out tail calibration statistic reliably predicts where a cross-fitted budget cap helps. It does about as well as rank accuracy, and the cap is competitive with a Soft-BoN baseline. Tail calibration therefore does not replace ranking metrics. It is an audit that is unchanged by any monotone rescaling of the reward scores and shows which budgets are at risk.

1 INTRODUCTION

Inference-time alignment is the pragmatic answer to a stubborn problem: retraining a policy is expensive, but compute at serving time is cheap Stiennon et al. (2020); Nakano et al. (2021). The default recipe is best-of-N (BoN) selection: sample candidates from a frozen policy, score them with a reward model, return the winner. No gradients, embarrassingly parallel, and quality that improves as the budget grows. It is the closest thing the field has to a free lunch.

The lunch has a catch. BoN does not sample the reward model, it searches for its maximum. Push N high enough and the proxy score keeps climbing while verifiable correctness turns over and falls. We call this *inference-time reward hacking*: a regression caused entirely by the selection budget, with not a single weight updated. It arrives by spending more compute and is fixed by spending less, which leaves a practitioner two questions. **1) Diagnosis:** is this reward model exposed at all? **2) Budgeting:** if so, at what N?

Neither can be answered from the score distribution as a whole, because BoN never looks at it as a whole. Mapping scores to within-prompt quantiles $u = F_S(S)$, the quantile BoN returns has density $p_N(u) = Nu^{N-1}$, which collapses onto a vanishing neighborhood of $u = 1$; at $N = 64$, half the selection mass sits above the 99th percentile. **A defect confined to the top percentile can dictate an entire deployment while shifting any average by an amount no benchmark would flag.** Yet reward models are graded by preference accuracy, calibration, or mean margin Lambert et al. (2024c); Frick et al. (2024), all averaged over the region selection never visits, and mitigated by KL penalties, soft selection, or ensembling Gao et al. (2023); Eisenstein et al. (2023); Coste et al. (2023b), all applied at a uniform rate whether or not a task is at risk. A team can pass every leaderboard check and still ship a budget past its own optimum.

A useful diagnostic must therefore read only the top of the distribution, depend on ranks alone, separate exposure from scale, and end in a budget cap that survives held-out data. All but the last

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

follow from one identity. With R the true utility and $Q(u) = \mathbb{E}[R \mid F_S(S) = u]$ the reward response in quantile space,

$$V(N) \;=\; \int_0^1 Q(u)\, N u^{N-1}\, du, \tag{1}$$

so **every question about budget becomes a question about the upper end of a single curve**: whether V peaks depends on whether Q bends down, and where it peaks on how soon.

That shape has a clean cause. Under $S = R + \varepsilon$ with bounded R, the tail class of ε decides everything. Gaussian errors drive Q toward the top of the support of R and Laplace tails settle into a fixed exponential tilt, but regularly varying tails eventually make an extreme score *uninformative* about R, so Q falls. Only the last case admits a finite optimal budget, and it does so as soon as some intermediate budget already helps. Correlation strength leaves the tail class untouched, moving only the quantile at which informativeness runs out. **The tail decides *whether* hacking is possible, correlation decides *where* it appears.**

This tells us what to measure: a *tail excess* statistic confined to the upper rank region,

$$\mathcal{T}_\tau \;=\; \frac{1}{1-\tau} \int_\tau^1 \big(Q(u) - \mathbb{E}[R]\big)\, du, \tag{2}$$

whose sign flags exposure while ignoring the bulk, recovered by cross-fitting so the same data cannot both propose and confirm a defect. One half of a task then estimates $\mathcal{T}_\tau$ and picks a budget cap; the other half checks what that cap buys. The output is not a damping coefficient applied everywhere, but a named task, a named model, and a number.

We validate by releasing our own assumptions one layer at a time, from synthetic settings that match the analysis, through variants that strain and finally break its prior and a study perturbing only the reward model's input, to public correctness data spanning several domains and learned reward models. Tail failure turns out to be sharply localized rather than ambient, and a reading from the top percentile converts into a budget cap that still pays off on data never used to choose it.

- We formalize a blind spot in reward-model evaluation, namely that BoN concentrates on a vanishing top quantile while standard metrics report bulk behavior, and show that the resulting risk factorizes into a tail class governing whether a finite optimal budget exists and a correlation scale governing where.
- We turn this into a practical audit: a monotone-invariant tail excess statistic, a cross-fitted reward response curve, and an estimate-then-validate procedure returning a budget cap. Unlike uniform mitigations such as rank-Soft-BoN, it names the task and model at risk and confirms the decision out of sample.
- We validate at three levels of externality, from assumption-matched synthetics with deliberately violated variants, through an input-controlled study on a frozen candidate pool, to public reward-model data independent of our generator and model choices.

2 Related Work

Reward-model overoptimization. Proxy optimization can improve predicted reward while degrading human or oracle utility Gao et al. (2023). BoN policies have also been analyzed through distribution shift, KL control, coverage, and win-rate bounds Beirami et al. (2024); Huang et al. (2025). Reward-model ensembles can mitigate some overoptimization regimes Coste et al. (2023a). We study a narrower object: the joint calibration of proxy-score rank and true utility inside a fixed candidate pool.

Inference-time reward hacking and tail-aware selection. Kwa et al. Kwa et al. (2024) analyze catastrophic Goodhart effects under heavy-tailed misspecification. Khalaf et al. Khalaf et al. (2025) derive the probability-integral-transform and Beta order-statistic scaffold used here, study single-peaked inference-time hacking, and propose Best-of-Poisson and HedgeTune. We do not claim that scaffold as a new theorem: our contribution is to turn it into a score-transform-invariant joint-tail audit, separate tail regime from finite-budget localization, and validate a held-out budget rule. Huang et al. Huang et al. (2025) propose InferenceTimePessimism (ITP), which trades optimistic

selection for a scaling-monotone pessimistic rule. Most recently, Hsu et al. Hsu et al. (2026) propose Best-of-Tails, using a Hill estimate of normalized reward-tail shape to choose a Tsallis-divergence interpolation between optimistic and pessimistic selection. Our contribution is complementary: $E(u)$ estimates the *joint* relation between true utility and score rank, is invariant to strictly increasing score transformations, and is evaluated as an audit-set budget trigger. It requires labeled outcomes, whereas a score-tail method can operate without them.

Reward-model evaluation. RewardBench, RM-Bench, and RMB expose capability, style, and bias failures Lambert et al. (2024b); Liu et al. (2024); Zhou et al. (2024). RewardBench 2 explicitly connects multi-skill benchmark accuracy to downstream BoN and RLHF performance Malik et al. (2025). PPE asks which offline reward-model metrics predict post-RLHF outcomes and releases correctness data over MATH, GPQA, MMLU-Pro, MBPP+, and IFEval Frick et al. (2024); we reuse those public fixed candidate pools for external validation. At process level, PRM-BiasBench applies semantics-preserving and semantics-altering perturbations to reasoning traces and finds that fluent but corrupted steps can retain high reward Bamba et al. (2025). Our masked and shuffled controls apply the same general principle of controlled input perturbation to outcome-level reward models; we do not evaluate intermediate-step PRMs.

3 Problem Setup and Tail Representation

We separate three questions usually conflated in discussions of best-of-N (BoN): *whether* its value can turn down as N grows, *where*, and *how fast*. Recommendations are collected in Section 3.6; proofs are in Appendix A.

3.1 Score space, quantile space, and true utility

For a prompt x, let $Y_1, \dots, Y_N$ be conditionally i.i.d. responses with true utility $R_i = R(x, Y_i)$ and proxy score $S_i = S(x, Y_i)$ ($R_i = C_i \in \{0, 1\}$ in verifiable experiments, real valued in the synthetic study). BoN selects

$$I_N = \operatorname*{arg\,max}_{1 \le i \le N} S_i, \qquad V_N = \mathbb{E}[R_{I_N}], \tag{3}$$

with uniform random tie breaking.

For a continuous score CDF F_S, define

$$G(s) = \mathbb{E}[R \mid S = s], \qquad U = F_S(S), \qquad Q(u) = \mathbb{E}[R \mid U = u] = G(F_S^{-1}(u)), \tag{4}$$

with $\mu = \mathbb{E}[R]$ and excess utility $E(u) = Q(u) - \mu$; $B_{a,b}$ denotes a $\mathrm{Beta}(a, b)$ variable. Empirically, Q is estimated from within-prompt ranks (tied blocks averaged) and averaged over prompts; all statements below hold for this aggregate curve.

Proposition 3.1 (Two identities)**.** *Let S be continuous and Q bounded.*

(i) (BoN quantile identity)

$$V_N = \int_0^1 Q(u)\, N u^{N-1}\, \mathrm{d}u = \mathbb{E}\big[Q(B_{N,1})\big], \qquad V_N - \mu = \int_0^1 E(u)\, N u^{N-1}\, \mathrm{d}u. \tag{5}$$

(ii) (Marginal-value identity) *If Q also has bounded variation, then for every $N \ge 1$*

$$\Delta V_N := V_{N+1} - V_N = \int_0^1 u^N (1 - u)\, \mathrm{d}Q(u), \tag{6}$$

with $\mathrm{d}Q$ the Lebesgue–Stieltjes measure of Q; for absolutely continuous Q, $\Delta V_N = \mathbb{E}\big[Q'(B_{N+1,2})\big] \big/ \big((N+1)(N+2)\big)$.

Equation (5) is standard (cf. Khalaf et al., 2025): BoN sees the reward model only through Q, so one pool gives V_N at every budget up to its size, invariant to strictly increasing recalibration. By (6), the $(N+1)$-st sample earns exactly the increments of Q it jumps over, essentially near quantile $1 - 1/N$ (Appendix A.1).

3.2 CALIBRATION SHAPE CONSTRAINS BON

For a signed measure κ on $(0,1)$, $S^-(\kappa)$ is the smallest k such that $(0,1)$ splits into $k+1$ consecutive intervals on which κ alternates in sign (for piecewise monotone Q, $S^-(\mathrm{d}Q)$ counts turning points); for sequences, S^- counts sign changes ignoring zeros.

Theorem 3.2 (Shape transfer). *Let S be continuous and Q of bounded variation. Then*

$$S^-\big((\Delta V_N)_{N\geq 1}\big) \;\leq\; S^-(\mathrm{d}Q), \tag{7}$$

with identical sign order under equality. Consequently:

- *(a) if Q is nondecreasing, so is V_N;*
- *(b) if Q is single-peaked (nondecreasing on $[0,u_\star]$, nonincreasing on $[u_\star,1]$), so is $N\mapsto V_N$, with a finite peak $N_\star$ if Q strictly decreases on some $(u_1,1)$;*
- *(c) $N\mapsto V_N$ has no more local extrema than Q.*

Under $t=-\log u$, ΔV_N is a Laplace transform of $\mathrm{d}Q$, so by Descartes' rule of signs BoN can remove turning points of Q but never create them (Appendix A.3). For single-peaked Q, stopping at the first $\Delta V_N \leq 0$ is globally optimal.

The same shape constraint holds for finite pools: every turn in the without-replacement BoN curve requires a turn in the tied-rank profile (Proposition A.1, Appendix A.2).

If the top of Q returns to the prompt mean after some finite budget beats random, BoN has a finite maximizer (Corollary A.3, Appendix A.4).

3.3 ADDITIVE NOISE AND THE PRIOR TILT

Consider $S=R+\varepsilon$, with $R\in[r_{\min},r_{\max}]$ independent of ε, whose positive density p_ε has locally absolutely continuous logarithm (e.g. C^1, or Laplace), and *hazard slope* $\psi_\varepsilon(t)=-\frac{\mathrm{d}}{\mathrm{d}t}\log p_\varepsilon(t)$ (a.e.). The *exponential tilt* of the prior,

$$m(\lambda)=\frac{\mathbb{E}[R\,e^{\lambda R}]}{\mathbb{E}[e^{\lambda R}]}=\frac{\mathrm{d}}{\mathrm{d}\lambda}\log\mathbb{E}\big[e^{\lambda R}\big],\qquad \lambda\in[0,\infty], \tag{8}$$

is nondecreasing with $m(0)=\mu$, $m(\infty)=r_{\max}$. The posterior at score s is the prior tilted by the average hazard slope over $[s-r,s]$; the posterior formula and binary specialization are in Appendix A.5 (Lemma A.4).

Log-concave additive errors make G, Q, and V_N nondecreasing for every prior. This follows from the covariance identity (32) in Appendix A.5; for Gaussian error, $G'(s)=\mathrm{Var}(R\mid S=s)/\sigma^2$.

Under additive noise, BoN can thus turn down only if ψ_ε decreases somewhere, i.e. under heavy tails; log-concavity is also necessary for monotonicity under every two-point prior (Appendix A.5).

Theorem 3.3 (Hazard trichotomy). *Assume $\lambda_\star := \lim_{t\to\infty}\psi_\varepsilon(t)\in[0,\infty]$ exists and $\Pr(R>r_{\max}-\delta)>0$ for every $\delta>0$. Then*

$$\lim_{s\to\infty}G(s)=\lim_{u\uparrow 1}Q(u)=m(\lambda_\star)\,: \tag{9}$$

an extreme score reveals the prior tilted by the asymptotic hazard slope. Explicitly,

- diverging influence, $\lambda_\star=\infty$ *(Gaussian, generalised Gaussian $e^{-|t|^\beta}$, $\beta>1$): $Q(u)\to r_{\max}$;*
- constant influence, $\lambda_\star\in(0,\infty)$ *(Laplace with scale b, $\lambda_\star=1/b$; logistic; $t^k e^{-\lambda_\star t}$; $\beta=1$): $Q(u)\to m(\lambda_\star)\in(\mu,r_{\max})$, a plateau;*
- vanishing influence, $\lambda_\star=0$ *(Student-t, Pareto, log-normal, $\beta<1$, and mixtures with such a component, Proposition A.9): $Q(u)\to\mu$, so Corollary A.3 applies once a finite budget beats random.*

Table 1: Tail regimes for bounded rewards. The first three rows assume independent additive error (Theorem 3.3); Gaussian and Laplace monotonicity follows from Proposition A.5. The last row assumes a unique scale maximizer (Proposition 3.4). Rates and further families are given in Appendix A.7.

Error model	$\lambda_\star$	Limiting utility	Budget behavior
Gaussian	∞	$r_{\max}$	Nondecreasing
Laplace (scale b)	$1/b$	$m(1/b)$	Nondecreasing; plateau
Student-t_ν	0	μ	Returns to prior
Heteroskedastic Gaussian	—	r_σ	Can end below μ

$\lambda_\star$ is the log-likelihood price of the extra error needed to explain an extreme score without a high reward (Appendix A.6). As it is also the limiting slope of $-\log \bar{F}_\varepsilon$, a log-survival plot of incorrect responses' scores identifies the regime, which bulk metrics such as AUC cannot.

Rates of approach are given in Proposition A.7 (Appendix A.7).

3.4 Heteroskedastic noise: scale beats shift

Neural reward-model errors may depend on R and the prompt; the simplest such departure already reverses the Gaussian conclusion.

Proposition 3.4 (Scale beats shift)**.** *Let $\varepsilon \mid R = r \sim \mathcal{N}\big(0, \sigma^2(r)\big)$ with $\sigma \geq 0$ continuous on $[r_{\min}, r_{\max}]$, $\sigma(R) > 0$ a.s., and a unique maximiser r_σ in the support of R. Then*

$$\lim_{s\to\infty} G(s) = r_\sigma, \qquad \textit{hence} \qquad \lim_{N\to\infty} V_N = r_\sigma, \tag{10}$$

which lies below random selection *whenever $r_\sigma < \mu$.*

The scale enters the log-likelihood at order s^2, the reward only at order s (Appendix A.8). E.g. $S = R(1+\varepsilon)$ has $\sigma(r) \propto |r|$ and drives BoN to $r_{\min}$ when $|r_{\min}| > r_{\max}$; our conditional-variance and multiplicative stress tests probe these violations.

The model-free binary-reward characterization through $\Lambda = p_1/p_0$ is given in Proposition A.8 (Appendix A.8).

3.5 Tail location sets the relevant budget

Where V_N turns is set by how far into the top quantiles the offending tail sits.

Lemma 3.5 (Invisible tails)**.** *Let Q, Q' take values in $[r_{\min}, r_{\max}]$ and agree on $[0, 1-\delta]$. Then for every N,*

$$|V_N - V'_N| \leq (r_{\max} - r_{\min})\big(1 - (1-\delta)^N\big) \leq (r_{\max} - r_{\min})\, N\delta. \tag{11}$$

Equation (11) shows that budget N reads Q essentially on the top $1/N$ band, so the top δ of Q is invisible until $N \sim 1/\delta$ (Appendix A.9).

A heavy-tailed contaminating component can therefore preserve the observed curve yet force an eventual return to the prior. No finite pool certifies monotonicity beyond its size (Proposition A.9 and Corollary A.10, Appendix A.9).

3.6 Prescriptions

Before choosing N, identify $\lambda_\star$ from the log-survival of low-reward scores; when Q is single-peaked, stop at the first budget where the estimated ΔV_N (Proposition A.1) is nonpositive. Never extrapolate a rising curve beyond the pool size (Corollary A.10): in the plateau and heavy-tailed regimes, cap N near the empirical saturation point or peak, and pursue the Gaussian regime's $\sqrt{\log N}$ gains only when samples are cheap.

4 EXPERIMENTS

We test three implications of Section 3: error tails determine limiting utility, calibration localizes finite-budget failures, and a held-out tail audit can guide a budget cap. Detailed settings, estimators, and uncertainty procedures are in Appendix B.

4.1 EXPERIMENTAL SETUP

Synthetic pools use a bounded Uniform prior, Gaussian, Laplace, and Student-t errors, six correlations, and budgets up to 512; prior and multiplicative stress tests give 162 cells, with 12 additional heteroskedastic cells. Each cell uses 20,000 candidate sets. For controlled tests, Qwen2.5-7B-Instruct (Qwen Team, 2024) generates 256 answers per problem on MATH-500 and GSM8K (Hendrycks et al., 2021; Cobbe et al., 2021); four reward models score identical pools under QA and answer-only inputs, with masked and shuffled MATH controls for two models. PPE (Frick et al., 2024) supplies 32 responses per prompt from four base policies on five tasks, giving 65 learned task–RM cells and five verbosity stress cells. We compute exact without-replacement BoN curves with tied blocks averaged. A downturn requires a positive simultaneous 95% interval for the selected interior peak-to-endpoint drop. In PPE, negative top-four tail excess on a calibration fold triggers a calibration-selected budget cap; the opposite fold evaluates it. Policy tests use BH correction over all 70 cells; diagnostic intervals bootstrap cells, preserving both folds.

4.2 SYNTHETIC TESTS OF TAIL REGIME AND BUDGET SCALE

Figure 1 shows all 162 cells. The bounded Uniform primary prior satisfies the assumptions of Theorem 3.3, so every additive error row has a predicted regime, and each is borne out.

Light-tailed additive errors. Gaussian cells are monotone over $N \leq 512$, and by Proposition A.5 this holds at every budget. Laplace never produces a downturn: five cells are inconclusive plateaus and one a boundary maximum. Since Proposition A.5 rules out a true interior peak for Laplace error, its inconclusive interior peaks are Monte Carlo fluctuations on the plateau $m(\lambda_\star)$ of Theorem 3.3, which Proposition A.7(ii) says is reached geometrically.

Heavy-tailed additive errors. The $t_{2.5}$, t_3, t_4, and t_6 rows have supported downturns at all six correlations, as the vanishing-influence regime and Corollary A.3 require. t_{10} lies at the finite-window transition and t_{30} is monotone. Both belong to the same regime, but their return has rate $N^{-1/\nu}$ (Proposition A.7(i)): at $N = 512$ this factor is still ≈ 0.81 for $\nu = 30$, against ≈ 0.08 for $\nu = 2.5$, so the turn lies beyond the window (Corollary A.10). Overall, 24/54 primary cells are downturns, 20 are monotone, and 10 are plateau/boundary.

Increasing correlation shifts and attenuates the heavy-tailed downturn without changing its regime (Lemma 3.5, Theorem 3.3); detailed peak locations and intervals are in Appendix B.1.

Prior and multiplicative stress tests are reported in Appendix B.3; they clarify why the bounded-prior and unique-scale-maximizer assumptions matter.

Heteroskedastic scale. Proposition 3.4 predicts $V_N \to \sqrt{3}$ for $\beta = +1$ and $V_N \to -\sqrt{3} < \mu = 0$ for $\beta = -1$. All six increasing-scale cells peak at $N = 512$, whereas four decreasing-scale cells have supported downturns. The largest drop, 1.939, already takes BoN below random selection; rank accuracy still reaches 0.828 in a failing cell. Full contrasts are in Appendix B.2.

4.3 THE CONTROLLED MATH FAILURE AND FORMAT CONTROLS

Across the original 16 task–RM–format conditions, no QA curve has a statistically supported downturn. The clear original failure is Tulu on MATH-500 answer-only: its peak moves from $N = 64$ under QA to $N = 4$, followed by a 0.0597 correctness drop at $N = 256$ with simultaneous 95% interval $[0.0316, 0.0878]$. As Theorem 3.2 requires, the downturn is matched by a turn in the calibration curve: the top-1% excess changes sign from $E_1 = +.054$ under QA to $E_1 = -.031$ (Table 2), so the top of Q falls below the prompt mean. GSM8K shows no supported downturn in either original format, so information removal is not by itself sufficient; as for every monotone curve here, this concerns only the window $N \leq 256$ (Corollary A.10).

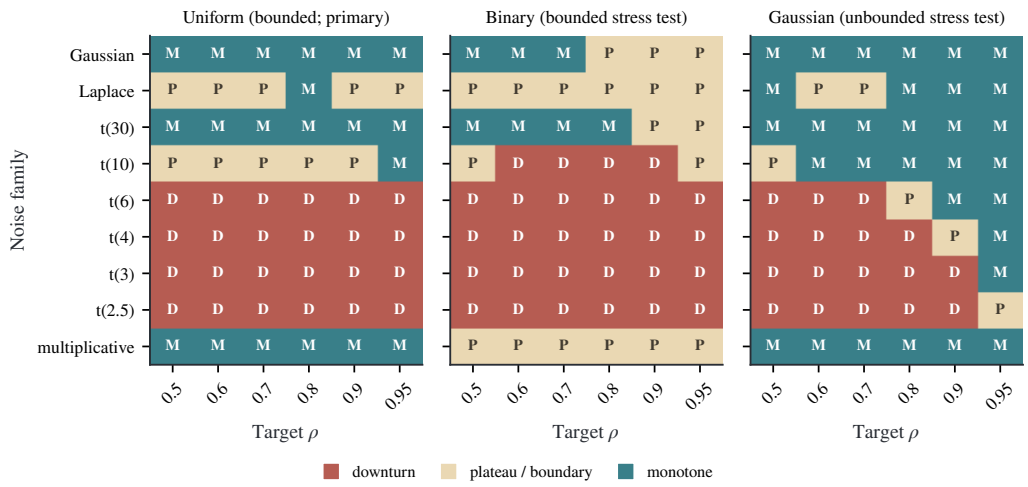

Figure 1: **Synthetic phase maps.** D denotes an interior peak whose peak-to-$N = 512$ simultaneous max-t interval is above zero; P denotes an inconclusive interior or saturated boundary; M denotes a maximum at 512. The bounded Uniform map is primary. Prior shape changes finite-window boundaries but preserves the broad separation between light and heavy additive tails.

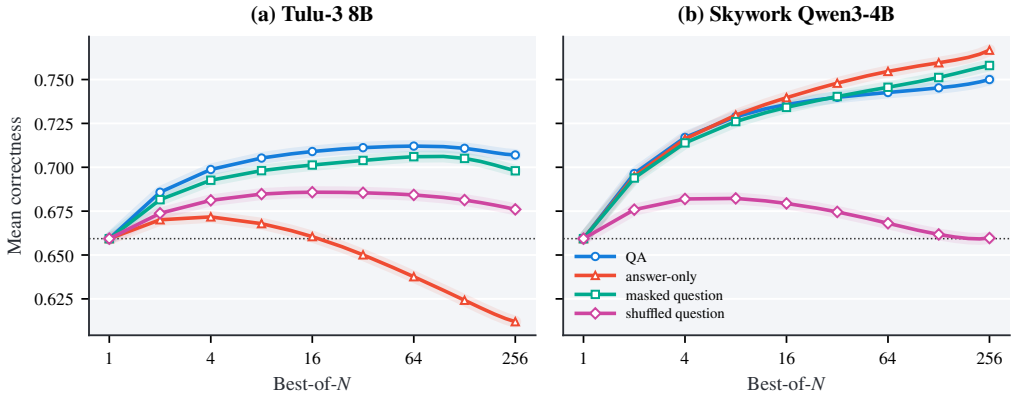

Figure 2: **MATH-500 input-format controls on a fixed candidate pool.** Each panel holds the reward model and all 500×256 answers fixed while varying only the score input: QA (blue circles), answer-only (orange-red triangles), masked question (teal squares), or deterministically shuffled question (magenta diamonds). The thin horizontal dotted line is base correctness.

Section 3 identifies the change in Q but not its cause, and the format controls address the cause. The failure does not survive a role-preserving missing-question control: Tulu masked-question peaks at $N = 96$ with a nonsignificant 0.0082 drop $[-0.0069, 0.0232]$ and positive top-1% excess. A shuffled question reduces rank accuracy and moves the empirical peak to 16, but its 0.0098 drop $[-0.0143, 0.0339]$ is also unsupported. Thus the answer-only result cannot be attributed to missing question semantics alone, and the chat role/template change is an important confound in that condition. Skywork is monotone when masked. When shuffled, its top-1% excess falls to $+.005$, so the top of Q has nearly returned to the prompt mean, the situation of Corollary A.3; the resulting peak at 8 is unsupported, providing a second model-level negative control.

Exact tied-rank budget estimates and smoothing sensitivity are reported in Appendix D.1. The answer-only optimum remains near $N = 3$–4 across visual estimators; unsupported conditions are less stable.

Table 2: MATH-500 input-format controls on the same 500×256 candidate pool. E_1 is top-1% tail excess; drop intervals are simultaneous within each condition. — indicates that the empirical maximum is at $N = 256$, so there is no selected interior peak contrast or corresponding interval.

RM	Input	Rank acc.	E_1	Peak N	Drop [95% CI]
Tulu-3 8B	QA	.741	+.054	64	.0051 [-.0118,.0220]
	answer-only	.581	-.031	4	.0597 [.0316,.0878]
	masked question	.698	+.049	96	.0082 [-.0069,.0232]
	shuffled question	.613	+.023	16	.0098 [-.0143,.0339]
Skywork Qwen3-4B	QA	.812	+.084	256	—
	answer-only	.799	+.101	256	—
	masked question	.792	+.089	256	—
	shuffled question	.639	+.005	8	.0226 [-.0028,.0480]

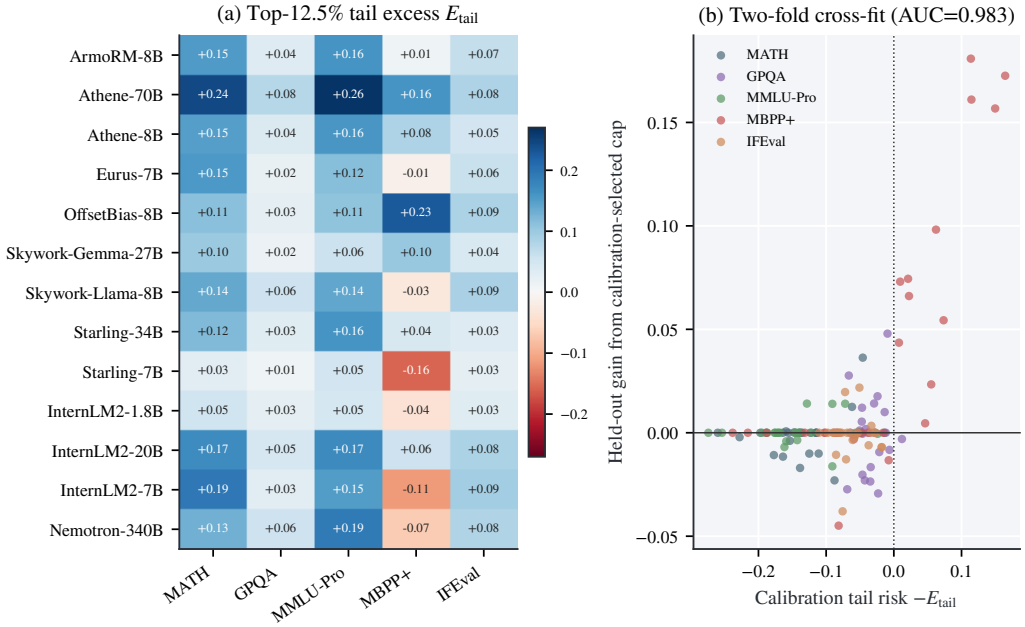

Figure 3: **External PPE validation.** (a) Top-12.5% tail excess for 13 learned reward models across five tasks. Negative cells concentrate on MBPP+. (b) Calibration-fold tail risk versus held-out gain from the calibration-selected cap, with two folds per learned cell. Cell-clustered bootstrap intervals are reported in Table 3.

4.4 External validation: failures concentrate in code

Figure 3a shows all 65 learned-model cells. The six negative top-12.5% tail excesses all occur on MBPP+, and the other four tasks have none. By Proposition A.8, a negative E_{tail} means that incorrect programs outweigh correct ones in the top eighth of scores ($\Lambda < 1$ there). This public pool thus identifies a code-generation failure domain, not a general upper-tail failure of current reward models.

The negative-tail cells show the two sign patterns that Proposition A.1 allows when the rank profile turns once. If the profile first rises and then falls, the increments go $+$ then $-$, giving an interior peak. Eurus-RM-7B is the one learned cell satisfying our strict interior-downturn definition: it peaks at $N = 2$ and loses 0.059 correctness by $N = 32$, with simultaneous interval $[0.018, 0.100]$. If the profile falls from the start, every increment is nonpositive and the peak is the boundary $N = 1$. Three other negative-tail models are already harmful at $N > 1$ in this way: Starling-7B (drop 0.165), InternLM2-7B (0.171), and Nemotron-340B (0.076). Keeping the two patterns apart prevents us from hiding monotone harmful scaling behind the term "downturn."

Table 3: Diagnostic validation on 65 learned PPE cells. Intervals resample task–RM cells and keep both cross-fit folds together. Differences are tail risk minus the comparator.

Predictor	ROC AUC [95% CI]	AP [95% CI]	ρ_s [95% CI]
$-E_{\text{tail}}$	.983 [.954,.998]	.825 [.449,.973]	.254 [.020,.459]
Low rank accuracy	.972 [.927,.995]	.747 [.337,.953]	.258 [.023,.453]
Score-tail spread	.355 [.133,.589]	.057 [.026,.117]	-.107 [-.292,.074]
Tail − rank	.012 [-.003,.042]	.079 [-.015,.215]	-.004 [-.043,.039]
Tail − spread	.628 [.384,.858]	.769 [.382,.915]	.361 [.060,.656]

Policy-stratified results and descriptive features of the bad code tail are reported in Appendix E; they show that risk depends on the policy–RM pairing and do not establish a causal shortcut.

4.5 Held-out budget caps and diagnostic comparisons

Here we test the prescription of Section 3.6: read the top of Q, then cap N if it has turned. The two-fold analysis contains 130 learned-cell folds, nine of which have a calibration-selected cap with a positive held-out interval: eight are MBPP+ and one is GPQA, for prevalence 0.069. Tail risk predicts this rare event with ROC AUC 0.983 [0.954, 0.998] and average precision (AP) 0.825 [0.449, 0.973], versus the prevalence baseline.

Rank accuracy is also strong (AUC/AP $0.972/0.747$); paired tail-minus-rank intervals include zero for AUC and AP (Table 3). The broad AP intervals expose the rarity and reject any claim that bulk ranking quality cannot predict the problem in practice; the connection to non-identifiability is discussed in Appendix C.1.

What tail excess adds over rank accuracy is structural: it localizes the failure to the budgets that read it (Lemma 3.5), it links exactly to V_N (Equation (5)), and it is invariant to monotone rescaling of scores. A score-only tail spread performs poorly (AUC 0.355), as Proposition A.8 explains, because the marginal score shape does not determine Λ.

Relative to BoN-32, the cap and rank-Soft-BoN gain 0.0085 [0.0019, 0.0168] and 0.0068 [-0.0003, 0.0156] over all 65 cells (paired difference 0.0017 [-0.0003, 0.0036]). On six negative-tail cells they gain 0.0924 and 0.0916: the hard rule is competitive, not materially superior.

Eurus, Starling-7B, InternLM2-7B, and Nemotron retain BH-corrected cap gains; Skywork-Llama and InternLM2-1.8B have positive point gains only. Unflagged folds remain at $N = 32$; consistent with Corollary A.10, this certifies nothing beyond the pool size (cell details: Appendix Table 5).

Tail-width and tie-order checks preserve the main pattern (Appendix C). Missing endpoint normalization and policy likelihoods prevent controlled Best-of-Tails and ITP comparisons; rank-Soft-BoN is the feasible same-pool baseline.

5 Conclusion

We studied inference-time reward hacking through the tail calibration of best-of-N selection. The quantile representation explains how calibration shape governs whether BoN can decline and how the location of tail defects determines the relevant budget scale. For bounded rewards with independent additive errors, the error tail distinguishes continued improvement, saturation, and return to the prior mean. Synthetic and controlled experiments support this account. In PPE, negative tail excess concentrates in code generation, and cross-fitted budget caps yield held-out gains competitive with rank-Soft-BoN. Tail calibration thus complements rank accuracy with a direct connection between reward-model evaluation and search budgets. With labeled candidate pools and independent validation, this connection supports task–policy–reward-model-specific budget choices grounded in true utility over the tested range.

AI USE STATEMENT

Generative AI tools assisted with literature synthesis, manuscript drafting and editing, code organization, figure preparation, and LaTeX formatting. Experimental commands, retained artifacts, numerical claims, references, and rendered output were checked against repository or primary-source evidence. No model-generated numerical result was accepted without a corresponding analysis artifact. The authors are responsible for the final content, including text, claims, and artifacts produced with generative AI assistance.

ETHICS STATEMENT

The work uses no personal data and does not target deployed systems. Because a tail diagnostic could help optimize against a weak reward model, we frame it as a pre-deployment audit requiring independent ground-truth verification.

REPRODUCIBILITY STATEMENT

The assumptions and proof details are provided in Appendix A; data processing, estimators, and experimental settings are described in Appendix B, with sensitivity analyses in Appendix C. Complete controlled results and additional PPE results appear in Appendices D and E.

REFERENCES

Udbhav Bamba, Rishabh Tiwari, Heng Yang, Kurt Keutzer, and Amir Gholaminejad. Reward under attack: Evaluating the sensitivity of process reward models. In *ICML 2025 Workshop on AI for Math*, 2025. URL https://icml.cc/virtual/2025/52443.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024. URL https://arxiv.org/abs/2401.01879.

Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*, volume 27. Cambridge university press Cambridge, 1987.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, and Reiichiro Nakano. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023a. URL https://arxiv.org/abs/2310.02743.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023b. URL https://arxiv.org/abs/2310.02743.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023. URL https://arxiv.org/abs/2312.09244.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. *arXiv preprint arXiv:2410.14872*, 2024. URL https://arxiv.org/abs/2410.14872.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

Hsiang Hsu, Eric Lei, and Chun-Fu Chen. Best-of-tails: Bridging optimism and pessimism in inference-time alignment. *arXiv preprint arXiv:2603.06797*, 2026. URL https://arxiv.org/abs/2603.06797.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *arXiv preprint arXiv:2503.21878*, 2025. URL https://arxiv.org/abs/2503.21878.

Samuel Karlin. *Total Positivity*, volume 1. Stanford University Press, Stanford, California, 1968.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-time reward hacking in large language models. *arXiv preprint arXiv:2506.19248*, 2025. URL https://arxiv.org/abs/2506.19248.

Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. Catastrophic goodhart: Regularizing rlhf with kl divergence does not mitigate heavy-tailed reward misspecification. *arXiv preprint arXiv:2407.14503*, 2024. URL https://arxiv.org/abs/2407.14503.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024a. URL https://arxiv.org/abs/2411.15124.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, L. J. Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024b. URL https://arxiv.org/abs/2403.13787.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024c. URL https://arxiv.org/abs/2403.13787.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025. URL https://arxiv.org/abs/2507.01352.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. Rm-bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*, 2024. URL https://arxiv.org/abs/2410.16184.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*, 2025. URL https://arxiv.org/abs/2506.01937.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. URL https://arxiv.org/abs/2112.09332.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325*, 2020. URL https://arxiv.org/abs/2009.01325.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, et al. Rmb: Comprehensively benchmarking reward models in llm alignment. *arXiv preprint arXiv:2410.09893*, 2024. URL https://arxiv.org/abs/2410.09893.

A Proofs for Section 3

The results of Section 3 rest on two mechanisms, and the proofs are organised around them.

- *Order statistics and total positivity* (Appendices A.1–A.4). BoN reads the calibration curve at the random quantile $U_{(N)}$, and one extra sample reads its increments through the kernel $u^N(1-u)$. Because this kernel is totally positive in (N, u), aggregation over N cannot create sign changes.
- *Explaining away at extreme scores* (Appendices A.5–A.8). The posterior over reward at a large score is a tilt of the prior whose rate is the hazard slope of the error (additive noise) or, more generally, the growth rate of the class-conditional log-likelihoods.

Appendix A.9 combines the two to separate whether a failure occurs from where it becomes visible.

Conventions. (C1) *Versions.* Q is a conditional expectation and is defined only up to Lebesgue-null sets of u. Every quantity we compute (V_N, ΔV_N) integrates Q against a law with a density on $(0,1)$, so it does not depend on the version. When Q has bounded variation we use its right-continuous version on $(0,1)$; then $\mathrm{d}Q((a,b]) = Q(b) - Q(a)$, the limits $Q(0^+)$ and $Q(1^-)$ exist, and $|\mathrm{d}Q|$ has finite total mass $\mathrm{TV}(Q)$. (C2) *Sign changes.* For a signed measure κ on $(0,1)$ with $S^-(\kappa) = k < \infty$ we call the $k+1$ intervals of a minimal partition the *sign blocks* of κ; by minimality each block carries nonzero mass and consecutive blocks carry opposite signs. (C3) *Total positivity* (Karlin, 1968). A kernel $K(x,y)$ on ordered sets $X \times Y$ is totally positive (TP) if every minor $\det[K(x_i,y_j)]_{i,j\le m}$ with $x_1 < \cdots < x_m$, $y_1 < \cdots < y_m$ is nonnegative, and strictly totally positive (STP) if every such minor is positive. (C4) *Beta facts.* $U_{(N)} \sim \mathrm{Beta}(N,1)$ and $1 - U_{(N)} \sim \mathrm{Beta}(1,N)$. $B_{N+1,2}$ has mode $N/(N+1)$, mean $1-2/(N+3)$ and standard deviation $\sqrt{2(N+1)}/\big((N+3)\sqrt{N+4}\big) = \frac{\sqrt{2}}{N}(1+O(1/N))$.

A.1 Proof of Proposition 3.1

Preliminaries. Because S is continuous, $\Pr(S_i = S_j) = 0$ for $i \neq j$, so ties have probability zero and the tie-breaking rule plays no role. The variables $U_i = F_S(S_i)$ are i.i.d. Uniform$[0,1]$, and $S_i > S_j$ implies $U_i \ge U_j$, with equality on a null event. Hence, almost surely, I_N is also the index of $U_{(N)}$. Nothing beyond the ranking of the scores is used below, which is the source of the invariance of every statement under strictly increasing transformations of S.

Proof of Equation (5). By exchangeability of the pairs (R_i, U_i),

$$V_N = \sum_{i=1}^{N} \mathbb{E}\big[R_i\,\mathbf{1}\{I_N = i\}\big] = N\,\mathbb{E}\big[R_1\,\mathbf{1}\{U_1 > U_j,\ 2 \le j \le N\}\big]. \tag{12}$$

The variables $U_2, \ldots, U_N$ are independent of (R_1, U_1), so conditioning on (R_1, U_1) gives $\Pr(U_1 > U_j\ \forall j \ge 2 \mid R_1, U_1) = U_1^{N-1}$ and

$$V_N = N\,\mathbb{E}\big[R_1 U_1^{N-1}\big] = N\,\mathbb{E}\big[\mathbb{E}[R_1 \mid U_1]\,U_1^{N-1}\big] = \int_0^1 Q(u)\,Nu^{N-1}\,\mathrm{d}u. \tag{13}$$

The second form follows from $\int_0^1 Nu^{N-1}\,\mathrm{d}u = 1$. Equation (13) says that the winner's reward has the law of a reward read at the random quantile $U_{(N)} \sim \mathrm{Beta}(N,1)$: the budget acts only by moving this reading point towards $u = 1$.

Proof of Equation (6)**: the crossing argument.** Realise the $(N+1)$-sample experiment by adding one fresh sample to the N-sample experiment, so that $U_{(N+1)} = \max(U_{(N)}, U_{N+1})$ with U_{N+1} independent of $U_{(N)}$. By (5) applied to N and $N+1$,

$$\Delta V_N = \mathbb{E}\big[Q(U_{(N+1)}) - Q(U_{(N)})\big] = \mathbb{E}\Big[\big(Q(U_{N+1}) - Q(U_{(N)})\big)\,\mathbf{1}\{U_{N+1} > U_{(N)}\}\Big]. \tag{14}$$

On the event $\{U_{N+1} > U_{(N)}\}$ the bracket equals $\mathrm{d}Q\big((U_{(N)}, U_{N+1}]\big) = \int_{(0,1)} \mathbf{1}\{U_{(N)} < u \le U_{N+1}\}\,\mathrm{d}Q(u)$. Since $|\mathrm{d}Q|$ is finite, Fubini's theorem applies and

$$\Delta V_N = \int_{(0,1)} \Pr\big(U_{(N)} < u \le U_{N+1}\big)\,\mathrm{d}Q(u) = \int_{(0,1)} u^N(1-u)\,\mathrm{d}Q(u). \tag{15}$$

The proof explains the kernel. A new sample changes the outcome only if it dethrones the incumbent, and it then earns the increments of Q over the quantile interval it jumps across. The kernel $u^N(1-u)$ is the probability that level u is jumped across: the incumbent is still below u (probability u^N, which is small unless u is close to 1) and the newcomer clears u (probability $1-u$, which vanishes at $u=1$). The product peaks at $u = N/(N+1)$, so the marginal sample can change the value only through the part of Q near the current record.

Second proof of Equation (6) **and the limit of** V_N**.** For Q of bounded variation and the continuous function $g(u) = u^N$, Lebesgue–Stieltjes integration by parts has no jump correction and gives $\int_{(0,1)} Q\,\mathrm{d}g = Q(1^-)g(1) - Q(0^+)g(0) - \int_{(0,1)} g\,\mathrm{d}Q$, that is,

$$V_N = Q(1^-) - \int_{(0,1)} u^N\,\mathrm{d}Q(u). \tag{16}$$

Differencing (16) in N gives (6) again. Since $u^N \to 0$ on $(0,1)$ and $|\mathrm{d}Q|$ is finite, dominated convergence in (16) gives $V_N \to Q(1^-)$: the BoN limit is the top of the calibration curve, and the remaining gap $\int u^N\,\mathrm{d}Q$ is a Laplace-type transform of $\mathrm{d}Q$ concentrated near $u=1$.

Beta form and the local question. If Q is absolutely continuous, $\mathrm{d}Q = Q'(u)\,\mathrm{d}u$ and $\int_0^1 u^N(1-u)Q'(u)\,\mathrm{d}u = \mathrm{B}(N+1,2)\,\mathbb{E}[Q'(B_{N+1,2})]$, where $\mathrm{B}(N+1,2) = \Gamma(N+1)\Gamma(2)/\Gamma(N+3) = 1/((N+1)(N+2))$. By (C4), $B_{N+1,2}$ sits at $1-\Theta(1/N)$ with spread $\asymp 1/N$, so $(N+1)(N+2)\,\Delta V_N$ is an average of the slope Q' over a window of width $\asymp 1/N$ around $1-1/N$; this is the "local question" of the main text.

Prompt averaging. If $V_N(x) = \int_0^1 Q_x(u)\,Nu^{N-1}\,\mathrm{d}u$ for each prompt x and the Q_x are uniformly bounded, Fubini's theorem gives $\mathbb{E}_x V_N(x) = \int_0^1 \bar{Q}(u)\,Nu^{N-1}\,\mathrm{d}u$ with $\bar{Q} = \mathbb{E}_x Q_x$. If moreover $\mathbb{E}_x \mathrm{TV}(Q_x) < \infty$, then $\mathrm{d}\bar{Q} = \mathbb{E}_x\,\mathrm{d}Q_x$ and (6) holds for the average. □

A.2 Finite pools: proof of Proposition A.1

Proposition A.1 (Finite pools)**.** *Let $\bar{r}_{(1)}, \dots, \bar{r}_{(K)}$ be the prompt-averaged rewards at ascending within-prompt score ranks (tied blocks averaged, pool size K), and $\hat{V}_N$ the without-replacement BoN curve. Then*

$$\hat{V}_{N+1} - \hat{V}_N = \sum_{j=1}^{K-1} k_N(j)\,\big(\bar{r}_{(j+1)} - \bar{r}_{(j)}\big), \qquad k_N(j) = \frac{\binom{j}{N}(K-j)}{\binom{K}{N}(K-N)} \ge 0, \tag{17}$$

where $k_N(j)$ is the probability that the best of the first N draws has rank at most j and the next draw rank above j; moreover $S^-\big((\hat{V}_{N+1} - \hat{V}_N)_{N<K}\big) \le S^-\big((\bar{r}_{(j+1)} - \bar{r}_{(j)})_{j<K}\big)$.

Every turn in a reported BoN curve is thus witnessed by a turn in the rank profile (Appendix A.2).

Fix a prompt with K responses and let $r_{(1)}, \dots, r_{(K)}$ be their rewards ordered by ascending score, with tied blocks averaged. BoN at budget $N \le K$ draws N of the K responses uniformly without replacement and returns the reward of the highest-ranked draw.

Lemma A.2 (Finite-pool identities). *Let $d_j = r_{(j+1)} - r_{(j)}$ and $W_N(j) = \binom{j}{N}/\binom{K}{N}$ (with $\binom{j}{N} = 0$ for $j < N$). Then*

(i) $\hat{V}_N = \sum_{j=N}^{K} \binom{j-1}{N-1}\binom{K}{N}^{-1} r_{(j)} = r_{(K)} - \sum_{j=1}^{K-1} W_N(j)\, d_j$;

(ii) for $N \le K-1$, $\hat{V}_{N+1} - \hat{V}_N = \sum_{j=1}^{K-1} k_N(j)\, d_j$ *with* $k_N(j) = W_N(j)\,\frac{K-j}{K-N} \ge 0$;

(iii) $S^-\big((\hat{V}_{N+1} - \hat{V}_N)_{N=1}^{K-1}\big) \le S^-\big((d_j)_{j=1}^{K-1}\big)$.

All three statements continue to hold for prompt averages with a common K, with $r_{(j)}$ replaced by $\bar{r}_{(j)}$.

Proof. (i) The top draw has rank j exactly when it is j and the other $N-1$ draws come from the $j-1$ responses below it, which has probability $\binom{j-1}{N-1}/\binom{K}{N}$. With uniform tie breaking the expected selected reward is linear in the reward vector, and averaging over the random order inside a tied block replaces each reward of the block by the block average. The second form is summation by parts, using $W_N(j) - W_N(j-1) = \binom{j-1}{N-1}/\binom{K}{N}$, $W_N(K) = 1$ and $W_N(0) = 0$.

(ii) By (i), $\hat{V}_{N+1} - \hat{V}_N = \sum_j \big(W_N(j) - W_{N+1}(j)\big) d_j$. Draw the $N+1$ responses one at a time. $W_N(j)$ is the probability that the first N all have rank at most j, and $W_{N+1}(j)$ is the probability that all $N+1$ do, so the difference is the probability that the first N have rank at most j and the next one has rank above j. Given the first event, the next draw is uniform over the $K-N$ remaining responses, of which $K-j$ lie above j; hence $W_N(j) - W_{N+1}(j) = W_N(j)(K-j)/(K-N)$. This is the exact finite-pool counterpart of the crossing kernel $u^N(1-u)$, to which it converges when $j/K \to u$ with N fixed.

(iii) As a matrix indexed by (N, j), $k_N(j) = a_N\, b_j \binom{j}{N}$ with $a_N = 1/\big(\binom{K}{N}(K-N)\big) > 0$ and $b_j = K - j > 0$ for $j \le K-1$. Multiplying rows and columns by positive numbers does not change the sign of any minor, so k is TP whenever the Pascal kernel $(N, j) \mapsto \binom{j}{N}$ is. The Pascal matrix is totally nonnegative: each minor counts families of non-intersecting lattice paths by the Lindström–Gessel–Viennot lemma. The variation-diminishing property of TP kernels (Karlin, 1968, Ch. 5, Thm. 3.1) then gives (iii). Prompt averaging is linear with a common kernel, which proves the last claim. □

The lemma makes the diagnostic of Theorem 3.2 exact at every sample size. A turn in the reported BoN curve must be matched by a turn in the reported rank profile, whatever the sampling noise, because the curve is a TP transform of the profile's increments. Noise can only add sign changes to the profile; it cannot hide the witness of a turn.

A.3 Proof of Theorem 3.2

We give a self-contained proof in three steps. Step 2 is Laguerre's proof of Descartes' rule of signs, transplanted to the kernel $u^\lambda(1-u)$; an alternative is to note that $(N, u) \mapsto u^N(1-u) = (1-u)e^{N \log u}$ is STP, because e^{xy} is STP and $\log u$ is increasing, and to invoke the variation-diminishing theorem (Karlin, 1968, Ch. 5, Thm. 3.1).

Let $\mathcal{M}$ be the class of signed measures κ on $(0,1)$ with $\int u^\lambda(1-u)\,|\kappa|(\mathrm{d}u) < \infty$ for every $\lambda > 0$, and for $\kappa \in \mathcal{M}$ put

$$\Phi_\kappa(\lambda) = \int_{(0,1)} u^\lambda (1-u)\, \kappa(\mathrm{d}u), \qquad \lambda > 0. \tag{18}$$

Every finite signed measure, in particular $\mathrm{d}Q$, lies in $\mathcal{M}$, and by Equation (6) $\Delta V_N = \Phi_{\mathrm{d}Q}(N)$. The idea is to embed the sequence (ΔV_N) in the real-analytic function $\Phi_{\mathrm{d}Q}$ and count zeros of the function instead of sign changes of the sequence.

Step 1: sign changes of the sequence are zeros of the function. Φ_κ is continuous on $(0, \infty)$ by dominated convergence. If $\Phi_\kappa(N_1)$ and $\Phi_\kappa(N_2)$ have opposite signs, with $N_1 < N_2$ and $\Phi_\kappa(N) = 0$ for the integers strictly between them, then Φ_κ vanishes somewhere in the open interval (N_1, N_2),

either at one of those integers or by the intermediate value theorem. Distinct sign changes of the sequence give zeros in disjoint intervals, so

$$S^-\big((\Phi_\kappa(N))_{N\geq 1}\big) \leq \#\{\lambda > 0 : \Phi_\kappa(\lambda) = 0\}. \tag{19}$$

Step 2: the function has at most $S^-(\kappa)$ zeros. We show by induction on k that if $\kappa \in \mathcal{M}$ is nonzero and $S^-(\kappa) \leq k$, then Φ_κ has at most k distinct zeros on $(0,\infty)$. For $k = 0$, κ or $-\kappa$ is a nonzero nonnegative measure and the integrand $u^\lambda(1-u)$ is strictly positive on $(0,1)$, so Φ_κ never vanishes. For the induction step let $S^-(\kappa) = k \geq 1$ and let $c \in (0,1)$ be the point that separates the last two sign blocks. Put

$$\Psi(\lambda) = c^{-\lambda}\Phi_\kappa(\lambda) = \int (u/c)^\lambda (1-u)\,\kappa(\mathrm{d}u), \tag{20}$$

which has the same zeros as Φ_κ. For λ in a compact subinterval $[\lambda_0,\lambda_1] \subset (0,\infty)$,

$$\big|\partial_\lambda (u/c)^\lambda\big| = |\log(u/c)|\,(u/c)^\lambda \leq C\,u^{\lambda_0/2} \qquad \text{on } (0,1), \tag{21}$$

so differentiation under the integral is permitted and

$$\Psi'(\lambda) = c^{-\lambda}\,\Phi_{\kappa'}(\lambda), \qquad \kappa'(\mathrm{d}u) = \log(u/c)\,\kappa(\mathrm{d}u). \tag{22}$$

The same bound shows $\kappa' \in \mathcal{M}$. The factor $\log(u/c)$ is negative on $(0,c)$ and positive on $(c,1)$. It therefore flips the sign of every block to the left of c and keeps the last block, so the last two blocks now carry the same sign and merge, while all other sign changes survive. Hence $S^-(\kappa') \leq k-1$. Moreover $\kappa' \neq 0$, because κ, having $k \geq 1$ sign changes, is not concentrated at the single point c. By the induction hypothesis Ψ' has at most $k-1$ zeros. If Φ_κ, and hence Ψ, had $k+1$ distinct zeros, Rolle's theorem would give k distinct zeros of Ψ', a contradiction.

Multiplying by $\log(u/c)$ is the familiar "multiply by $e^{c\lambda}$ and differentiate" step of Descartes' rule, written in the variable $t = -\log u$: each differentiation removes exactly one sign change, and Rolle's theorem says it removes at most one zero. Steps 1 and 2 give $S^-\big((\Delta V_N)_{N\geq 1}\big) \leq S^-(\mathrm{d}Q)$, which is the Laplace-transform version of the same sign-change bound.

Step 3: the sign at large N, and equality. Let $S^-(\mathrm{d}Q) = k$, let the last sign block be $(c_k, 1)$, and let $\epsilon_{k+1} \in \{-1,+1\}$ be its sign. Split $\Phi_{\mathrm{d}Q}(\lambda)$ into the integral over $(0,c_k]$, bounded in absolute value by $c_k^\lambda\,\mathrm{TV}(Q)$, and the integral over $(c_k,1)$, which has sign ϵ_{k+1}. The last block has nonzero mass, so there are $c_k < v < v' < 1$ with $|\mathrm{d}Q|((v,v')) > 0$, and the second integral is at least $v^\lambda(1-v')\,|\mathrm{d}Q|((v,v'))$ in absolute value. Since $v > c_k$, the second integral dominates:

$$\operatorname{sign}\Delta V_N = \epsilon_{k+1} \qquad \text{for all large } N. \tag{23}$$

The tail of the BoN increments thus carries the sign of the top of the calibration curve. If $S^-(\Delta V) = k$ as well, the two sign sequences have the same length and the same last element, and alternating sequences are determined by these, so the patterns coincide.

Consequences. (a) If Q is nondecreasing, $\mathrm{d}Q \geq 0$ and $\Delta V_N \geq 0$ directly from Equation (6); no total positivity is needed. (c) The local extrema of $N \mapsto V_N$ are the sign changes of (ΔV_N), and for piecewise monotone Q the turning points of Q are the sign changes of $\mathrm{d}Q$.

(b) The single-peaked case has a one-line proof that shows the mechanism directly. Let $\mathrm{d}Q \geq 0$ on $(0,u_\star]$ and ≤ 0 on $(u_\star,1)$, with $u_\star > 0$; the case $u_\star = 0$ is trivial. Dividing by $u_\star^N$,

$$\frac{\Delta V_N}{u_\star^N} = \underbrace{\int_{(0,u_\star]} \Big(\frac{u}{u_\star}\Big)^N (1-u)\,\mathrm{d}Q(u)}_{A_N \geq 0} \;-\; \underbrace{\int_{(u_\star,1)} \Big(\frac{u}{u_\star}\Big)^N (1-u)\,|\mathrm{d}Q|(\mathrm{d}u)}_{B_N \geq 0}. \tag{24}$$

Since $(u/u_\star)^N$ is nonincreasing in N for $u \leq u_\star$ and nondecreasing for $u > u_\star$, A_N is nonincreasing and B_N is nondecreasing, so $A_N - B_N$ is nonincreasing in N. The set $\{N : \Delta V_N \leq 0\}$ is therefore an up-set: once one more sample stops helping, it never helps again, V_N is single-peaked, and $N_\star = \min\{N : \Delta V_N \leq 0\}$ is a global maximiser. This is the justification of the stopping rule in the main text. Each additional sample reweights the calibration curve towards its falling side.

For finiteness, strict decrease of Q on $(u_1,1)$ forces $u_\star \leq u_1$, because Q cannot be both nondecreasing and strictly decreasing on $(u_1,u_\star)$. Pick $u_1 < v < v' < 1$. Then $A_N \leq \mathrm{TV}(Q)$, while $B_N \geq (v/u_\star)^N(1-v')\,|\mathrm{d}Q|((v,v')) \to \infty$ because $|\mathrm{d}Q|((v,v')) = Q(v^+) - Q(v'^-) > 0$ and $v > u_\star$. So $\Delta V_N < 0$ for all large N and $N_\star < \infty$.

On heterogeneity. Averaging single-peaked per-prompt curves whose peaks lie at different quantiles can produce an averaged $\bar{Q}$ with several turning points, for instance when the prompts split into an easy group peaking early and a hard group peaking late. Theorem 3.2 then applies to $\bar{Q}$, which is the curve that the aggregate BoN value reads (Appendix A.1). Per-prompt or per-group rank profiles are the right tool when the averaged profile is not single-peaked. □

A.4 Proof of Corollary A.3, and the speed of return

Corollary A.3 (Return-to-prior certificate). *If Q is bounded, $\lim_{u\uparrow 1} Q(u) = \mu$, and $V_{N_0} > \mu$ for some N_0, then $\{V_N : N \in \mathbb{N}\}$ has a maximizer at finite N.*

Once the top of Q returns to the prompt mean, more samples are no longer a safe default (Appendix A.4).

Since $\Pr(U_{(N)} \le 1-\eta) = (1-\eta)^N \to 0$ for every $\eta > 0$ and $U_{(N)}$ is nondecreasing in N, $U_{(N)} \uparrow 1$ almost surely. As $Q(u) \to \mu$ for $u \uparrow 1$, $Q(U_{(N)}) \to \mu$ almost surely, and bounded convergence gives $V_N = \mathbb{E}[Q(U_{(N)})] \to \mu$. Choose N_1 with $V_N < (\mu + V_{N_0})/2$ for all $N > N_1$. Then $\sup_N V_N = \max_{N \le \max(N_0, N_1)} V_N$, a maximum over a finite set, which is attained. □

How fast the return happens. Writing $W = 1 - U_{(N)} \sim \mathrm{Beta}(1, N)$,

$$V_N - \mu = \mathbb{E}\big[E(1-W)\big], \qquad NW \xrightarrow{d} \mathrm{Exp}(1). \tag{25}$$

So $V_N - \mu$ is the excess of the rank profile at quantile $1 - \Theta(1/N)$, averaged with nearly exponential weights. If $w \mapsto E(1-w)$ is regularly varying at 0 with index $\alpha \ge 0$, the Abelian argument of Appendix A.7(i) gives

$$V_N - \mu \sim \Gamma(1+\alpha)\, E(1 - 1/N). \tag{26}$$

The speed of the return is therefore read directly off the top of $E(u)$. By (16) with $Q(1^-) = \mu$, we also have $V_N - \mu = \int u^N \, \mathrm{d}(-Q)(u)$: at large N, BoN beats random selection exactly when the top of Q approaches μ from above.

A.5 Additive noise: the tilt representation and Proposition A.5

Throughout Appendices A.5–A.7, $S = R + \varepsilon$ with $R \perp \varepsilon$, P_R is the law of R on $[r_{\min}, r_{\max}]$, $\bar{r} = \max(|r_{\min}|, |r_{\max}|)$, and $p_\varepsilon > 0$ has a locally absolutely continuous logarithm. Then $p_S(s) = \int p_\varepsilon(s-r)\, P_R(\mathrm{d}r) > 0$ for every s, $P(R \in \mathrm{d}r \mid S = s) = p_\varepsilon(s-r) P_R(\mathrm{d}r)/p_S(s)$ is a regular posterior, G is continuous, and $Q = G \circ F_S^{-1}$ is a genuine continuous version.

Then

$$G(s) = \frac{\int r\, p_R(r)\, p_\varepsilon(s-r)\, \mathrm{d}r}{\int p_R(r)\, p_\varepsilon(s-r)\, \mathrm{d}r}. \tag{27}$$

Since $p_\varepsilon(s-r)/p_\varepsilon(s) = \exp\big(\int_{s-r}^{s} \psi_\varepsilon\big)$, the posterior at s is the prior tilted by the average hazard slope over $[s-r, s]$ (Lemma A.4); for binary reward, exactly

$$\mathrm{logit}\, G(s) = \mathrm{logit}\, \mu + \int_{s-1}^{s} \psi_\varepsilon(t)\, \mathrm{d}t, \qquad R \in \{0,1\}. \tag{28}$$

Lemma A.4 (Tilt representation). *For all s, r, $p_\varepsilon(s-r)/p_\varepsilon(s) = \exp\big(\int_{s-r}^{s} \psi_\varepsilon(t)\,\mathrm{d}t\big)$, with the convention $\int_a^b = -\int_b^a$. Consequently*

$$G(s) = \frac{\mathbb{E}[R\, w_s(R)]}{\mathbb{E}[w_s(R)]}, \qquad w_s(r) = \exp\Big(\int_{s-r}^{s} \psi_\varepsilon(t)\,\mathrm{d}t\Big). \tag{29}$$

If $R \in \{a, b\}$ with $a < b$ and $\Pr(R = b) = \mu_b \in (0,1)$, then

$$\mathrm{logit}\, \Pr(R = b \mid S = s) = \mathrm{logit}\, \mu_b + \int_{s-b}^{s-a} \psi_\varepsilon(t)\, \mathrm{d}t, \tag{30}$$

and (28) *is the case $a = 0$, $b = 1$.*

Proof. Since $\log p_\varepsilon$ is locally absolutely continuous with derivative $-\psi_\varepsilon$, $\log p_\varepsilon(s) - \log p_\varepsilon(s-r) = -\int_{s-r}^{s} \psi_\varepsilon$. Dividing the numerator and denominator of $G(s)$ by $p_\varepsilon(s)$ gives the second claim. For (30), the posterior odds are the prior odds times $p_\varepsilon(s-b)/p_\varepsilon(s-a)$, and $\log p_\varepsilon(s-b) - \log p_\varepsilon(s-a) = \int_{s-b}^{s-a} \psi_\varepsilon$. □

The weight w_s is the exponential tilt $e^{\lambda r}$ with the constant rate λ replaced by the average hazard slope over the window between $s-r$ and s. When ψ_ε is nearly constant on windows of length $\bar{r}$, which is the situation at large $|s|$ under a convergent hazard, $G(s) \approx m(\psi_\varepsilon(s))$. Every result in Sections 3.3 and 3.5 is a statement about how this effective rate behaves.

Properties of m. With $M(\lambda) = \mathbb{E}e^{\lambda R}$ and the tilted law $P_\lambda(\mathrm{d}r) \propto e^{\lambda r} P_R(\mathrm{d}r)$, we have $m = (\log M)'$ and $m'(\lambda) = \mathrm{Var}_{P_\lambda}(R) \geq 0$, with strict inequality unless R is degenerate; also $m(0) = \mu$. If $r_{\max}$ is in the support of R, then for every $\delta > 0$

$$\frac{P_\lambda(R \leq r_{\max} - \delta)}{P_\lambda(R > r_{\max} - \delta/2)} \leq \frac{e^{-\lambda\delta/2}}{\Pr(R > r_{\max} - \delta/2)} \to 0, \tag{31}$$

so $m(\lambda) \to r_{\max}$, and $m(\lambda) < r_{\max}$ for finite λ unless $R \equiv r_{\max}$. For binary R, (30) with $\psi_\varepsilon \equiv \lambda$ gives $\mathrm{logit}\, m(\lambda) = \mathrm{logit}\, \mu + \lambda$.

Proposition A.5 (Covariance identity)**.** *Whenever differentiation under the integral is permitted (e.g. ψ_ε continuous),*

$$G'(s) = -\operatorname{Cov}\big(R,\ \psi_\varepsilon(s-R) \,\big|\, S = s\big). \tag{32}$$

Hence if p_ε is log-concave, G and Q are nondecreasing and, by Theorem 3.2(a), so is V_N for every *prior p_R. For Gaussian error, $G'(s) = \mathrm{Var}(R \mid S = s)/\sigma^2$.*

Proof of the covariance identity (32)**.** Assume ψ_ε is continuous. For s in a compact interval J and $r \in [r_{\min}, r_{\max}]$, the derivative $\partial_s p_\varepsilon(s-r) = -\psi_\varepsilon(s-r)p_\varepsilon(s-r)$ is bounded, so differentiation under the integral is permitted in both the numerator and the denominator of G. First, $p_S'(s) = -\int \psi_\varepsilon(s-r)p_\varepsilon(s-r)\, P_R(\mathrm{d}r)$, so the hazard of the score is the posterior mean of the hazard of the error: $\psi_S(s) := -(\log p_S)'(s) = \mathbb{E}[\psi_\varepsilon(s-R) \mid S = s]$. Then

$$\begin{aligned} G'(s) &= -\mathbb{E}\big[R\,\psi_\varepsilon(s-R) \mid S = s\big] + \mathbb{E}[R \mid S = s]\,\mathbb{E}\big[\psi_\varepsilon(s-R) \mid S = s\big] \\ &= -\operatorname{Cov}\big(R, \psi_\varepsilon(s-R) \mid S = s\big). \end{aligned} \tag{33}$$

If ψ_ε is nondecreasing, $r \mapsto \psi_\varepsilon(s-r)$ is nonincreasing, and Chebyshev's association inequality gives a nonpositive covariance: for X' an independent copy of X, f nondecreasing and g nonincreasing, $\mathrm{Cov}(f(X), g(X)) = \frac{1}{2}\mathbb{E}\big[(f(X) - f(X'))(g(X) - g(X'))\big] \leq 0$ because the integrand is pointwise nonpositive. For Gaussian error, $\psi_\varepsilon(s-R) = (s-R)/\sigma^2$ and the covariance is $-\mathrm{Var}(R \mid S = s)/\sigma^2$. Finally F_S^{-1} is increasing, so Q inherits the monotonicity of G.

The same computation localises the failure. $G'(s) < 0$ requires $r \mapsto \psi_\varepsilon(s-r)$ to fail to be nonincreasing on $[r_{\min}, r_{\max}]$, that is, ψ_ε must decrease somewhere on $[s - r_{\max}, s - r_{\min}]$. For a two-point prior, (30) makes this exact:

$$G'(s) = (b-a)\,\pi_s(1-\pi_s)\,\big[\psi_\varepsilon(s-a) - \psi_\varepsilon(s-b)\big], \qquad \pi_s = \Pr(R = b \mid S = s). \tag{34}$$

The calibration curve falls exactly where the larger error $s-a$, which explains s by the low reward, is cheaper per unit than the smaller error $s-b$, which explains it by the high reward.

Sharpness, without differentiability.

Proposition A.6 (Log-concavity is exactly the monotonicity condition)**.** *For p_ε as above the following are equivalent:*

(a) p_ε is log-concave;

(b) for every law of R on a bounded interval, the posterior $P(R \in \cdot \mid S = s)$ is nondecreasing in s in the likelihood-ratio order; in particular $\mathbb{E}[\varphi(R) \mid S = s]$ is nondecreasing in s for every bounded nondecreasing φ, and G is nondecreasing;

(c) G is nondecreasing for every two-point law of R.

Proof. (a)⇒(b). Let $s < s'$ and $f = \log p_\varepsilon$. The posterior at s' has density proportional to $\rho(r) = p_\varepsilon(s'-r)/p_\varepsilon(s-r)$ with respect to the posterior at s, and $\log \rho(r) = f(x+d) - f(x)$ with $x = s-r$ and $d = s'-s > 0$. For concave f the increment $f(x+d)-f(x)$ is nonincreasing in x; as r increases, x decreases, so ρ is nondecreasing in r. This is the likelihood-ratio order. With $\tilde\rho = \rho/\mathbb{E}_s[\rho]$ (mean one under the posterior at s), $\mathbb{E}_{s'}[\varphi(R)]-\mathbb{E}_s[\varphi(R)] = \mathrm{Cov}_s(\varphi(R),\tilde\rho(R)) \geq 0$ by Chebyshev's inequality with two nondecreasing functions. (b)⇒(c) is immediate. (c)⇒(a). Take $R \in \{a,b\}$ with $a < b$. By (30), G is nondecreasing if and only if $s \mapsto f(s-b) - f(s-a)$ is nondecreasing. Writing $x = s-b$ and $d = b-a$, this says that $x \mapsto f(x+d) - f(x)$ is nonincreasing, for every $d > 0$ since $a < b$ are arbitrary. Comparing x with $x+d$ gives $f(x+d) - f(x) \geq f(x+2d) - f(x+d)$, i.e. $f(x+d) \geq \frac{1}{2}\big(f(x) + f(x+2d)\big)$. Thus f is midpoint concave, and being continuous it is concave. □

The equivalence has two readings. First, the failure mechanism identified in Section 3.3 is the exact boundary and not merely a convenient sufficient condition: for every non-log-concave error there is a binary reward whose calibration curve decreases somewhere. Second, log-concavity of p_ε is equivalent to total positivity of order two of the location kernel $(s,r) \mapsto p_\varepsilon(s-r)$ (Karlin, 1968). The total positivity that drives Theorem 3.2 on the side of N is therefore the same property that drives monotonicity of G on the side of the score. Under (b) the conclusion holds for $\mathbb{E}[\varphi(R) \mid S = s]$ with any nondecreasing φ, so it does not depend on the scale on which utility is measured. □

A.6 Proof of Theorem 3.3

Step 0: $\lambda_\star \geq 0$, and $\lambda_\star$ is a log-survival slope. If $\lambda_\star < 0$, then $\psi_\varepsilon \leq \lambda_\star/2$ on some $[M,\infty)$, so $\log p_\varepsilon(t) \geq \log p_\varepsilon(M) + |\lambda_\star|(t-M)/2$, which contradicts integrability. Integrability together with the existence of $\lambda_\star$ also forces $p_\varepsilon(t) \to 0$. For $\lambda_\star > 0$ this is clear. For $\lambda_\star = 0$, note that $p_\varepsilon(t+h)/p_\varepsilon(t) = \exp(-\int_t^{t+h} \psi_\varepsilon) \to 1$ for every fixed h; if $p_\varepsilon(t_n) \geq c > 0$ along $t_n \to \infty$, then $p_\varepsilon \geq c/2$ on infinitely many disjoint intervals of fixed length, which is impossible. L'Hôpital's rule applied to $p_\varepsilon/\bar F_\varepsilon$ then gives

$$-\frac{\mathrm{d}}{\mathrm{d}t}\log \bar F_\varepsilon(t) = \frac{p_\varepsilon(t)}{\bar F_\varepsilon(t)} \;\longrightarrow\; \lim_{t\to\infty} \frac{-p_\varepsilon'(t)}{p_\varepsilon(t)} = \lambda_\star. \tag{35}$$

This is why $\lambda_\star$ can be estimated from a log-survival plot, as suggested in the main text.

Step 1: $\lambda_\star < \infty$. Let $\eta_M = \operatorname{ess\,sup}_{t\geq M} |\psi_\varepsilon(t) - \lambda_\star|$, so $\eta_M \to 0$. For $|r| \leq \bar r$ the interval between $s-r$ and s lies in $[s-\bar r, s+\bar r]$, so for $s \geq M + \bar r$

$$\Big| \int_{s-r}^{s} \psi_\varepsilon - \lambda_\star r \Big| \leq \bar r\, \eta_M. \tag{36}$$

Hence $w_s(r) \to e^{\lambda_\star r}$ uniformly in $r \in [r_{\min}, r_{\max}]$, and w_s is bounded above and away from zero uniformly for large s. Lemma A.4 and bounded convergence give $G(s) \to \mathbb{E}[Re^{\lambda_\star R}]/\mathbb{E}[e^{\lambda_\star R}] = m(\lambda_\star)$. If R is nondegenerate, the properties of m give $\mu < m(\lambda_\star) < r_{\max}$ for $\lambda_\star \in (0,\infty)$.

Step 2: $\lambda_\star = \infty$. Fix $\delta > 0$ and let $A = [r_{\min}, r_{\max} - \delta]$ and $B = (r_{\max} - \delta/2, r_{\max}]$. For $r \in A$ and $r' \in B$ we have $r' - r \geq \delta/2$ and $s - r' \geq s - r_{\max}$, so

$$\frac{p_\varepsilon(s-r)}{p_\varepsilon(s-r')} = \exp\Big(-\int_{s-r'}^{s-r} \psi_\varepsilon\Big) \leq \exp\Big(-\tfrac{\delta}{2} \operatorname*{ess\,inf}_{t\geq s-r_{\max}} \psi_\varepsilon(t)\Big) \;\longrightarrow\; 0. \tag{37}$$

Integrating over A and B, $\Pr(R \in A \mid S=s)/\Pr(R \in B \mid S=s) \leq \big[\Pr(R\in A)/\Pr(R \in B)\big] \cdot o(1) \to 0$, where $\Pr(R \in B) > 0$ by the support assumption. Hence $G(s) \geq (r_{\max} - \delta) - (r_{\max} - r_{\min})\Pr(R \in A \mid S = s) \to r_{\max} - \delta$; since δ is arbitrary and $G \leq r_{\max}$, $G(s) \to r_{\max}$.

Step 3: transfer to Q. $p_S > 0$ on $\mathbb{R}$, so F_S is a continuous increasing bijection onto $(0,1)$ and $F_S^{-1}(u) \to \infty$ as $u \uparrow 1$.

Step 4: the examples. Gaussian: $\psi_\varepsilon(t) = t/\sigma^2 \to \infty$. Generalised Gaussian $p_\varepsilon \propto e^{-|t|^\beta}$ (log-density locally absolutely continuous for every $\beta > 0$): $\psi_\varepsilon(t) = \beta t^{\beta-1}$ for $t > 0$, which tends to ∞, 1 or 0 as $\beta > 1$, $= 1$ or < 1. Laplace with scale b: $\psi_\varepsilon = \operatorname{sign}(t)/b$. Logistic with scale b: $\psi_\varepsilon(t) = \tanh(t/(2b))/b \to 1/b$. Tails $\propto t^k e^{-\lambda t}$: $\psi_\varepsilon(t) = \lambda - k/t \to \lambda$. Student-$t_\nu$: $\psi_\varepsilon(t) = (\nu+1)t/(\nu+t^2) \sim (\nu+1)/t$. Pareto-type tails $\propto t^{-(\alpha+1)}$: $\psi_\varepsilon(t) = (\alpha+1)/t$. Log-normal-type tails: $\psi_\varepsilon(t) = \big(1 + (\log t - m)/s^2\big)/t \to 0$.

The mechanism: explaining away. Steps 1 and 2 are a likelihood-ratio contest between explanations of the same extreme score. Crediting one more unit of reward to the response saves one unit of error, which is worth about $\psi_\varepsilon(s)$ nats of likelihood. If this saving grows without bound, only the highest reward is a credible explanation; if it converges to $\lambda_\star$, every reward level keeps a posterior weight proportional to $e^{\lambda_\star r}$; if it vanishes, the score carries no information about which reward produced it and the posterior falls back to the prior. By (28), in the binary case the log-odds lift at the top of the ranking is exactly $\int_{s-1}^{s} \psi_\varepsilon \to \lambda_\star$. □

A.7 Rates of approach: Proposition A.7

Proposition A.7 (Rates of approach)**.** *(i)* (Heavy tails) *If p_ε is regularly varying at $+\infty$ with index $-(\nu+1)$, $\nu > 0$, and $\psi_\varepsilon(t) \sim (\nu+1)/t$, then $G(s) - \mu = (\nu+1)\operatorname{Var}(R)\, s^{-1}(1+o(1))$ and*

$$V_N - \mu \;\sim\; \frac{(\nu+1)\,\Gamma(1+1/\nu)\,\operatorname{Var}(R)}{s_N}, \qquad s_N := \bar{F}_\varepsilon^{-1}(1/N) \asymp N^{1/\nu}\ell(N), \tag{38}$$

with ℓ slowly varying: heavier tails return to the prior faster.

(ii) (Exponential tails) *If $p_\varepsilon(t) \propto e^{-\lambda_\star t}$ for $t > t_0$, then $G(s) = m(\lambda_\star)$ for $s > t_0 + r_{\max}$ and $|V_N - m(\lambda_\star)| \le (r_{\max} - r_{\min})\, F_S(t_0 + r_{\max})^N$.*

(iii) (Gaussian tails, binary reward) *If $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $R \in \{0, 1\}$, $\mu \in (0, 1)$, then*

$$1 - V_N = \exp\Big(-(1 + o(1))\, \sqrt{2 \log N}\, /\, \sigma \Big), \tag{39}$$

slower than any power of N.

(iv) (Gaussian tails, continuous reward) *If $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and R has density $\asymp (r_{\max} - r)^\beta$ near $r_{\max}$, then $r_{\max} - V_N \sim (\beta+1)\,\sigma/\sqrt{2\log N}$.*

All four follow from $G(s_N) \approx m\big(\psi_\varepsilon(s_N)\big)$ (Appendix A.7): extra samples are wasted beyond a geometric horizon at a plateau, help at super-polynomial cost under Gaussian tails, and hurt beyond the peak under heavy tails.

(i) Heavy tails. Assume $\operatorname{Var}(R) > 0$; otherwise both sides vanish.

Step 1: calibration. Since $t\,\psi_\varepsilon(t) \to \nu+1$, $\psi_\varepsilon(t) = (\nu+1)/s + o(1/s)$ uniformly for $t \in [s - \bar{r}, s + \bar{r}]$, so $\int_{s-r}^{s} \psi_\varepsilon = (\nu+1)r/s + o(1/s)$ uniformly in $|r| \le \bar{r}$ and $w_s(r) = 1 + (\nu+1)r/s + o(1/s)$. By Lemma A.4,

$$G(s) = \frac{\mu + (\nu+1)\mathbb{E}[R^2]/s + o(1/s)}{1 + (\nu+1)\mu/s + o(1/s)} = \mu + \frac{(\nu+1)\operatorname{Var}(R)}{s} + o(1/s). \tag{40}$$

Equivalently $G(s) = m(\psi_\varepsilon(s)) + o(1/s)$, because $m(\lambda) = \mu + \operatorname{Var}(R)\lambda + O(\lambda^2)$: at extreme scores the posterior is the prior tilted by the local hazard slope.

Step 2: the tail of S. By Karamata's theorem (Bingham et al., 1987, Thm. 1.5.11), $\bar{F}_\varepsilon$ is regularly varying with index $-\nu$. By the uniform convergence theorem for regularly varying functions, $\bar{F}_\varepsilon(s - r)/\bar{F}_\varepsilon(s) \to 1$ uniformly in $|r| \le \bar{r}$, so $\bar{F}_S(s) = \mathbb{E}[\bar{F}_\varepsilon(s - R)] \sim \bar{F}_\varepsilon(s)$. Asymptotically equivalent regularly varying functions of negative index have asymptotically equivalent inverses (Bingham et al., 1987, Thm. 1.5.12), so $\bar{F}_S^{-1}(w) \sim \bar{F}_\varepsilon^{-1}(w) = w^{-1/\nu}/L(w)$ as $w \downarrow 0$, with L slowly varying at 0.

Step 3: the rank profile. Combining Steps 1 and 2,

$$E(1 - w) = G\big(\bar{F}_S^{-1}(w)\big) - \mu = (\nu+1)\operatorname{Var}(R)\, w^{1/\nu} L(w)\,(1 + o(1)), \qquad w \downarrow 0. \tag{41}$$

Step 4: the Abelian step. Let $W = 1 - U_{(N)}$, with density $N(1-w)^{N-1}$. Fix a small $\eta > 0$. The event $\{W > \eta\}$ contributes at most $\|E\|_\infty (1-\eta)^N$, which is exponentially small. On $\{W \le \eta\}$ substitute $w = x/N$:

$$\mathbb{E}\big[W^{1/\nu} L(W);\, W \le \eta\big] = N^{-1/\nu} L(1/N) \int_0^{\eta N} x^{1/\nu}\, \frac{L(x/N)}{L(1/N)} \Big(1 - \frac{x}{N}\Big)^{N-1} \mathrm{d}x. \tag{42}$$

Pointwise, $L(x/N)/L(1/N) \to 1$ and $(1-x/N)^{N-1} \to e^{-x}$. Potter's bounds (Bingham et al., 1987, Thm. 1.5.6) give, for any $\gamma > 0$ and η small enough, $L(x/N)/L(1/N) \le 2\max(x^\gamma, x^{-\gamma})$, and $(1 - x/N)^{N-1} \le e^{-x/2}$ for $N \ge 2$. Taking $\gamma < 1/\nu$, dominated convergence yields $\int_0^\infty x^{1/\nu} e^{-x}\, \mathrm{d}x = \Gamma(1 + 1/\nu)$. Because the factor $1 + o(1)$ in Step 3 is uniform on $w \le \eta$, we obtain

$$\begin{aligned} V_N - \mu = \mathbb{E}[E(1-W)] &\sim (\nu+1)\Gamma(1+1/\nu)\,\mathrm{Var}(R)\, N^{-1/\nu} L(1/N) \\ &= \frac{(\nu+1)\Gamma(1+1/\nu)\,\mathrm{Var}(R)}{s_N}, \end{aligned} \tag{43}$$

since $s_N = \bar{F}_\varepsilon^{-1}(1/N) = N^{1/\nu}/L(1/N)$. The constant is accurate at moderate budgets. For Student-t_3 noise and $R \in \{0, 1\}$ with $\mu = 1/2$, direct evaluation of Equation (5) gives ratios of $V_N - \mu$ to the asymptotic expression of 0.81, 0.95, 0.99 and 0.998 at $N = 10^2, 10^3, 10^4, 10^5$.

The exponent $1/\nu$ appears for a transparent reason: the budget-N winner sits at score $s_N \asymp N^{1/\nu}$, and there the local tilt is $(\nu+1)/s_N$. Heavier tails move the winner out faster and so make the tilt vanish sooner.

(ii) Exponential tails. For $s > t_0 + r_{\max}$ we have $s - r > t_0$ for every r in the support, so $p_\varepsilon(s-r) = c\, e^{-\lambda_\star (s-r)}$ and the posterior is exactly $\propto e^{\lambda_\star r} P_R(\mathrm{d}r)$, i.e. $G(s) = m(\lambda_\star)$. Since $V_N = \mathbb{E}[G(S_{(N)})]$ and both G and m take values in $[r_{\min}, r_{\max}]$,

$$|V_N - m(\lambda_\star)| \le (r_{\max} - r_{\min})\, \Pr\big(S_{(N)} \le t_0 + r_{\max}\big) = (r_{\max} - r_{\min})\, F_S(t_0 + r_{\max})^N, \tag{44}$$

with $F_S(t_0 + r_{\max}) < 1$ because $p_S > 0$. Solving $(r_{\max} - r_{\min}) F_S(t_0 + r_{\max})^N \le \delta$ for N gives the horizon stated in the main text.

(iii) Gaussian tails, binary reward. By (28) with $\psi_\varepsilon(t) = t/\sigma^2$, $\int_{s-1}^{s} t\, \mathrm{d}t/\sigma^2 = (2s-1)/(2\sigma^2)$, so exactly

$$1 - G(s) = \Big(1 + \tfrac{\mu}{1-\mu}\, e^{(2s-1)/(2\sigma^2)}\Big)^{-1}, \qquad -\log\big(1 - G(s)\big) = \frac{s}{\sigma^2}\,(1 + o(1)). \tag{45}$$

Let $a_N = \bar{F}_S^{-1}(1/N)$. Since $\bar{F}_S(s) = (1-\mu)\bar{\Phi}(s/\sigma) + \mu\bar{\Phi}((s-1)/\sigma)$ and $-\log\bar{\Phi}(z) = \frac{z^2}{2}(1 + o(1))$, we have $-\log \bar{F}_S(s) = \frac{s^2}{2\sigma^2}(1 + o(1))$ and $a_N = \sigma\sqrt{2\log N}\,(1 + o(1))$. G is nondecreasing (Proposition A.5) and $0 \le 1 - G \le 1$. Fix $\delta \in (0, 1)$.

Upper bound. Splitting on $\{S_{(N)} \le (1-\delta)a_N\}$,

$$1 - V_N \le 1 - G\big((1-\delta)a_N\big) + \Pr\big(S_{(N)} \le (1-\delta)a_N\big). \tag{46}$$

The first term is $\exp\big(-(1-\delta)(1+o(1))\sqrt{2\log N}/\sigma\big)$. For the second, $\Pr(S_{(N)} \le x) = F_S(x)^N \le \exp(-N\bar{F}_S(x))$ and $N\bar{F}_S((1-\delta)a_N) = N^{1-(1-\delta)^2+o(1)}$, which grows polynomially. The second term is therefore $\exp(-N^c)$ for some $c > 0$, which is negligible.

Lower bound. Restricting to $\{S_{(N)} \le (1+\delta)a_N\}$,

$$1 - V_N \ge \big[1 - G\big((1+\delta)a_N\big)\big]\, \Pr\big(S_{(N)} \le (1+\delta)a_N\big), \tag{47}$$

and $\Pr(S_{(N)} \le (1+\delta)a_N) \ge 1 - N\bar{F}_S((1+\delta)a_N) = 1 - N^{1-(1+\delta)^2+o(1)} \to 1$.

Letting $\delta \downarrow 0$ gives $-\log(1 - V_N) = (1 + o(1))\sqrt{2\log N}/\sigma$. The rate is slow because the tilt at the winning score is a_N/σ^2, and light tails make a_N grow only like $\sqrt{\log N}$: raising the top score by one unit requires multiplying N by roughly e^{a_N/σ^2}.

(iv) Gaussian tails, continuous reward. Assume $P_R(\mathrm{d}r) = p_R(r)\,\mathrm{d}r$ with $p_R(r) = c\,(r_{\max} - r)^{\beta}(1 + o(1))$ as $r \uparrow r_{\max}$, $\beta > -1$. For Gaussian error, $p_\varepsilon(s - r) \propto e^{-s^2/(2\sigma^2)}\, e^{-r^2/(2\sigma^2)}\, e^{(s/\sigma^2)r}$, so $G(s) = \tilde{m}(s/\sigma^2)$ *exactly*, where $\tilde{m}$ is the tilt function of the modified prior $\tilde{p}_R \propto p_R\, e^{-r^2/(2\sigma^2)}$. This prior has the same $(r_{\max} - r)^{\beta}$ behaviour at the top. Under the tilted law the gap $y = r_{\max} - R$ has density $\propto y^{\beta}(1 + o(1))e^{-\lambda y}$ near $y = 0$, so by Laplace's method it is asymptotically $\mathrm{Gamma}(\beta + 1, \lambda)$ and $r_{\max} - \tilde{m}(\lambda) \sim (\beta + 1)/\lambda$. Hence $r_{\max} - G(s) \sim (\beta + 1)\sigma^2/s$. The two-sided argument of (iii), with $a_N = \sigma\sqrt{2\log N}(1 + o(1))$ unchanged because R is bounded, gives $r_{\max} - V_N \sim (\beta + 1)\sigma^2/a_N = (\beta + 1)\sigma/\sqrt{2\log N}$. □

A.8 Proof of Proposition 3.4 and of Proposition A.8

Proposition A.8 (Binary reward: the tail likelihood ratio decides)**.** *Let $R = C \in \{0, 1\}$ and p_1, p_0 the densities of S given $C = 1, 0$, with no further assumption. Then*

$$\operatorname{logit} Q(u) = \operatorname{logit} \mu + \log \Lambda\big(F_S^{-1}(u)\big), \qquad \Lambda = p_1/p_0, \tag{48}$$

so $\lim_{u\uparrow 1} Q(u)$ is 1, μ, or 0 as $\Lambda(s) \to \infty$, 1, or 0, and a plateau in $(0, 1)$ for any other finite positive limit.

Theorem 3.3 is the case $\log \Lambda(s) = \int_{s-1}^{s} \psi_\varepsilon$ of (28), Proposition 3.4 that of Gaussian classes with unequal variances (Appendix A.8); in verifiable tasks Λ at the top of the score range can be estimated directly.

Write $\varphi_\sigma(x) = (2\pi\sigma^2)^{-1/2} e^{-x^2/(2\sigma^2)}$ (and $\varphi_0 \equiv 0$ off the origin), $\sigma_{\max} = \sigma(r_\sigma)$, and $P(R \in \mathrm{d}r \mid S = s) \propto \varphi_{\sigma(r)}(s - r)\, P_R(\mathrm{d}r)$.

A monotonicity fact. For fixed $x \neq 0$, $\partial_\sigma \log \varphi_\sigma(x) = -1/\sigma + x^2/\sigma^3$, which is positive for $\sigma < |x|$. So when the discrepancy $|x|$ exceeds every available scale, a wider conditional law always assigns it more density. We take $s \geq 2\bar{r} + 2\sigma_{\max}$, so that $|s - r| \geq s - \bar{r} > \sigma_{\max}$ for all r.

Concentration at the noisiest level. Fix $\eta \in (0, \sigma_{\max})$ and put $A_\eta = \{r : \sigma(r) \leq \sigma_{\max} - \eta\}$ and $D_\eta = \{r : \sigma(r) \geq \sigma_{\max} - \eta/2\}$. By continuity D_η contains a neighbourhood of r_σ in $[r_{\min}, r_{\max}]$, so $\Pr(R \in D_\eta) > 0$ because r_σ is in the support. For $r \in A_\eta$, monotonicity in σ and then in $|x|$ gives

$$\varphi_{\sigma(r)}(s - r) \leq \varphi_{\sigma_{\max} - \eta}(s - r) \leq \frac{1}{\sqrt{2\pi}(\sigma_{\max} - \eta)} \exp\Big(- \frac{(s - \bar{r})^2}{2(\sigma_{\max} - \eta)^2}\Big), \tag{49}$$

and this bound needs no lower bound on $\sigma(r)$. For $r \in D_\eta$, likewise,

$$\varphi_{\sigma(r)}(s - r) \geq \varphi_{\sigma_{\max} - \eta/2}(s + \bar{r}) = \frac{1}{\sqrt{2\pi}(\sigma_{\max} - \eta/2)} \exp\Big(- \frac{(s + \bar{r})^2}{2(\sigma_{\max} - \eta/2)^2}\Big). \tag{50}$$

The ratio of the two bounds is $\exp(-c_\eta s^2 + O(s))$ with $c_\eta = \frac{1}{2}\big[(\sigma_{\max} - \eta)^{-2} - (\sigma_{\max} - \eta/2)^{-2}\big] > 0$, so $\Pr(R \in A_\eta \mid S = s)/\Pr(R \in D_\eta \mid S = s) \to 0$. Hence $\Pr(\sigma(R) > \sigma_{\max} - \eta \mid S = s) \to 1$ for every $\eta > 0$.

From scale to location. Because σ is continuous on the compact interval $[r_{\min}, r_{\max}]$ with a unique maximiser, for every $\delta > 0$ there is $\eta > 0$ with $\{\sigma > \sigma_{\max} - \eta\} \subset (r_\sigma - \delta, r_\sigma + \delta)$. Otherwise some sequence r_n with $\sigma(r_n) \to \sigma_{\max}$ and $|r_n - r_\sigma| \geq \delta$ would have a limit point that is a second maximiser. So the posterior concentrates at r_σ and $G(s) \to r_\sigma$. Finally $V_N = \mathbb{E}[G(S_{(N)})]$, $S_{(N)} \to \infty$ almost surely because S has unbounded upper support (each conditional law is Gaussian with $\sigma(R) > 0$), and bounded convergence gives $V_N \to r_\sigma$. □

Why scale beats shift. The log-likelihood is

$$\log \varphi_{\sigma(r)}(s - r) = -\frac{s^2}{2\sigma^2(r)} + \frac{s\,r}{\sigma^2(r)} - \frac{r^2}{2\sigma^2(r)} - \log \sigma(r) + \text{const.} \tag{51}$$

The order-s^2 term depends on r only through the scale, and the reward enters first at order s. With homoskedastic noise the s^2 term is common to all r, the order-s term sr/σ^2 is maximised at $r_{\max}$, and we recover Theorem 3.3 with $\lambda_\star = \infty$. With heteroskedastic noise the s^2 term decides.

The multiplicative model. If $S = R(1+\varepsilon)$ with $\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$, then $S \mid R = r \sim \mathcal{N}(r, \sigma_0^2 r^2)$ and $\sigma(r) = \sigma_0|r|$. If $\Pr(R = 0) = 0$, $|r_{\min}| > r_{\max}$ and $r_{\min}$ is in the support, the unique maximiser is $r_{\min}$, and the proposition gives $V_N \to r_{\min}$: BoN converges to the worst reward level. The vanishing of σ at $r = 0$ is harmless because the upper bound on A_η does not use a lower bound on σ.

Proof of Proposition A.8. Bayes' rule gives $\Pr(C = 1 \mid S = s)/\Pr(C = 0 \mid S = s) = \mu\, p_1(s)/\big((1-\mu)p_0(s)\big)$; take logarithms and compose with F_S^{-1}. The limits follow from continuity of the logistic function on $[-\infty, \infty]$. In the additive model $p_1(s)/p_0(s) = p_\varepsilon(s-1)/p_\varepsilon(s) = \exp(\int_{s-1}^{s} \psi_\varepsilon)$ by Lemma A.4. For Gaussian classes with scales σ_1, σ_0, $\log \Lambda(s) = s^2\big(\sigma_0^{-2} - \sigma_1^{-2}\big)/2 + O(s)$, which tends to $+\infty$ or $-\infty$ according as $\sigma_1 > \sigma_0$ or $\sigma_1 < \sigma_0$.

The lexicographic principle for binary reward. Suppose $-\log p_c(s) = a_c s^{\beta_c}\big(1 + o(1)\big)$ as $s \to \infty$ for $c \in \{0, 1\}$. If $\beta_0 < \beta_1$, incorrect responses have the heavier tail, $\log \Lambda \to -\infty$ and $Q(u) \to 0$, so BoN converges to the *worst* outcome even if correct responses score higher in the bulk. If $\beta_0 = \beta_1$ and $a_0 < a_1$, the wider incorrect class wins in the same way. Only when shape and scale agree do lower-order terms, among them the location shift, decide; in the additive model they decide through $\int_{s-1}^{s} \psi_\varepsilon$, which is Theorem 3.3. An extreme score is thus explained first by the heaviest tail, then by the largest scale, and only then by the largest reward.

A.9 Proofs for Section 3.5

Proposition A.9 (Contamination)**.** *Let $p_\varepsilon = (1-\eta)\, p_0 + \eta\, p_1$, $\eta \in (0, 1]$, with p_0 having generalised Gaussian or exponential tails ($\beta \geq 1$) and p_1 satisfying the condition of Proposition A.7(i). Then $\lambda_\star = 0$ for every $\eta > 0$, so $\lim_{u \uparrow 1} Q(u) = \mu$ and Corollary A.3 applies once a finite budget beats random.*

Any heavy-tailed component, however small, eventually forces a return to the prior; by Lemma 3.5, smaller η only delays the turn (Appendix A.9).

Corollary A.10 (No finite certificate of monotonicity)**.** *Let Q take values in $[r_{\min}, r_{\max}]$, $K \in \mathbb{N}$, $\gamma > 0$. Then there exist $\delta > 0$ and Q' with values in $[r_{\min}, r_{\max}]$, equal to Q on $[0, 1-\delta]$ and constant on $(1-\delta, 1)$ with value $\mu' = \int_0^1 Q'$, such that $|V'_N - V_N| \leq \gamma$ for all $N \leq K$ and $\lim_{N\to\infty} V'_N = \mu'$.*

No BoN curve up to budget K rules out a later return, so our diagnostics target $\lambda_\star$ and Λ directly (Appendix A.9).

The localisation heuristic. ΔV_N weights $\mathrm{d}Q$ by the $\mathrm{Beta}(N+1, 2)$ kernel, whose mode is $N/(N+1)$ (C4). A feature at quantile $u_\star$ therefore receives maximal weight when $N/(N+1) = u_\star$, i.e. at $N_{\mathrm{loc}} = u_\star/(1-u_\star)$.

Proof of Lemma 3.5, and tightness. By Equation (5), $V_N - \tilde{V}_N = \mathbb{E}\big[(Q - \tilde{Q})(U_{(N)})\big] = \mathbb{E}\big[(Q-\tilde{Q})(U_{(N)})\, \mathbf{1}\{U_{(N)} > 1-t\}\big]$, because the integrand vanishes on $[0, 1-t]$. Hence $|V_N - \tilde{V}_N| \leq \|Q - \tilde{Q}\|_\infty \Pr(U_{(N)} > 1-t) = \|Q - \tilde{Q}\|_\infty \big(1 - (1-t)^N\big)$, and $1 - (1-t)^N \leq Nt$ by Bernoulli's inequality. The bound is attained: if $Q - \tilde{Q} \equiv c$ on $(1-t, 1]$, then $V_N - \tilde{V}_N = c\,\big(1 - (1-t)^N\big) \geq c\,(1 - e^{-Nt})$. A tail feature of height c on the top-t region is therefore invisible for $Nt \ll 1$ and fully expressed for $Nt \gg 1$; the switch happens at $N \approx 1/t$, consistent with N_{loc} for $u_\star = 1 - t$.

Worked example: Student-t errors. Let $S = R + \sigma T_\nu$. Then $\psi_\varepsilon(t) = (\nu+1)t/(\nu\sigma^2 + t^2)$, which is increasing on $[-\sigma\sqrt{\nu}, \sigma\sqrt{\nu}]$ and decreasing on $[\sigma\sqrt{\nu}, \infty)$. By the localisation in Appendix A.5, G is nondecreasing whenever the window $[s - r_{\max}, s - r_{\min}]$ lies in $[-\sigma\sqrt{\nu}, \sigma\sqrt{\nu}]$, i.e. for $s \in [r_{\max} - \sigma\sqrt{\nu},\ r_{\min} + \sigma\sqrt{\nu}]$, and nonincreasing for $s \geq r_{\max} + \sigma\sqrt{\nu}$. If $r_{\max} - r_{\min} \leq 2\sigma\sqrt{\nu}$, the upper turning point of G therefore lies in the band $[r_{\min} + \sigma\sqrt{\nu},\ r_{\max} + \sigma\sqrt{\nu}]$. When the reward range is small compared with $\sigma\sqrt{\nu}$, the score exceedance probability there is $\bar{F}_S \approx \Pr(T_\nu > \sqrt{\nu})$, and the localisation heuristic gives

$$N_{\mathrm{loc}} \approx \frac{1}{\Pr(T_\nu > \sqrt{\nu})} \approx 4,\ 11,\ 26 \qquad (\nu = 1, 3, 5), \tag{52}$$

independently of σ to first order. An exact evaluation of Equation (5) for $R \in \{0, 1\}$, $\mu = 1/2$ and $\sigma \in \{1, 3\}$ gives optimal budgets $\arg\max_N V_N = 5$, 11 and 22–24 respectively. The tail index places the peak; the noise scale mainly sets its height. As $\nu \to \infty$, $\Pr(T_\nu > \sqrt{\nu}) \to 0$ and the peak recedes to infinity, which recovers the monotone Gaussian case.

Proof of Proposition A.9. (a) The contamination label is independent of R and of both errors, so $S - R$ has the mixture density $p_\varepsilon = (1-\pi)p_{\varepsilon_0} + \pi p_{\varepsilon_1}$ independently of R, and the model is additive. Its hazard slope is

$$\psi_\varepsilon = \frac{(1-\pi)(-p'_{\varepsilon_0}) + \pi\, \psi_{\varepsilon_1} p_{\varepsilon_1}}{(1-\pi)p_{\varepsilon_0} + \pi\, p_{\varepsilon_1}}. \tag{53}$$

After dividing numerator and denominator by p_{ε_1}, the numerator is $(1-\pi)\, o(1) + \pi\, \psi_{\varepsilon_1} \to 0$ by the assumptions and the denominator is at least π, so $\psi_\varepsilon \to 0$. Theorem 3.3 with $\lambda_\star = 0$, whose Step 1 needs no support assumption, gives $Q_\pi(u) \to m(0) = \mu$. The mechanism is that the heavy component, however rare, owns the extreme scores: the posterior probability that a response at score s is contaminated tends to one.

(b) Build both models on one probability space. Let $(R_i, \varepsilon_{0,i}, \varepsilon_{1,i}, B_i)_{i \le N}$ be independent across i, with $B_i \sim \text{Bernoulli}(\pi)$ the contamination flags. Set $S_i^0 = R_i + \varepsilon_{0,i}$ and $S_i^\pi = R_i + \varepsilon_{0,i}$ if $B_i = 0$, and $S_i^\pi = R_i + \varepsilon_{1,i}$ if $B_i = 1$. Each score vector has the correct law. On $\{B_1 = \cdots = B_N = 0\}$, which has probability $(1-\pi)^N$, the score vectors coincide, and so do the selected indices (ties have probability zero) and the selected rewards. Otherwise the two selected rewards differ by at most $r_{\max} - r_{\min}$. Hence $|V_N^\pi - V_N^0| \le (r_{\max} - r_{\min})\big(1 - (1-\pi)^N\big) \le (r_{\max} - r_{\min})N\pi$. □

Part (b) uses no property of ε_1 whatsoever: rare mis-scoring is invisible at small budgets whatever its form. Part (a) uses only the tail of ε_1 and is insensitive to π. The rate π therefore plays exactly the role of t in Lemma 3.5.

Proof of Corollary A.10. For $N \le K$ and $\pi \le \delta/((r_{\max} - r_{\min})K)$, part (b) gives $|V_N^\pi - V_N^0| \le (r_{\max} - r_{\min})N\pi \le \delta$. By part (a), $Q_\pi(u) \to \mu$, and the argument of Appendix A.4 gives $V_N^\pi = \mathbb{E}[Q_\pi(U_{(N)})] \to \mu$. □

If ε_0 is Gaussian and $r_{\max}$ is in the support, then $V_N^0 \to r_{\max}$ while $V_N^\pi \to \mu$. The two curves are indistinguishable to within δ on $N \le K$, yet their eventual gap is $r_{\max} - \mu$, the entire value of selection. The window of guaranteed invisibility, $N \le \delta/((r_{\max} - r_{\min})\pi)$, is linear in $1/\pi$, which is the precise sense in which halving the contamination rate doubles it.

B Detailed Experimental Protocol

Synthetic phase maps. The primary prior is bounded $R \sim \text{Uniform}[-\sqrt{3}, \sqrt{3}]$ ($\mu = 0$, $r_{\max} = \sqrt{3}$), as Theorem 3.3 assumes; Bernoulli(.5) (bounded, discrete) and $\mathcal{N}(0, 1)$ (unbounded, outside Theorem 3.3) are stress tests. For each prior, six target correlations $\rho \in \{.5, .6, .7, .8, .9, .95\}$ are crossed with Gaussian, variance-standardized Laplace, t_ν for $\nu \in \{30, 10, 6, 4, 3, 2.5\}$, and multiplicative $\varepsilon = aRZ$ error. Gaussian, Laplace, and t_ν are the diverging-, constant-, and vanishing-influence regimes of Theorem 3.3; the multiplicative error is the model $S = R(1 + aZ)$ of Section 3.4. Noise scale a matches the pooled target correlation, so ρ changes only where a tail becomes visible. Each of 162 cells uses 20,000 separately drawn $K = 512$ candidate sets. An interior peak is a downturn only when its selected peak-to-512 contrast has a 95% simultaneous max-t interval above zero over the tested grid. Full grids and phase definitions follow below. A separate 12-cell bounded-prior experiment uses conditionally Gaussian errors with $\sigma_\beta(r) = \exp(\beta r/\sqrt{3})$ for $\beta \in \{-1, +1\}$ at the same six correlations; these satisfy the assumptions of Proposition 3.4, with unique scale maximiser $r_\sigma = r_{\max}$ for $\beta = +1$ and $r_\sigma = r_{\min}$ for $\beta = -1$.

Controlled real-model study. Qwen2.5-7B-Instruct (Qwen Team, 2024) generated 256 answers for each of 500 MATH-500 problems (Hendrycks et al., 2021) and 1,319 GSM8K test questions (Cobbe et al., 2021). Four open reward models score the same fixed candidates: Skywork-Reward-V2 Llama-3.2 1B/3B and Qwen3-4B (Liu et al., 2025), plus Tulu-3 8B (Lambert et al., 2024a). Because the candidates, and hence μ, are held fixed, by Equation (5) every difference between conditions is a difference in the calibration curve Q. Besides QA and answer-only, MATH controls for Tulu and

Skywork Qwen3-4B preserve the user–assistant template while either replacing the question with a fixed mask or deterministically shuffling questions across answers. Section 3 says what a change in Q implies, not what causes it; these controls separate missing information from role/template shift as candidate causes.

External PPE validation. PPE (Frick et al., 2024) contributes 32 responses per prompt from four base policies on MATH, GPQA, MMLU-Pro, MBPP+, and IFEval. Thirteen learned RMs on each task give 65 cells; NaiveVerbosityModel is retained only as one explicit stress cell per task. We exclude 31 nonnumeric Nemotron rows and record the PPE revision, source hashes, exact exclusions, and policy strata.

Estimation and uncertainty. If $c_{(j)}$ is correctness at increasing score rank j among K candidates, the exact finite-pool BoN expectation is

$$\widehat{f}(N) = \sum_{j=N}^{K} \frac{\binom{j-1}{N-1}}{\binom{K}{N}} c_{(j)}, \tag{54}$$

which is unbiased for V_N of Proposition 3.1 for every $N \le K$; its increments have the nonnegative-kernel form of Proposition A.1, so the rank profile both estimates Q and witnesses any turn. Ties are block averaged, which is exactly the expectation under the uniform random tie breaking of Section 3; we also test candidate-index, reverse-index, and repeated random tie orders. Prompt bootstrap intervals preserve all candidates; 2,000-replicate max-t intervals control selected interior contrasts within a cell. Exact $K = 256$ rank profiles determine controlled $N^\star$. PPE uses top-12.5% excess E_{tail} (top four minus prompt base correctness), the empirical mean of $E(u)$ over the top eighth of scores, i.e. the band that budgets $N \approx 8$ read by Lemma 3.5; for binary correctness its sign is the sign of $\log \Lambda$ averaged over that band (Proposition A.8).

For the PPE intervention, two base-policy-stratified folds alternate as calibration and test. A negative calibration E_{tail} triggers the calibration-optimal cap in $\{1, 2, 4, 8, 12, 16, 24, 32\}$, a finite-pool version of the stopping rule of Theorem 3.2(b); otherwise $N = 32$ is retained. One-sided cell tests are BH corrected over all 70 cells. Diagnostic AUC, average precision, and Spearman correlation compare $-E_{\mathrm{tail}}$, low rank accuracy, and score-tail spread; confidence intervals resample cells while retaining both folds. As a same-pool hedging baseline, rank-Soft-BoN uses all 32 candidates, maps scores to tied empirical quantiles (and so, like BoN, is invariant to monotone recalibration, Proposition 3.1), and chooses its inverse temperature on the calibration fold before held-out evaluation.

Synthetic grid. The complete budget grid is

$$\{1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64\} \cup \{96, 128, 192, 256, 384, 512\}. \tag{55}$$

Student errors are variance standardized and the empirical second-moment equation is solved for noise scale a in every cell. The selected-peak interval uses 1,000 Gaussian-multiplier max-t draws with seed 0. A non-downturn cell is called monotone when its empirical peak is $N = 512$ and plateau/boundary otherwise; these labels describe the observed window only. Binary multiplicative scores have an atom at zero, making that row an explicit tie/continuity stress test rather than an instantiation of Proposition 3.1. The 12 heteroskedastic cells use the same bounded Uniform prior, budget grid, 20,000 candidate sets, six target correlations, and 1,000-draw simultaneous procedure. Their model is $S = R + a \exp(\beta R/\sqrt{3})Z$ for $\beta \in \{-1, +1\}$; a is solved per cell, and seed 314159 deterministically seeds the cell sequence.

Controlled generation and scoring. Generation uses temperature 0.8, top-$p = .95$, a 1,024-token output limit, a 4,096-token context limit, and fixed per-prompt seeds. Correctness follows the repository's boxed-answer extraction and exact matching. RMs use their released chat templates, bfloat16 inference, and maximum length 4,096. The answer-only condition packages the answer as the sole user message, changing both information and role relative to QA. The masked prompt is exactly "[Question omitted]". Shuffling uses a fixed nonzero cyclic offset determined by seed 42, ensuring that no answer receives its original question. For Tulu, visual $N^\star$ estimates over 16, 32, 50, 64, and 100 histogram bins and local-linear bandwidths $.02, .05, .10$ are 3–4 for answer-only, 6–45 for QA, 6–59 when masked, and 5–25 when shuffled; the exact tied-rank values are 3, 68, 82, and 18, respectively.

PPE provenance. The four base policies are Meta-Llama-3-8B-Instruct, Claude-3-Haiku-20240307, Gemma-2-9B-IT, and GPT-4o-mini-2024-07-18. Nonnumeric Nemotron exclusions are 23 MATH, 4 GPQA, 2 MMLU-Pro, 2 IFEval, and 0 MBPP+. Every compact cell stores its source-relative path and SHA-256 hash. Source files contain 512 prompts per task except 507 for MBPP+ before RM-specific exclusions. Within-prompt tied score blocks receive mean correctness before exact finite-pool integration.

Tail and cross-fit definitions. For PPE with $K = 32$ and P prompts,

$$E_{\text{tail}} = \frac{1}{P}\sum_{i=1}^{P}\left(\frac{1}{4}\sum_{j=K-3}^{K} c_{i,(j)} - \frac{1}{K}\sum_{j=1}^{K} c_{i,(j)}\right). \tag{56}$$

The two folds are stratified within base policy. A calibration-selected cap is evaluated only on the opposite fold. AUC labels a fold positive when the held-out gain at its calibration-selected cap has a normal 95% lower bound above zero. Average precision integrates precision over the distinct risk thresholds and is compared with the empirical prevalence baseline. Diagnostic intervals use 2,000 cell-clustered bootstrap draws; cross-fitted policy p-values are one-sided and BH corrected across all 70 cells. Width sensitivity freezes these folds and uses $m \in \{2, 4, 8\}$; leave-one-cell-out selection maximizes Spearman correlation with continuous held-out gain in the other cells. Tie sensitivity uses candidate-index, reverse-index, and 200 uniformly randomized boundary-tie orders.

For rank-Soft-BoN, tied empirical quantiles $u_{ij} \in [0, 1]$ give candidate probabilities

$$p_{ij}(\lambda) = \frac{\exp(\lambda u_{ij})}{\sum_{k=1}^{32}\exp(\lambda u_{ik})}. \tag{57}$$

The calibration fold selects the smallest maximizing $\lambda \in \{0, .125, .25, .5, 1, 2, 4, 8, 16, 32, 64\}$ and the opposite fold reports its exact expected correctness. This quantile normalization follows the score-rank step of HedgeTune but uses direct calibration utility rather than its residual root finder.

B.1 Correlation and finite-budget localization

Correlation sets location, not regime. For Uniform $t_{2.5}$, the detected peak shifts from $N = 12$ at $\rho = .5$ to $N = 32$ at $\rho = .95$; the corresponding drop contracts from 0.495 [0.471, 0.520] to 0.296 [0.282, 0.310]. For t_6, the peak moves to $N = 128$ and the drop falls to 0.0177 [0.0118, 0.0235] at $\rho = .95$. A smaller noise scale pushes the quantiles where the heavy tail dominates further up, so larger budgets are needed to read them (Lemma 3.5), while $\lambda_\star = 0$ and hence the eventual return (Theorem 3.3) are unchanged. Better global quality thus delays and attenuates hacking without changing the error family.

B.2 Heteroskedastic scale results

Heteroskedastic scale. Proposition 3.4 predicts $V_N \to \sqrt{3}$ for $\beta = +1$ and $V_N \to -\sqrt{3} < \mu = 0$ for $\beta = -1$. With $\beta = +1$, all six cells peak empirically at $N = 512$; with $\beta = -1$, four have supported downturns. For $\rho = .5$–$.8$, peaks move through $N = 4, 6, 12, 32$ and drops range from 1.939 to 0.600, with simultaneous lower bounds at least 0.573. Since no value exceeds $r_{\max} = \sqrt{3} < 1.939$, the largest drop already leaves BoN below random selection, on the way to the predicted limit $r_{\min}$. At $\rho = .9, .95$ the failure lies beyond the window. Rank accuracy is still 0.828 at $\rho = .8$: high bulk quality coexists with a regime that ends below random, which is the non-identifiability behind Corollary A.10.

B.3 Prior and multiplicative stress tests

Prior stress tests. Binary rewards remain within Theorem 3.3 through (28) and yield 28 downturn, 19 plateau, and 7 monotone cells. The unbounded Gaussian prior lies outside Theorem 3.3, because the prior's own tail competes with the error's, and yields 17/6/31. Proposition A.5 still holds for both priors: Gaussian rows remain non-downturn, as does t_{30}, and $t_{2.5}$ is the most robust downturn row. Beyond the bounded case, phase depends jointly on reward prior, error tail, and budget.

Multiplicative control. The multiplicative control is monotone for Uniform and Gaussian priors and plateaus for binary rewards. Its scale $\sigma(r) \propto |r|$ is maximised at *both* edges of a symmetric prior, so the uniqueness condition of Proposition 3.4 fails, and the example $S = R(1 + \varepsilon)$ of Section 3.4 sits exactly at its boundary $|r_{\min}| = r_{\max}$. With the $O(s^2)$ scale terms tied, the $O(s)$ shift term breaks the tie in favour of $r_{\max}$, so extreme positive scores still favour the good edge. A scale that grows with $|R|$ is therefore insufficient; what matters is which reward level dominates the conditional upper tail.

C PPE Sensitivity Details

At fixed folds, top-m tail risk has AUC/AP $0.972/0.749$, $0.983/0.825$, and $0.976/0.771$ for $m = 2, 4, 8$. By Lemma 3.5, these bands (the top 1/16, 1/8, and 1/4 of scores) are those read at budgets of roughly 16, 8, and 4. All six negative signs survive two deterministic and 200 random tie orders. Nested selection favors $m = 2$ without improving AP, supporting a robust range, not an optimal width; tune $m \geq 2$ on disjoint data. With $K = 32$ candidates per prompt we do not estimate $\lambda_\star$ from log-survival plots on PPE; E_{tail} targets Λ at the top of the score range directly (Proposition A.8). Best-of-Tails needs endpoint-normalized rewards and ITP reference-policy likelihoods, absent from these signed fixed-pool logits (Hsu et al., 2026; Huang et al., 2025); rank-Soft-BoN is the feasible same-pool comparator. The median learned cell has any tied score in 73.2% of prompts and a top-four-boundary tie in 7.0%. All six primary negative MBPP+ cells remain negative for $m = 2$; one additional GPQA cell crosses below zero. For $m = 8$, five remain negative and Eurus moves to $+0.003$. Leave-one-cell-out continuous-gain selection chooses $m = 2$ for 126/130 folds and $m = 8$ for four, with AP 0.749.

C.1 Relation to bulk ranking metrics

Rank accuracy is also strong (AUC/AP $0.972/0.747$); paired tail-minus-rank intervals include zero for AUC and AP. This does not contradict Section 3, whose claim is non-identifiability, not non-correlation. By the argument of Lemma 3.5, curves that agree below quantile $1 - \delta$ have bulk metrics within $O(\delta)$ of one another yet may end in different regimes, and the synthetic $\beta = -1$, $\rho = .8$ cell (rank accuracy 0.828) is such a case. On PPE, however, the bad-tail cells are also low-rank-accuracy cells. The broad AP intervals expose the rarity and reject any claim that bulk ranking quality cannot predict the problem in practice.

D Complete Controlled Results

Table 2 reports the two structure-aware controls alongside QA and answer-only. Comparing masked-question with answer-only separates question content from the chat-role shift; shuffled questions retain the template while supplying deliberately mismatched semantics.

D.1 Budget localization

Tulu's exact tied-rank $N^\star$ is 68 for QA, 3 for answer-only, 82 when masked, and 18 when shuffled. Because the kernel $u^N(1-u)$ of Equation (6) peaks at $u = N/(N+1)$, these place the turn of Q near the 98.6th, 75th, 98.8th, and 94.7th score percentiles: the answer-only failure sits deep in the bulk, whereas the other conditions turn, if at all, only at the extreme top (Section 3.5). Answer-only visual estimates stay in $\{3, 4\}$ across five histogram grids and three bandwidths; unsupported conditions vary more (Appendix B). Top-1% excess is $+.0538, -.0310, +.0495, +.0231$ in the same order.

E PPE Reward Models and Additional Results

Base-policy strata show joint policy–RM effects: Eurus tail excess on MBPP+ is $+.130$ for Meta-Llama-3-8B responses but $-.127$ for Claude-3-Haiku, with peak-to-32 drops .011 and .178. The other five negative cells also reverse sign across policies. Section 3 applies to the aggregate curve, which averages these strata, so the top of Q depends on the policy mix. In a way analogous to Proposition A.9, though not an instance of it, a minority stratum with a bad tail can decide the top quantiles, which precludes a task-independent safe budget here.

Table 4: All original controlled conditions. Drop is the selected interior peak minus $N = 256$ with a within-condition simultaneous max-t interval. — denotes a boundary maximum at $N = 256$, with no interior peak contrast.

Task	Input	RM	Rank acc.	Peak N	Drop [95% CI]
MATH-500	QA	Skywork 1B	.746	24	.0049 [-.0180,.0278]
	QA	Skywork 3B	.772	160	.0014 [-.0087,.0115]
	QA	Skywork Qwen3-4B	.812	256	—
	QA	Tulu-3 8B	.741	64	.0051 [-.0118,.0220]
	answer-only	Skywork 1B	.712	24	.0109 [-.0141,.0360]
	answer-only	Skywork 3B	.735	256	—
	answer-only	Skywork Qwen3-4B	.799	256	—
	answer-only	Tulu-3 8B	.581	4	.0597 [.0316,.0878]
GSM8K	QA	Skywork 1B	.706	12	.0070 [-.0026,.0167]
	QA	Skywork 3B	.731	12	.0050 [-.0038,.0138]
	QA	Skywork Qwen3-4B	.803	24	.0028 [-.0044,.0100]
	QA	Tulu-3 8B	.726	16	.0031 [-.0064,.0126]
	answer-only	Skywork 1B	.605	8	.0046 [-.0053,.0145]
	answer-only	Skywork 3B	.622	12	.0042 [-.0064,.0149]
	answer-only	Skywork Qwen3-4B	.732	32	.0005 [-.0076,.0087]
	answer-only	Tulu-3 8B	.569	4	.0016 [-.0105,.0136]

What occupies the MBPP+ bad tail? In terms of Proposition A.8, this asks what drives the upper tail of p_0, the score density of incorrect responses. Released responses show polished rather than malformed top-tail errors. They are about 13% longer, add three lines, and across all six RMs are more likely to parse, define the target function, contain a docstring, and place prose outside the fence; repeated examples/assertions are 13.3 points less common. These shared-pool correlates are descriptive, not causal.

The 13 learned PPE score functions are ArmoRM-Llama3-8B-v0.1, Athene-RM-70B, Athene-RM-8B, Eurus-RM-7B, Llama-3-OffsetBias-RM-8B, Skywork-Reward-Gemma-2-27B, Skywork-Reward-Llama-3.1-8B, Starling-RM-34B, Starling-RM-7B-alpha, InternLM2 reward models at 1.8B, 7B, and 20B, and Nemotron-4-340B-reward. PPE's NaiveVerbosityModel is retained only as a stress baseline. It has negative tail excess on all five tasks and a supported BH-corrected cap gain on MATH, MMLU-Pro, and MBPP+, confirming that the pipeline detects a deliberately shallow score rule without conflating it with learned-RM prevalence.

Table 5: Learned PPE cells with negative primary tail excess. "D" marks the strict interior-downturn definition; q is for the cross-fitted cap policy over all 70 cells.

MBPP+ RM	Rank acc.	E_{tail}	Peak	Drop	Cap gain [95% CI]	q
Eurus-7B	.537	-.014	2	.059 D	.059 [.024,.093]	.0047
Skywork-Llama-8B	.514	-.028	2	.043	.039 [.001,.078]	.192
Starling-7B	.387	-.157	1	.165	.165 [.128,.202]	$< 10^{-4}$
InternLM2-1.8B	.464	-.039	1	.045	.045 [.002,.088]	.161
InternLM2-7B	.452	-.114	1	.171	.171 [.132,.213]	$< 10^{-4}$
Nemotron-340B	.477	-.068	1	.076	.076 [.030,.119]	.0027

Table 6: Held-out mitigation comparison against hard BoN-32. Intervals resample learned task–RM cells and keep both folds together.

Population	Cells	Tail cap Δ	Rank-Soft-BoN Δ	Cap $-$ soft
All learned	65	.0085 [.0019,.0168]	.0068 [-.0003,.0156]	.0017 [-.0003,.0036]
Negative tail	6	.0924 [.0519,.1383]	.0916 [.0511,.1339]	.0009 [.0000,.0020]

Among learned cells, the median prompt-level frequency of any tied score is 73.2%, but the median frequency of a tie intersecting the top-four boundary is only 7.0%. The six negative cells remain negative under both deterministic orders and all 200 randomized orders.

F Descriptive Extreme-Answer Features

For the controlled MATH pool, we report four predefined descriptive answer features: log character count, number of equation markers, presence of a `boxed` marker, and whether the final answer is extractable. For each score format, the analysis gives within-prompt Spearman correlation with score and correctness contrasts in the top 1%, 5%, and 10% of score ranks, with prompt-bootstrap intervals. Table 7 summarizes the Tulu top-1% comparison. In the failing answer-only condition, extreme answers are shorter, contain fewer equation markers (contrast -0.88 $[-1.56, -0.23]$), and are less likely to contain boxed or extractable final answers. Masking reverses the correctness and final-answer-marker contrasts. The failure is therefore not a simple verbosity effect. These associations only describe which response forms occupy an extreme tail; they do not identify a causal reward-model shortcut.

Table 7: Descriptive Tulu MATH top-1% features. ρ_L is mean within-prompt Spearman correlation between score and log character count; Δ_L and Δ_B compare top-tail versus remaining log length and boxed-marker frequency; E_1 is correctness excess.

Input	ρ_L	Δ_L [95% CI]	Δ_B [95% CI]	E_1 [95% CI]
QA	-.106	-.020 [-.031,-.009]	+.084 [.069,.100]	+.053 [.034,.073]
Answer-only	-.144	-.029 [-.040,-.018]	-.026 [-.041,-.011]	-.031 [-.048,-.013]
Masked	-.042	+.006 [-.005,.018]	+.067 [.052,.082]	+.048 [.029,.066]
Shuffled	+.050	+.027 [.016,.038]	+.024 [.011,.037]	+.023 [.006,.038]

For MBPP+, code is extracted from the longest Python fence (or the full response if unfenced) and parsed with Python's AST. The expected function name comes from the released reference code. Table 8 reports RM-cell-level contrasts; no confidence interval treats the six repeated views of the shared response pool as independent.

Table 8: MBPP+ incorrect top-tail minus other incorrect response features over the six negative-RM cells. Binary-feature contrasts are proportions.

Feature	Median contrast	RM-cell range
Log character count	+.121	[.025,.150]
Line count	+2.986	[.424,3.850]
Syntax valid	+.018	[.007,.022]
Defines target function	+.021	[.000,.036]
Target-function docstring	+.029	[.007,.037]
Examples or assertions	-.133	[-.217,-.112]
Prose outside code fence	+.128	[.105,.148]
Comment fraction	-.004	[-.038,-.001]