CERTIFY THE EVENT: TARGET GEOMETRY AND MEASUREMENT DEPTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Suppose a frozen generator is randomized by evidence order and decoding, and let Q denote its ordering-specific probability of passing a declared verifier. A service may require $\mathbb{E}Q > p^\star$, while an acquisition heuristic averages a decreasing uncertainty score $\phi(Q)$. We characterize when thresholding that mean score certifies the required event. Using one-moment geometry, we show that the largest distribution-free safe cutoff is the lower convex envelope of ϕ at $p^\star$, and that a mean-score gate is lossless for every ordering law if and only if ϕ is strictly decreasing affine. Support or shape restrictions can recover nonlinear calibration, for which we derive sharp completeness margins. We then turn from population targets to their measurement. With n verifier draws per ordering, the binomial count law identifies exactly the polynomial functionals of degree at most n; thus, increasing the number of orderings at fixed n cannot uniformly recover a nonpolynomial log mean. We derive an unbiased variance decomposition and a count-based jackknife correction, then use homogeneous controls to separate finite-depth bias from genuine ordering heterogeneity. Controlled probability-law experiments exhibit both phenomena without claiming their prevalence in language models. The resulting protocol is target aligned: certify fresh-draw success from fresh verifier bits, and use nonlinear score means only with explicit population calibration and an ordering-clustered measurement audit.

1 INTRODUCTION

A retrieval or agent system can become more confident without becoming reliable enough for its service contract. The converse can also occur: a generator can pass a verifier with high average probability while receiving a poor nonlinear uncertainty score because a small set of evidence orderings performs badly. These observations concern different population functionals, so a confidence signal becomes a certificate only after its target is specified.

We fix an input, model checkpoint, evidence state, decoder, verifier, and law for randomizing the evidence. Once these objects are frozen, only evidence ordering and output sampling remain random. If Q is the success probability conditional on an ordering, our service target is the probability that a *fresh* output under a *fresh* ordering passes the verifier, namely $\mathbb{E}Q$. This is neither the correctness probability of an already selected answer nor automatically the target of an acquisition heuristic.

This distinction is relevant without attributing a certification claim to existing heuristics. FLARE, DRAGIN, Adaptive-RAG, and TASR use model behavior to decide when or how to retrieve (Jiang et al., 2023; Su et al., 2024; Jeong et al., 2024; Kieback et al., 2026); semantic entropy detects an important class of confabulations (Farquhar et al., 2024). They target retrieval utility or error prediction rather than a distribution-free lower bound on fresh verifier-success mass. Our question is instead a reduction problem: if a system thresholds the *mean* of such a score, what event does that threshold protect, and what information is required to measure it?

The paper gives two main answers. First, classical moment geometry yields the sharp safe cutoff for any continuous nonincreasing score and shows that only a strictly decreasing affine score preserves all success-qualified ordering laws. Restricted support, two-level, and fixed-shape families recover useful nonlinear calibrations, with an explicit completeness cost when they do not. Second, finite verifier depth creates a separate failure mode: the distribution of n Bernoulli outcomes per ordering does not identify a general nonlinear functional, and the plug-in log mean can cross the wrong side

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.

of its cutoff even when the ordering law is homogeneous. We connect this fact to a variance decomposition, an interior bias expansion, and a count jackknife. A direct Bernoulli e-process supplies the matched-target operational baseline; standard sequential bounds then quantify the additional evidence cost of choosing a stricter surrogate null.

The numerical evidence is intentionally controlled. It consists of exact calculations and simulations under fully specified probability laws, not a prevalence estimate for LLMs. A fifty-law stress panel and all homogeneous controls are preserved in Appendix M; machine-readable counts and a reproduction script accompany the paper. Appendix N separately records the frozen Qwen/EvalPlus audit protocol and four executed audit tables covering task-level decisions, aggregate checkpoint statistics, mean-matched homogeneous controls and ordering-sensitivity diagnostics.

2 TARGET, OBSERVATION UNIT, AND MATCHED COMPARISONS

At evaluation round t, an acquisition policy freezes its evidence multiset E_t. An ordering π and output $Y \sim S_{t,\pi}$ are then sampled from declared laws. For a fixed verifier $g : \mathcal{Y} \to \{0, 1\}$, define

$$Q = q_{t,\pi} = \mathbb{P}\{g(Y) = 1 \mid \mathcal{H}_t, \pi\}, \qquad m_t = \mathbb{E}_\pi q_{t,\pi}. \tag{1}$$

The law P of Q is a Borel probability law on $[0, 1]$. Fresh evaluation draws are conditionally independent within a frozen round, and the verifier, ordering law, and decoder do not change during evaluation. We suppress t for one state. The success null is $\mathcal{N}_p = \{P : \mathbb{E}_P Q \le p^\star\}$. An error level α controls the probability of declaring $m_t > p^\star$ under this closed null; it does not assert that every accepted output is correct.

One fresh ordering and output per trial yields $X_j \sim \mathrm{Ber}(m_t)$. For a prior Π_t fixed before evaluation,

$$T_{t,n} = \int \prod_{j=1}^{n} \left(\frac{\theta}{p^\star}\right)^{X_j} \left(\frac{1-\theta}{1-p^\star}\right)^{1-X_j} \Pi_t(d\theta), \qquad \theta \in (p^\star, 1]. \tag{2}$$

Accepting at $T_{t,n} \ge 1/(\alpha_d w_t)$, with $w_t = 1/[(t+1)(t+2)]$, controls false certification over all rounds and adaptive within-round sample counts by α_d. A new evidence state starts a new test. For acceptance event A and an independent served output,

$$\mathbb{P}(A, \text{failure}) \le (1 - p^\star)\mathbb{P}(A) + \alpha_d. \tag{3}$$

Appendix G proves both statements. The direct gate is a classical matched-target baseline: it observes exactly the event in Eq. (1). The joint bound in Eq. (3) is distinct from conditional risk, which also depends on certificate coverage.

Table 1 locates nearby methods by population unit and guarantee. C-RAG controls bounded generation risk across calibrated inputs, while TRAQ and MiCP give conformal coverage for retrieval or multi-turn prediction sets (Kang et al., 2024; Li et al., 2024; Zhou et al., 2026). E-valuator tests trajectory success from sequential verifier scores, and certified self-consistency targets modal-answer selection (Sadhuka et al., 2025; Cordero-Encinar & Duncan, 2025). An always-valid release wrapper calibrates adaptive verifier scores against hard negatives and controls release on infeasible workflows, with an MBPP+ case study (Cho & Sun, 2026). These are complementary guarantees, not numerical baselines for Eq. (1). Within our target, Appendix L compares the direct mixture with fixed-look exact binomial procedures on the identical Bernoulli stream.

The foundations are classical moment-set geometry, sequential testing, betting inference, nested Monte Carlo, and binomial-mixture estimation (Winkler, 1988; Wald, 1945; Ville, 1939; Howard et al., 2021; Waudby-Smith & Ramdas, 2024; Rainforth et al., 2018; Tian et al., 2017; Vinayak et al., 2019). Our contribution is their target-aligned synthesis for ordering-dependent generation, not a new conformal method, e-process, Jensen inequality, or jackknife.

3 POPULATION TARGET GEOMETRY

Let $\phi : [0, 1] \to \mathbb{R}$ be continuous, bounded, and nonincreasing, with smaller values indicating greater confidence. A surrogate-mean gate seeks evidence that $\mathbb{E}\phi(Q) < c$. Its largest safe threshold is

$$c^\star_\phi(p^\star) = \inf_{P : \mathbb{E}_P Q \le p^\star} \mathbb{E}_P \phi(Q). \tag{4}$$

Table 1: Positioning by estimand rather than by method name. A valid comparison must match the population unit, null, and observable.

Method	Population unit and target	Statistical object
C-RAG	Future input under a calibration/test law; bounded generation risk	Conformal risk bound
TRAQ / MiCP	Future QA example or reasoning turn; answer-set coverage	Conformal prediction set
E-valuator	Actions within one trajectory; trajectory-level false alarms	Sequential score test
Release wrapper	Adaptive workflow on a task; release under infeasibility	Calibrated e-process
Semantic entropy	Repeated meanings for a prompt; confabulation detection	Predictive uncertainty score
This work	Fresh ordering–output pair from one frozen state; $\mathbb{E}Q > p^\star$	Direct Bernoulli or calibrated mean test

This definition optimizes over the entire ordering law, not only a homogeneous probability. With the strict gate, protecting $\mathcal{N}_\phi = \{P : \mathbb{E}_P\phi(Q) \geq c^\star_\phi\}$ also protects $\mathcal{N}_p$. Equality belongs to both protected nulls.

Theorem 1 (Calibration frontier and lossless mean certificates)**.** *For $p^\star \in (0,1)$, the cutoff in Eq. (4) is the lower convex envelope of ϕ at $p^\star$. It is attained by a distribution supported on at most two probabilities. Any larger cutoff is unsafe for some ordering distribution. Moreover, a threshold c satisfies*

$$\mathbb{E}_P\phi(Q) < c \quad \Longleftrightarrow \quad \mathbb{E}_P Q > p^\star \quad \textit{for every distribution } P \tag{5}$$

if and only if $\phi(q) = a - bq$ with $b > 0$ and $c = a - bp^\star$. Consequently, every score not of this strictly decreasing affine form has a success-qualified distribution inside $\mathcal{N}_\phi$. Every uniformly α-valid test of that surrogate null accepts such a distribution with probability at most α, at any sample size.

Score-only observational non-identifiability is treated separately in Appendix F. The convex-envelope calculation is a classical one-moment problem (Winkler, 1988). The proof pushes every null mean to $p^\star$, applies Jensen's inequality to the greatest convex minorant, and realizes its boundary by a two-point chord. Exact losslessness forces every mean-$p^\star$ chord to have the same slope. Appendix A gives the full argument. The theorem concerns a gate that thresholds $\mathbb{E}\phi(Q)$; it does not say that a nonlinear score is uninformative. An invertible score can be inverted before directly testing $\mathbb{E}Q$.

For clipped log loss and $0 < \delta < 1/2$, put

$$\begin{aligned} u(q) &= -\log \operatorname{clamp}(q; \delta, 1-\delta), \quad \mu = \mathbb{E}u(Q), \\ a_\delta &= -\log(1-\delta), \quad B_\delta = -\log\delta, \quad R_\delta = B_\delta - a_\delta. \end{aligned} \tag{6}$$

When $\mathrm{e}\delta \leq p^\star \leq 1-\delta$, its optimal cutoff is $c^\star_u = -\log p^\star$. The convex envelope is linear up to $q = \mathrm{e}\delta$, agrees with $-\log q$ until $1-\delta$, and is constant thereafter. Appendix A derives this envelope. Thus optimal calibration removes an avoidable clipping correction, but it does not remove the completeness gap created by nonlinear averaging.

3.1 Restricted ordering families

The unrestricted lossless characterization does not exclude useful nonlinear certificates on a smaller ordering family. Consider a justified support $Q \in [a,b]$ with $0 < a < p^\star < b < 1$, and suppose clipping is inactive on this interval. Write $m = \mathbb{E}Q$, $s^2 = \operatorname{Var}(Q)$, and $J = \mathbb{E}[-\log Q] + \log m$.

Proposition 1 (Sharp completeness margin and lossless families)**.** *Over all laws supported on $[a,b]$, the largest safe log cutoff is $-\log p^\star$. Define*

$$m^\star_{a,b} = a + (b-a)\frac{\log(p^\star/a)}{\log(b/a)}. \tag{7}$$

Every such law with $m > m^\star_{a,b}$ satisfies the log criterion. For every $m \in (p^\star, m^\star_{a,b}]$, a law with that mean violates the strict log criterion. Moreover,

$$\frac{s^2}{2b^2} \le J \le \frac{s^2}{2a^2}, \qquad m > p^\star \exp\left(\frac{\sigma^2}{2a^2}\right),\ s^2 \le \sigma^2 \implies \mathbb{E}[-\log Q] < -\log p^\star. \tag{8}$$

If instead the support is known to be exactly a subset of $\{a, b\}$, the family-specific cutoff

$$c_{\{a,b\}}(p^\star) = \frac{b - p^\star}{b - a}(-\log a) + \frac{p^\star - a}{b - a}(-\log b) \tag{9}$$

is lossless. A fixed-shape family $Q = mR$, with a known positive law of R, $\mathbb{E}R = 1$, and admissible m keeping $mR \in [a, b]$, is also lossless at cutoff $-\log p^\star + \mathbb{E}[-\log R]$, provided $p^\star$ is admissible.

The secant bound is the classical Edmundson–Madansky upper bound for a convex function (**?**); the curvature bounds are standard Jensen-gap controls (Gao et al., 2019). For a known two-level family, the expected log score is affine in m; for a fixed-shape family, the Jensen gap is constant. Homogeneous orderings are a special case. Appendix A.1 gives the proof and the symmetric-support expansion. These are population alignment statements, not free access to an ordering's latent success probability.

The distinction between interval and two-level support is essential. Enlarging $\{a, b\}$ to $[a, b]$ admits homogeneous mean-$p^\star$ null states and requires the smaller cutoff $-\log p^\star$. For symmetric support $[p^\star - h, p^\star + h]$, the worst-case excluded success margin is $h^2/(2p^\star) + O(h^4/(p^\star)^3)$, second order in support width. Sample extrema do not establish population support. A restriction must be justified externally, on split data, or with simultaneous error control. Appendix E replaces hard support by explicit uncertainty about tail mass.

4 Evidence cost and finite-depth measurement

4.1 Target-induced cost with exact ordering probabilities

Once the population target is fixed, standard sequential theory separates its intrinsic cost from a conservative confidence boundary. Normalize a nonconstant decreasing score to $Z = (\phi(Q) - \phi(1))/(\phi(0) - \phi(1)) \in [0, 1]$ and write $\eta = (c^\star_\phi - \phi(1))/(\phi(0) - \phi(1))$. The protected null is $\mathbb{E}Z \ge \eta$. For a declared alternative margin $0 \le v < \eta$, let $\mathfrak{T}_\alpha(v, \eta)$ be the smallest worst-case expected acceptance time among sequential tests that are uniformly α-valid under the null and accept almost surely for every iid law with $\mathbb{E}Z \le v$:

$$\mathfrak{T}_\alpha(v, \eta) = \inf_\tau \sup_{P:\mathbb{E}_P Z \le v} \mathbb{E}_P \tau, \qquad d(v, \eta) = \mathrm{KL}(\mathrm{Ber}(v) \| \mathrm{Ber}(\eta)). \tag{10}$$

Theorem 2 (Finite-sample mean-class bounds)**.** *For $0 < v < \eta < 1$ and $A = \log(1/\alpha)$,*

$$\frac{A}{d(v, \eta)} \le \mathfrak{T}_\alpha(v, \eta) \le \frac{A + \log[(1 - v)/(1 - \eta)]}{d(v, \eta)}. \tag{11}$$

The upper guarantee is achieved by stopping when the affine e-process

$$M_n(v) = \prod_{i=1}^{n} \left[Z_i \frac{v}{\eta} + (1 - Z_i) \frac{1 - v}{1 - \eta} \right] \tag{12}$$

reaches $1/\alpha$. The lower bound covers every valid sequential procedure, not only betting tests, and the sandwich is asymptotically sharp as $\alpha \downarrow 0$. At $v = 0$, the earliest deterministic acceptance time is

$$n_\phi = \left\lceil \frac{A}{-\log(1 - \eta)} \right\rceil, \qquad \mathbb{P}_{Z \equiv 0}(\tau \le n) \le \min\{1, \alpha(1 - \eta)^{-n}\}. \tag{13}$$

The endpoint Bernoulli alternative and null give the lower bound by stopped change of measure; an affine e-factor, a chord bound on log growth, and bounded overshoot give the upper bound. These are classical Wald–Ville ingredients (Wald, 1945; Ville, 1939; Waudby-Smith & Ramdas, 2024); Appendix B supplies the proof and comparison with recent general-alternative power results (Durand & Wintenberger, 2025; Sandoval et al., 2026).

Table 2: Exact all-success counts at local error .025, with every score using its largest distribution-free safe cutoff. Log uses $\delta = (1 - p^\star)/4$; near-linear is $(1 - q) + (1 - q)^2$. All columns observe the same exact-Q endpoint path but protect different population nulls.

$p^\star$	Direct	Near-linear	Calibrated log	Squared residual
.50	6	8	11	13
.80	17	29	62	91
.90	36	66	167	368
.95	72	139	415	1474
.99	368	729	2926	36,887
.999	3688	7369	40,766	3,688,878

Write $\varepsilon = 1 - p^\star$. Direct certification has $\eta = \varepsilon$. Optimally calibrated clipped log loss has

$$\eta_u = \frac{-\log p^\star - a_\delta}{R_\delta}. \tag{14}$$

On the common endpoint family $\mathbb{P}(Q = 0) = v$, exact Q reveals the same binary outcome to both gates. With $\delta = \kappa\varepsilon$, the log gate's all-success boundary cost is asymptotically larger by $\log(1/(\kappa\varepsilon))/(1 - \kappa)$; for $(1 - q)^r$, $r > 1$, it is larger by ε^{1-r}. These are target-induced costs, not evidence that one test is statistically inferior for the same null. Appendix K gives exact-recursion controls and finite-budget comparisons.

Table 2 makes the target effect finite-sample and concrete. The direct column is not an oracle shortcut: it is an exact, valid sequential test for the declared success event. The other columns spend more observations because their mean-score nulls contain additional success-qualified laws. Appendix K reports general-alternative bounds, finite-cap acceptance, and risk–coverage curves so that the all-success path is not mistaken for a complete power comparison.

4.2 What finite verifier counts identify

A fresh verifier bit for each fresh ordering is iid $\mathrm{Ber}(m)$, so one bit per ordering is sufficient for the direct target. A nonlinear ordering mean requires repeated outputs within an ordering. Let $S_j \mid Q_j \sim \mathrm{Bin}(n, Q_j)$ for K iid ordering clusters, and write $\widehat{q}_j = S_j/n$.

Theorem 3 (Fixed-depth identifiability)**.** *For fixed n and continuous ϕ, the complete count distribution identifies $\mathbb{E}\phi(Q)$ over all mixing laws on $[0, 1]$ if and only if ϕ is a polynomial of degree at most n. If ϕ is nonpolynomial, two ordering laws have identical count distributions but different ϕ expectations; no number of iid orderings at that same depth yields a uniformly consistent estimator of the functional.*

The binomial probabilities are expectations of the Bernstein basis, which spans the polynomials through degree n. Moment-matched laws give the converse. Appendix D proves the theorem and relates it to binomial-mixture recovery (Tian et al., 2017; Vinayak et al., 2019). The result does not preclude approximation, parametric recovery, or increasing n. The one-bit special case is given in Appendix I.

The law of total variance separates latent heterogeneity from generation noise:

$$\mathrm{Var}(\widehat{q}_j) = \mathrm{Var}(Q_j) + \frac{\mathbb{E}[Q_j(1 - Q_j)]}{n}. \tag{15}$$

For $K, n \geq 2$, an unbiased estimate of $s^2 = \mathrm{Var}(Q)$ is

$$\widehat{s}^2_{\mathrm{unb}} = \frac{1}{K - 1}\sum_{j=1}^{K}(\widehat{q}_j - \overline{q})^2 - \frac{1}{K(n - 1)}\sum_{j=1}^{K}\widehat{q}_j(1 - \widehat{q}_j). \tag{16}$$

Negative finite-sample values must be retained when checking a homogeneous control; truncating at zero changes the estimator.

For clipped log loss on a fixed interval inside the clipping boundaries,

$$\mathbb{E}[u(S/n) \mid Q = q] - u(q) = \frac{1 - q}{2nq} + O(n^{-2}). \tag{17}$$

Thus more orderings at fixed n estimate the wrong plug-in functional more precisely. The count-based delete-one jackknife

$$j_{n,\delta}(s) = nu(s/n) - (n-1)\left[\frac{s}{n}u\left(\frac{s-1}{n-1}\right) + \left(1-\frac{s}{n}\right)u\left(\frac{s}{n-1}\right)\right],$$
$$\widehat{\mu}_{\mathrm{JK}} = K^{-1}\sum_{j=1}^{K} j_{n,\delta}(S_j) \tag{18}$$

cancels the n^{-1} term on an interior interval (Jiao & Han, 2020); zero-coefficient endpoint terms are omitted. It is a bias correction, not an unbiased estimator or a one-sided certificate. Appendix D gives the exact finite-binomial identity and clustered-interval study. In a bits-only model, Appendices H–J also give a truncated inverse-binomial interval for a smoothed log residual and a worst-case $\Omega((\delta w^2)^{-1})$ expected-call lower bound for uniform width w, matching the construction's polynomial dependence up to logarithms. That lower bound concerns residual estimation, not direct target certification.

Not every useful decision requires reconstructing the whole ordering law. Suppose Q, Q' are fresh ordering probabilities and (X_1, X_2) and (Y_1, Y_2) are conditionally independent verifier pairs. Define

$$A = \frac{X_1 + X_2 + Y_1 + Y_2}{4}, \qquad W = \frac{(X_1 - Y_1)(X_2 - Y_2)}{2}. \tag{19}$$

Then $\mathbb{E}A = m$ and $\mathbb{E}W = s^2$. For K independent blocks, Hoeffding intervals with radius $r_K = \sqrt{\log(4/\alpha)/(2K)}$, intersected with the known ranges $[a, b]$ and $[0, (b-a)^2/4]$, give $[m_L, m_U]$ and $[v_L, v_U]$.

Proposition 2 (Observable sufficient decisions)**.** *Under the justified support condition of Proposition 1, the two intervals jointly cover m and s^2 with probability at least $1-\alpha$. On that event,*

$$-\log m_U + \frac{v_L}{2b^2} \le \mathbb{E}[-\log Q] \le -\log m_L + \frac{v_U}{2a^2}. \tag{20}$$

Thus $m_L > p^\star$ and a lower endpoint at least $-\log p^\star$ certify target mismatch; an upper endpoint below $-\log p^\star$ certifies the log criterion. Otherwise the audit remains unresolved. The total cost is $4K$ verifier calls.

Appendix H.1 proves the proposition and specifies its simulation. It gives sufficient decisions, not full identification: even exact knowledge of m and s^2 need not determine the log mean.

5 Controlled numerical evaluation

We now isolate the two failures under fully specified probability laws. Oracle experiments observe exact Q and report their cost in exact-Q ordering observations. Count experiments observe simulated verifier bits. Neither unit is a language-model generation, latency measurement, or deployment speedup. Appendix C records seeds, quadrature, uncertainty, and boundary-null controls; Appendix L records the exact sequential recursions.

5.1 Population mismatch without endpoint atoms

Fix $p^\star = .8$, mean success $m = .805$, and support $[a, b] = [.625, .985]$. For concentration κ, draw

$$Q = a + (b-a)B, \qquad B \sim \mathrm{Beta}(\kappa/2, \kappa/2). \tag{21}$$

Clipping at $\delta = .01$ is inactive. Every law has the same mean, but variance $(b-a)^2/[4(\kappa+1)]$. We evaluate $\kappa \in \{.5, 2, 8, 32, 128\}$. Both gates receive the same iid exact-Q path. The direct gate tests $\mathbb{E}Q > .8$; the log gate tests $\mathbb{E}[-\log Q] < -\log .8$. Each condition has 2000 paired paths, local error .05, and cap 8192.

The first two laws lie in the mismatch band without atoms at 0 or 1. At $\kappa = 2$, the direct target is true and its gate accepts .9310 of paths; the calibrated log target is false and its valid gate accepts .0165. This illustrates different protected events, not inferior testing. At $\kappa = 128$, both targets hold and both gates accept on every simulated path; their restricted mean exact-Q costs are 180.92 and

Table 3: Continuous-ordering simulation at fixed mean .805 and target .8. "Accept" is the pathwise acceptance fraction for each gate's own null; the log population criterion is false at $\kappa = .5, 2$. Costs are restricted mean exact-Q ordering observations $\pm$ standard error, with the cap charged to nonacceptance. Marginal intervals and boundary-null controls are in Appendix C.

κ	Jensen gap	Direct accept	Log accept	Direct cost	Log cost
0.5	0.017034	0.6305	0.0095	5635.94 ± 60.86	8114.41 ± 17.72
2	0.008461	0.9310	0.0165	3624.08 ± 51.45	8060.92 ± 22.72
8	0.002797	1.0000	0.9005	1260.04 ± 21.26	4002.27 ± 57.37
32	0.000759	1.0000	1.0000	311.71 ± 4.52	402.25 ± 5.89
128	0.000194	1.0000	1.0000	180.92 ± 1.10	208.33 ± 1.21

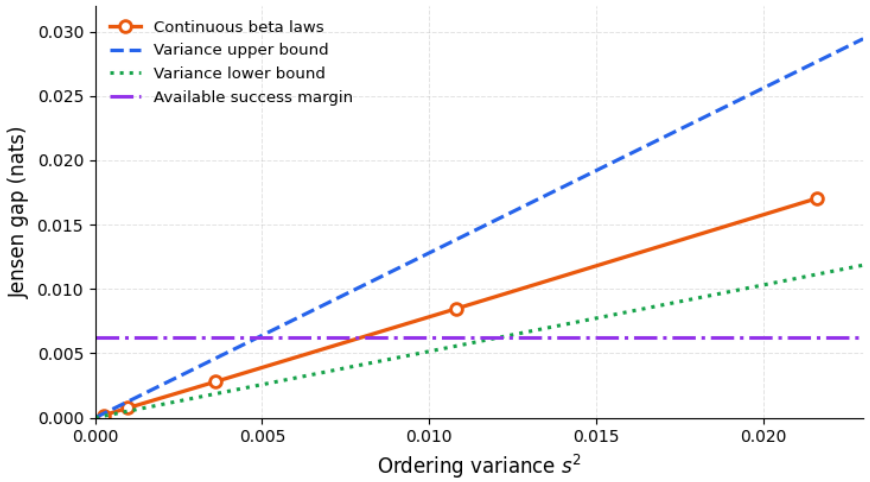

Figure 1: Population target geometry at fixed mean. The log criterion holds exactly when the Jensen gap J lies below $\log(.805/.8)$. Points are quadratures for Table 3; curves are the variance bounds in Proposition 1.

208.33, a ratio of 1.15. Boundary-null controls remain below local error in Table C.2; observed unit acceptance is not a claim of population probability one. Figure 1 illustrates the population target geometry at fixed mean.

For a declared grid of 140 means at each concentration, quadrature identifies $13, 6, 2, 0, 0$ mismatch laws as concentration increases. These counts characterize that designed grid only. Appendix C gives support margins. A complementary four-bit audit estimates the ordering mean and variance without estimating each Q: two fresh orderings and two verifier bits per ordering form unbiased block statistics for both moments. Curvature bounds then give a conservative interval for the log mean Appendix H (Appendix H.1). Table 4 shows the cost of making this audit observable.

The audit costs 4×10^6 simulated verifier bits per law yet resolves only two of five cases. This is a useful negative result: moment bounds can certify clear separation while correctly leaving near-boundary cases unresolved. The experiment does not estimate model prevalence, and the support interval is supplied by the designed law rather than inferred from sample extrema.

5.2 Finite-depth bias and homogeneous controls

Figure 2 evaluates the exact binomial expectation for the homogeneous law $Q = .805$, with clipping $\delta = .01$ and target .8. The true population loss is .216913, below the safe cutoff .223144. At depth $n = 20$, however, the expected plug-in score is .223381 and lies on the wrong side of the cutoff even though no ordering heterogeneity exists. Increasing K cannot remove this fixed-n shift. The jackknife is already close to the truth by $n = 50$, but its small finite-depth bias can have either sign. Figure 2 illustrates the finite-depth transformation bias and jackknife correction.

Table 4: Verifier-only audit of the five continuous laws. Each row uses 10^6 independent four-bit blocks and local error .01, giving simultaneous confidence .95 across the five laws. The cutoff is $-\log .8 = .223144$; every mean interval exceeds .8.

κ	Mean stat.	Variance stat.	Log-mean interval	Decision
0.5	0.804755	0.021803	[0.225413, 0.249493]	Mismatch
2	0.804738	0.010853	[0.219791, 0.235499]	Unresolved
8	0.805179	0.003704	[0.215560, 0.225799]	Unresolved
32	0.804747	0.001244	[0.215079, 0.223188]	Unresolved
128	0.805290	0.000342	[0.214406, 0.221357]	Log criterion

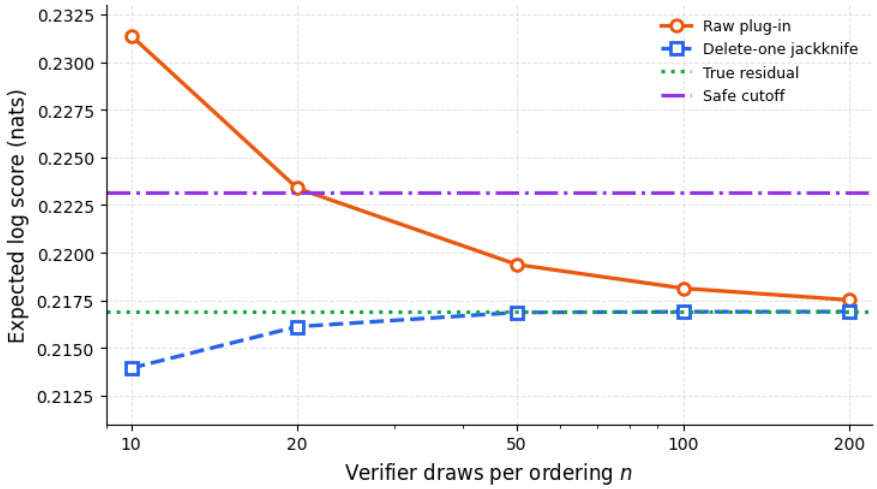

Figure 2: A homogeneous control isolates finite-depth transformation bias. Values are exact binomial expectations, not Monte Carlo estimates. The raw plug-in loss crosses the cutoff at $n = 20$ although the population criterion holds; the delete-one jackknife removes the leading interior bias but is not a certificate.

Outer uncertainty remains after bias correction. In 2000 independent datasets with $K = 50$ orderings and $n = 200$ bits per ordering, the homogeneous raw mismatch flag occurs 46 times and the corrected flag once. Across five heterogeneous laws and this control, ordering-clustered percentile intervals cover the true log mean in .9345–.9455 of datasets and the true variance in .9095–.9435; the nominal intervals therefore under-cover in some conditions. No dataset at the narrow .805 versus .8 margin has both lower bounds needed to establish mismatch. Full depth sweeps and clustered results appear in Appendix D.

5.3 Secondary finite-law stress panel

To check the measurement pipeline away from the narrow boundary, we use fifty designed probability laws D01–D50, each with $K = 100$ orderings and $n = 200$ simulated bits, plus a population-mean-matched homogeneous control. For target p, let $G_p = \log(m/p)$ and $R_p = \mu + \log p$. A descriptive mismatch requires both ordering-clustered lower endpoints to exceed $\gamma = .05$ nats; alignment requires $G_{p,L} > \gamma$ and $R_{p,U} < -\gamma$; below-target requires $G_{p,U} < -\gamma$; all other laws remain unresolved. These percentile classifications are marginal summaries, not simultaneous or anytime-valid certificates. Table 5 summarizes the stress-panel classifications.

At target .5, only D12 and D22 meet the separated mismatch rule, while 12 laws remain unresolved. No homogeneous control is classified as mismatch at any target. The median unbiased ordering-variance estimate is .000462 for the D laws; $32/50$ marginal variance intervals have a positive lower endpoint, compared with $0/50$ controls. The median raw-to-jackknife correction is .002682 nats. Appendix M gives the law families, complete protocol, and data hashes. These outcomes describe a designed finite collection, not a benchmark prevalence.

Table 5: Fifty-law stress panel and matched homogeneous controls at a .05-nat separation margin. Every row retains all 50 designed laws.

Arm / p	Mismatch	Aligned	Below	Unresolved	Unresolved fraction
D laws / .5	2	28	8	12	24%
D laws / .7	0	17	17	16	32%
D laws / .9	0	2	32	16	32%
Homogeneous / .5	0	37	8	5	10%
Homogeneous / .7	0	25	19	6	12%
Homogeneous / .9	0	2	35	13	26%

For the model-level audit, Appendix N specifies the checkpoint, benchmark version, task-selection rule, ordering law, decoding law, verifier, and cluster-level analysis. It also reports four executed audit tables covering task-level decisions, aggregate checkpoint statistics, mean-matched homogeneous controls, and ordering-sensitivity diagnostics. These results are analyzed under the stated audit protocol and are not interpreted as prevalence estimates beyond the evaluated checkpoint and benchmark setting.

The separation rule is stronger than population mismatch. When clipping is inactive, $G_p + R_p = J$, so both population margins can hold only for targets in

$$e^{\gamma-\mu} < p < me^{-\gamma}. \tag{22}$$

This interval is nonempty exactly when $J > 2\gamma$. At $\gamma = .05$, even exact population knowledge requires a Jensen gap above .10 nats. Unresolved outcomes therefore combine limited ordering precision with an intentionally strong separation requirement; they are not estimator failures by definition.

Optional stopping is a different issue. At target .8, reusing an unadjusted fixed-look exact binomial test crosses under the boundary null with probability .170818, versus .013890 for a sixteen-look Bonferroni correction and .036750 for the direct mixture. These exact-recursion results in Table L.2 compare procedures on the same null and observable, unlike Table 3.

6 Discussion and conclusion

A calibrated score can still change the event being certified. The convex-envelope frontier identifies the strongest safe mean cutoff, while restricted-family results specify when that cutoff is lossless or incurs a bounded completeness cost. Statistical efficiency begins only after this target choice. Better confidence sequences can shorten a valid test, but they cannot accept laws that the protected surrogate null intentionally contains. Conversely, Theorem 1 does not preclude nonlinear scores: an invertible score can be inverted, and a predictive score can remain useful for retrieval or ranking. The restriction concerns universal certification through its mean.

Measurement introduces a distinct layer. More independent orderings reduce uncertainty about the ordering distribution; more draws within an ordering reduce transformation bias and better resolve its latent probability. At fixed total bits, these objectives compete. For the pooled mean, the ordering component scales as s^2/K, whereas the within-order component is $\mathbb{E}[Q(1-Q)]/(Kn)$. The jackknife addresses the interior n^{-1} bias but does not create new ordering clusters or provide one-sided validity. Homogeneous controls, untruncated variance estimates, and ordering-clustered intervals therefore provide complementary safeguards for an empirical order-sensitivity audit.

The controlled results use fully specified probability laws, making the relevant support and shape assumptions explicit. For model data, such restrictions require external or error-controlled validation, and any verifier certificate remains relative to the chosen verifier. Returning one candidate that happens to pass and certifying the next randomized output are different service contracts. For the latter, fresh verifier outcomes align measurement directly with the certified event. A nonlinear surrogate mean is justified only when its population calibration, measurement depth, clustering, and unresolved cases are explicitly accounted for.

REPRODUCIBILITY STATEMENT

The appendices provide complete proofs, probability laws, finite sums, mixture weights, quadrature rules, random seeds, observation units, and cost definitions. The numerical experiments were executed only on these specified probability models. The accompanying CSV contains all $10{,}000$ ordering counts for the D01–D50 and homogeneous arms, and the accompanying Python script reproduces Table 5, per-law point estimates, bootstrap intervals, and classifications from that ledger. Appendix M records the corresponding hash, seed, and law definitions. Appendix N specifies the frozen model audit protocol and reports the executed checkpoint tables, including task-level decisions, aggregate statistics, homogeneous controls, and ordering-sensitivity diagnostics.

AI USE STATEMENT

Generative AI assisted with manuscript organization, literature comparison, mathematical derivation, proof checking, and numerical implementation. All reported probability-law programs, calculations, and model-audit analyses were executed as described in the manuscript. The authors reviewed and verified the resulting text, analyses, and reported results.

REFERENCES

Young Hyun Cho and Will Wei Sun. When should an AI workflow release? always-valid inference for black-box generate-verify systems. arXiv:2605.12947, 2026. URL https://arxiv.org/abs/2605.12947.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Paula Cordero-Encinar and Andrew B. Duncan. Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in LLMs. arXiv:2510.17472, 2025. URL https://arxiv.org/abs/2510.17472.

Amaury Durand and Olivier Wintenberger. Power comparison of sequential testing by betting procedures. *arXiv preprint arXiv:2504.00593*, 2025. URL https://arxiv.org/abs/2504.00593.

S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Xiang Gao, Meera Sitharam, and Adrian E. Roitberg. Bounds on the Jensen gap, and implications for mean-concentrated distributions. *The Australian Journal of Mathematical Analysis and Applications*, 16(2):1–16, 2019. URL https://arxiv.org/abs/1712.05267. Article 14.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. Ann. Statist., 49(2), 2021.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-Coder technical report. *arXiv preprint arXiv:2409.12186*, 2024. URL https://arxiv.org/abs/2409.12186.

S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented LLMs through question complexity. NAACL, 2024.

Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig. Active retrieval augmented generation. EMNLP, 2023.

Jiantao Jiao and Yanjun Han. Bias correction with jackknife, bootstrap, and taylor series. *arXiv preprint arXiv:1709.06183*, 2020. URL https://arxiv.org/abs/1709.06183v4. Version 4.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-RAG: Certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22963–23000, 2024. URL https://proceedings.mlr.press/v235/kang24a.html.

Adrian Kieback, Uyiosa Philip Amadasun, Aman Chadha, and Aaron Elkins. TASR: Training-free adaptive stopping for iterative retrieval. arXiv:2606.13814, 2026. URL https://arxiv.org/abs/2606.13814. Agent4IR Workshop, KDD 2026.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3799–3821, 2024. doi: 10.18653/v1/2024.naacl-long.210. URL https://aclanthology.org/2024.naacl-long.210/.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*, 2023. URL https://arxiv.org/abs/2305.01210.

Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting Monte Carlo estimators. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4267–4276. PMLR, 2018. URL https://proceedings.mlr.press/v80/rainforth18a.html.

Shuvom Sadhuka, Drew Prinster, Clara Fannjiang, Gabriele Scalia, Bonnie Berger, Aviv Regev, and Hanchen Wang. E-valuator: Reliable agent verifiers with sequential hypothesis testing. arXiv:2512.03109, 2025. URL https://arxiv.org/abs/2512.03109.

Ricardo J. Sandoval, Ian Waudby-Smith, and Michael I. Jordan. Multi-armed sequential hypothesis testing by betting. *arXiv preprint arXiv:2603.17925*, 2026. URL https://arxiv.org/abs/2603.17925.

W. Su, Y. Tang, Q. Ai, Z. Wu, and Y. Liu. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models. ACL, 2024.

Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. *arXiv preprint arXiv:1709.02707*, 2017. URL https://arxiv.org/abs/1709.02707.

Bas van Opheusden, Luigi Acerbi, and Wei Ji Ma. Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *PLOS Computational Biology*, 16(12):e1008483, 2020. doi: 10.1371/journal.pcbi.1008483. URL https://arxiv.org/abs/2001.03985.

J. Ville. Étude critique de la notion de collectif. Gauthier-Villars, 1939. URL https://www.numdam.org/item/THESE_1939__218__1_0/. Thèse, no. 218, 124 pages.

Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6448–6457. PMLR, 2019. URL https://proceedings.mlr.press/v97/vinayak19a.html.

A. Wald. Sequential Tests of Statistical Hypotheses. Ann. Math. Statist., 16:256–298, 1945. URL https://api.semanticscholar.org/CorpusID:222593486.

I. Waudby-Smith and A. Ramdas. Authors' reply to the Discussion of "Estimating means of bounded random variables by betting". J. Royal Stat. Soc. B, 86(1):53–61, 2024.

Gerhard Winkler. Extreme points of moment sets. *Mathematics of Operations Research*, 13(4):581–587, 1988. doi: 10.1287/moor.13.4.581. URL https://pubsonline.informs.org/doi/10.1287/moor.13.4.581.

Xiaofan Zhou, Huy Nguyen, Bo Yu, Chenxi Liu, and Lu Cheng. Adaptive stopping for multi-turn LLM reasoning. arXiv:2604.01413, 2026. URL https://arxiv.org/abs/2604.01413.

A CALIBRATION FRONTIER AND EXACT TARGET ALIGNMENT

We first make the convex-envelope statement explicit. For a continuous function ϕ, write $\operatorname{co}\phi$ for its greatest convex minorant on $[0,1]$. For a nonincreasing ϕ and any $p \in (0,1)$,

$$c_\phi^\star(p) = \min_{0\le x\le p\le y\le 1} \frac{(y-p)\phi(x) + (p-x)\phi(y)}{y-x}, \tag{23}$$

where the degenerate choice $x = y = p$ is interpreted as $\phi(p)$. This is the lower boundary of the convex hull of the graph at horizontal coordinate p. Compactness and continuity ensure attainment. Only two support points are needed for this one-moment constraint; this is a special case of the extremal structure of moment sets (Winkler, 1988).

Proof of Theorem 1. Let a distribution have mean $m < p$. Replacing Q by $Q' = Q + s(1-Q)$, with $s = (p-m)/(1-m)$, increases every probability and gives $\mathbb{E}Q' = p$. Since ϕ is nonincreasing, $\mathbb{E}\phi(Q') \le \mathbb{E}\phi(Q)$. Thus the infimum over means at most p equals the infimum over mean exactly p. For every such distribution, Jensen's inequality gives

$$\mathbb{E}\phi(Q) \ge \mathbb{E}[(\operatorname{co}\phi)(Q)] \ge (\operatorname{co}\phi)(p).$$

Conversely, a chord realizing the lower hull at p supplies a two-point distribution with mean p and expected score $(\operatorname{co}\phi)(p)$. This proves Eq. (23) and the first assertion. Increasing the cutoff above this attained infimum would accept that same mean-p distribution at the population level, so the resulting implication would be unsafe.

For the lossless characterization, suppose Eq. (5) holds. Every mean-p law has expected score at least c. Mixing it with any positive mass on $Q = 1$ produces a mean strictly above p, so the new expected score is strictly below c. Sending the added mass to zero shows that the original mean-p score is at most c, hence equal to c. In particular, $\phi(p) = c$.

For any $x < p < y$, the unique mean-p mixture on $\{x, y\}$ therefore satisfies

$$\frac{\phi(x) - c}{p - x} = \frac{c - \phi(y)}{y - p}.$$

Fixing one $y > p$ shows that the ratio on the left is the same for all $x < p$. Fixing one such x then shows the same constant for all $y > p$. Consequently, $\phi(q) = c + b(p - q)$ for every $q \in [0,1]$. For a point mass above p, Eq. (5) requires its score to be below c, so $b > 0$. The converse follows immediately by taking expectations of an affine function.

Finally, the cutoff $c_\phi^\star$ already makes $\mathcal{N}_p \subseteq \mathcal{N}_\phi$. If ϕ is not strictly decreasing affine and there were no law with $\mathbb{E}Q > p$ and $\mathbb{E}\phi(Q) \ge c_\phi^\star$, the two sets would be equal, contradicting the characterization. Such a law therefore exists. Uniform type-I control for the surrogate null gives acceptance probability at most α under that law, regardless of the stopping horizon. □

The last assertion concerns a certificate for the surrogate mean itself. It does not imply that all information carried by a nonlinear score is inadequate. If ϕ is invertible, observing $\phi(Q)$ reveals Q, and a procedure can test $\mathbb{E}Q$ directly. Nor does the characterization preclude a lossless threshold on a smaller family of ordering distributions. For example, restricting to homogeneous probabilities can remove the relevant mixtures. The unrestricted ordering family is essential to the theorem.

A.1 RESTRICTED CALIBRATION AND SHARP COMPLETENESS

Proof of Proposition 1. Let $f(q) = -\log q$. Convexity and the secant inequality imply, for every $m = \mathbb{E}Q \in [a,b]$,

$$-\log m \le \mathbb{E}f(Q) \le S_{a,b}(m), \qquad S_{a,b}(m) = \frac{b-m}{b-a} f(a) + \frac{m-a}{b-a} f(b). \tag{24}$$

The lower endpoint is attained by $Q = m$ almost surely. The upper endpoint is attained by the unique mixture on $\{a, b\}$ with mean m. A homogeneous mean-$p^\star$ null state proves that $-\log p^\star$ is the largest safe cutoff over all interval-supported laws.

The secant is strictly decreasing. Solving $S_{a,b}(m) = -\log p^\star$ gives Eq. (7). Strict convexity gives $S_{a,b}(p^\star) > -\log p^\star$ and $S_{a,b}(b) = -\log b < -\log p^\star$, so $p^\star < m^\star_{a,b} < b$. For any larger mean, the upper bound in Eq. (24) is below the cutoff. For every mean in $(p^\star, m^\star_{a,b}]$, the endpoint mixture attains a residual at least as large as the cutoff. This proves both completeness and sharpness, including the equality case under the protected residual null.

Since $f''(q) = q^{-2}$ on $[a, b]$, Taylor's formula with integral remainder gives

$$\frac{(q-m)^2}{2b^2} \le f(q) - f(m) - f'(m)(q-m) \le \frac{(q-m)^2}{2a^2}.$$

Taking expectations yields Eq. (8). If $s^2 \le \sigma^2$ and $\log(m/p^\star) > \sigma^2/(2a^2)$, then $\mathbb{E}f(Q) = -\log m + J < -\log p^\star$. This sufficient condition can be vacuous when its required mean exceeds b; it does not presume that every variance bound is informative.

On the family supported on $\{a, b\}$, the upper bound in Eq. (24) is always an equality. Consequently $\mathbb{E}f(Q) < S_{a,b}(p^\star)$ is equivalent to $m > p^\star$. The cutoff is maximal because the mean-$p^\star$ mixture is in this family. In fact,

$$\frac{f(Q) - f(b)}{f(a) - f(b)} = \frac{b - Q}{b - a} \qquad (Q \in \{a, b\}),$$

so the family-calibrated normalized residual and direct probability scores coincide pathwise. Identical bets and error allocations therefore give identical oracle stopping times on this family. This identity concerns exact ordering probabilities, not arbitrary black-box verifier access.

For $Q = mR$, positivity and bounded support make all log expectations finite, and $\mathbb{E}R = 1$ gives $\mathbb{E}Q = m$. The identity $\mathbb{E}[-\log(mR)] = -\log m + \mathbb{E}[-\log R]$ shows that subtracting the family-constant gap leaves exactly $-\log m$. Hence the stated cutoff is lossless. At admissible $m = p^\star$ it is attained, proving maximality. Any calibration uncertainty about the law of R must be handled separately rather than treated as known. □

For symmetric support, put $x = h/p^\star$. Equation (7) yields

$$m^\star_{p^\star - h, p^\star + h} - p^\star = h\left[\frac{2[-\log(1-x)]}{\log(1+x) - \log(1-x)} - 1\right] = \frac{h^2}{2p^\star} + \frac{h^4}{12(p^\star)^3} + O\left(\frac{h^6}{(p^\star)^5}\right).$$

This expansion holds as $h/p^\star \to 0$. A narrow interval therefore limits the population exclusion to a second-order mean margin without requiring an endpoint assumption. A known two-level family is stronger: its correctly recalibrated exclusion is exactly zero even when $b - a$ is large.

A.2 Optimal clipped-log calibration

Assume $\mathrm{e}\delta \le 1 - \delta$. The convex envelope of the clipped log loss in Eq. (6) is

$$(\mathrm{co}\, u)(q) = \begin{cases} B_\delta - q/(\mathrm{e}\delta), & 0 \le q \le \mathrm{e}\delta, \\ -\log q, & \mathrm{e}\delta \le q \le 1 - \delta, \\ a_\delta, & 1 - \delta \le q \le 1. \end{cases} \tag{25}$$

To verify this formula, the line from $(0, B_\delta)$ is tangent to $-\log q$ at $q = \mathrm{e}\delta$, since $\log[(\mathrm{e}\delta)/\delta] = 1$. It is below the flat lower-clipping segment and below the log curve by its supporting-line property. At $1 - \delta$, the derivative increases from $-1/(1-\delta)$ to zero. The displayed function is therefore convex and is a minorant of u. Every convex minorant is bounded by the displayed chord on $[0, \mathrm{e}\delta]$ and by u thereafter, proving maximality. In particular, $c^\star_u(p) = -\log p$ on the middle interval.

This calculation distinguishes the population threshold from an estimation correction. For a mean-p homogeneous state in the middle interval, $\mathbb{E}u(Q) = -\log p$, so no larger uniformly safe cutoff is possible. The sufficient clamping correction $-\log(p + \delta)$ is avoidable at these targets. The high-reliability separation in Corollary 1 uses the larger, optimal threshold and therefore cannot be attributed to that correction.

A.3 Additional surrogate families

For $\phi_r(q) = (1-q)^r$, $r \geq 1$, convexity gives $c_r^\star(p) = (1-p)^r$. The normalized endpoint cutoff is $\eta_r = \varepsilon^r$. For $\phi(q) = (1-q) + \lambda(1-q)^2$, $\lambda > 0$, it is

$$c_\phi^\star(p) = \varepsilon + \lambda\varepsilon^2, \qquad \eta_\phi = \frac{\varepsilon + \lambda\varepsilon^2}{1+\lambda}.$$

These scores are transforms of an ordering's success probability. In particular, the squared residual here is the loss relative to a positive reference event; it is not the expected Brier risk over observed success and failure labels. The distinction matters because the latter is a different functional.

For entropy, monotonicity fails. The general safe threshold remains $\inf_{P:\mathbb{E}Q\leq p} \mathbb{E}\phi(Q)$, but the restriction to mean exactly p need not preserve that infimum. With $\phi(q) = -q\log q - (1-q)\log(1-q)$, the null state $Q = 0$ gives the infimum zero. Since entropy is nonnegative, no strict lower-than-threshold entropy gate can certify any state at that uniformly safe cutoff. This is an event-identification obstruction, not a sample-size penalty.

B General-alternative minimax proof

Corollary 1 (High-reliability penalty after optimal calibration)**.** *Fix $\kappa \in (0, 1/2)$ and let $\delta = \kappa\varepsilon$. For fixed $\alpha \in (0,1)$, the optimal all-success counts satisfy*

$$\frac{n_u}{n_{\mathrm{dir}}} \sim \frac{\log(1/(\kappa\varepsilon))}{1-\kappa}. \tag{26}$$

For any fixed $\rho \in (0,1)$, the ratio of the mean-class minimax times at $v = \rho\eta_u$ converges, as $\alpha \downarrow 0$, to $d(v,\varepsilon)/d(v,\eta_u)$. Subsequently, as $\varepsilon \downarrow 0$,

$$\frac{d(\rho\eta_u, \varepsilon)}{d(\rho\eta_u, \eta_u)} \sim \frac{\log(1/(\kappa\varepsilon))}{(1-\kappa)(1-\rho+\rho\log\rho)}. \tag{27}$$

For the power residual $\phi_r(q) = (1-q)^r$, $r > 1$, the optimal cutoff is ε^r and the all-success ratio is asymptotic to ε^{1-r}.

Proof of Theorem 2 for $0 < v < \eta$. For the lower bound, choose the endpoint alternative $P_v = \mathrm{Ber}(v)$ and null $P_\eta = \mathrm{Ber}(\eta)$ for Z. Let τ be any test in $\mathcal{T}_\alpha$. If $\mathbb{E}_v\tau = \infty$, the lower bound is immediate. Otherwise, the stopped log likelihood has expectation

$$\mathbb{E}_v\left[\sum_{i=1}^{\tau} \log\frac{dP_v}{dP_\eta}(Z_i)\right] = d(v,\eta)\mathbb{E}_v\tau.$$

The increments are bounded because both probabilities lie strictly between zero and one; the stopped identity follows by summing their predictable inclusion indicators. Under the alternative, $\tau < \infty$ almost surely. Under the null, this event has probability at most α. Data processing of the stopped experiment to that event gives

$$d(v,\eta)\mathbb{E}_v\tau \geq d(1, \mathbb{P}_\eta\{\tau < \infty\}) \geq \log(1/\alpha).$$

One can equivalently apply the finite-horizon identity to $\tau \wedge n$ and the event $\{\tau \leq n\}$, then let n increase. Thus the lower bound covers arbitrary stopping rules, not only likelihood-ratio rules. Because the endpoint alternative belongs to the mean class, it also lower-bounds its minimax cost.

For the upper bound, define

$$f_v(z) = z\frac{v}{\eta} + (1-z)\frac{1-v}{1-\eta} = 1 + \frac{\eta - v}{\eta(1-\eta)}(\eta - z).$$

It is positive on $[0,1]$, and its expectation under any null law with $\mathbb{E}Z \geq \eta$ is at most one. Hence $M_n(v) = \prod_i f_v(Z_i)$ is a nonnegative supermartingale under every iid null law, and Ville's inequality gives uniform validity.

Concavity of $\log f_v$ gives the chord inequality

$$\log f_v(z) \geq z \log \frac{v}{\eta} + (1-z) \log \frac{1-v}{1-\eta}.$$

For an alternative with mean $m \leq v$, the right-hand expectation is at least its value at $m = v$, since $\log(v/\eta) < 0 < \log[(1-v)/(1-\eta)]$. Consequently, every alternative log-increment mean is at least $d(v, \eta) > 0$. Write $S_n = \log M_n(v)$ and $b = \log[(1-v)/(1-\eta)]$. For every finite n, predictable stopping gives

$$d(v, \eta)\mathbb{E}_P(\tau_v \wedge n) \leq \mathbb{E}_P S_{\tau_v \wedge n} \leq A + b.$$

The final inequality holds pathwise: before crossing $S_n < A$, and the crossing increment is at most b. Monotone convergence yields the upper bound on expected time. It also gives $\tau_v < \infty$ almost surely, so the displayed test belongs to $\mathcal{T}_\alpha$. The endpoint law makes the log-chord bound an equality, accounting for the matching information rate. □

For $v = 0$, the alternative class contains only $Z = 0$ almost surely. Under the least-favorable null $\mathrm{Ber}(\eta)$, a length-n all-zero history has probability $(1-\eta)^n$ and is observationally identical to that alternative history. Coupling the private seed therefore gives $\alpha \geq (1-\eta)^n \mathbb{P}_0(\tau \leq n)$. Summing over disjoint first-acceptance events gives $\alpha \geq \mathbb{E}_0(1-\eta)^\tau$ when alternative acceptance is almost sure. Jensen then proves the expected-time lower bound. The endpoint process $\prod_i (1-Z_i)/(1-\eta)$ is valid and reaches the boundary at the integer count in Eq. (13). Exact attainability is asserted for that deterministic count, not for a nonintegral expectation lower bound.

When the test observes Q rather than only its normalized score, the lower bound remains available for the continuous monotone score family: $Q = 1$ and $Q = 0$ realize the score endpoints, so the same two-point statistical experiment lies inside the allowed class. The upper-bound test simply uses the normalized score. Additional access to a completely known ordering-probability table is a different experiment and is not covered by this sampling lower bound.

Proof of Corollary 1. With $p^\star = 1 - \varepsilon$ and $\delta = \kappa\varepsilon$,

$$-\log p^\star - a_\delta = \log \frac{1-\kappa\varepsilon}{1-\varepsilon} = (1-\kappa)\varepsilon + O(\varepsilon^2), \qquad R_\delta \sim \log \frac{1}{\kappa\varepsilon}.$$

Thus $\eta_u \sim (1-\kappa)\varepsilon / \log(1/(\kappa\varepsilon))$. Substituting in Eq. (13), using $-\log(1-x) \sim x$, proves Eq. (26); integer rounding is negligible because both counts diverge.

For fixed $v < \eta$, Eq. (11) gives $\mathfrak{T}_\alpha(v, \eta)/\log(1/\alpha) \to 1/d(v, \eta)$ as $\alpha \downarrow 0$. Apply this separately to the residual threshold η_u and the direct threshold ε at $v = \rho\eta_u$. These are mean-class minimax quantities whose common least-favorable endpoint law has bad-order mass v; they are not an instancewise optimality assertion for every interior ordering distribution.

For fixed $\rho \in (0, 1)$, as $\eta \downarrow 0$,

$$d(\rho\eta, \eta) = \eta(1 - \rho + \rho \log \rho) + O(\eta^2).$$

Since η_u/ε is of order $1/\log(1/\varepsilon)$,

$$d(\rho\eta_u, \varepsilon) = \varepsilon + o(\varepsilon).$$

The first Bernoulli KL term is of order $\eta_u \log(\varepsilon/\eta_u) = o(\varepsilon)$, and expansion of the second term gives the leading ε. Their ratio proves Eq. (27).

For ϕ_r, the endpoint cutoff is exactly $\eta_r = \varepsilon^r$, so the boundary ratio is asymptotic to $\varepsilon/\eta_r = \varepsilon^{1-r}$. At margin $v = \rho\eta_r$, the same expansion yields the additional factor $(1 - \rho + \rho \log \rho)^{-1}$. For the near-linear family, $\eta_\phi \sim \varepsilon/(1+\lambda)$, giving boundary ratio $1 + \lambda$. □

At optimally calibrated log loss, the interval $\eta_u < v < \varepsilon$ is nonempty whenever $\mathrm{e}\delta \leq p^\star < 1 - \delta$ and $\delta > 0$. Indeed, the mean-$p^\star$ endpoint mixture has loss strictly above the convex-envelope value $-\log p^\star$, because the convex envelope is not affine on the full segment $[0, 1]$. Hence $\eta_u < \varepsilon$. Each such v is a strict target mismatch. Under a valid residual test its eventual acceptance probability is at most α, even though a consistent direct Bernoulli test can have power one.

C CONTINUOUS ORDERING LAWS AND RESTRICTED CALIBRATION

The continuous experiment uses the exact probabilities of a mathematical ordering model. Its sample unit is an independently drawn latent ordering probability Q, observed by both gates. An oracle observation is not counted as an ordinary generator call. Within a row, the paired procedures use exactly the same sequence of Q values, including a continued path after either stopping rule accepts; only its prefix through acceptance or the cap contributes to that procedure's cost.

For interval $[a, b]$ and target p, the normalized scores and thresholds are

$$Z_{\mathrm{dir}} = \frac{b-Q}{b-a}, \quad \eta_{\mathrm{dir}} = \frac{b-p}{b-a}, \qquad Z_{\log} = \frac{\log(b/Q)}{\log(b/a)}, \quad \eta_{\log} = \frac{\log(b/p)}{\log(b/a)}.$$

Every score lies in $[0, 1]$. Each gate uses the equal-weight mixture of Eq. (12) over $v = \eta(1 - z)$, where

$$z \in \{.01, .025, .05, .1, .25, .5, .75, .9, 1\}.$$

These values are fixed for all rows and both gates. Every observation is inspected. A run accepts at the first mixture value at least 20, or receives cost 8192 on nonacceptance. A threshold crossing at the last observation is counted as acceptance with that same capped cost. The endpoint component is zero after a zero factor and stays in the mixture; it is not reset or removed after observing failures.

In Table 3, clipping at $\delta = .01$ is inactive and $Q = .625 + .36B$ with $B \sim \mathrm{Beta}(\kappa/2, \kappa/2)$. We computed each log expectation by normalized 256-node Gauss–Jacobi quadrature, with beta exponents matching the weight function. Means and variances have their analytical beta-distribution values. The population cutoff is $-\log .8$, not a quantile chosen after the simulated paths. The gap values are also checked against both curvature inequalities of Proposition 1.

Concentrations are $.5, 2, 8, 32, 128$ in that order. Their independent NumPy random-generator seeds are $20260924 + i$ for zero-based row index i. Each row generates 2000 iid paths of length 8192 in blocks of 128 observations. The seed, block shape, and replication count determine the draw order. Both stopping times and acceptance indicators are evaluated for every replication. The displayed 40-bin histograms use the first block, consisting of 256000 ordering draws per condition.

We report the sample mean of capped costs, its standard error, and the paired cost difference with its standard error. Power intervals use exact binomial inversion (Clopper & Pearson, 1934). Neither the power interval nor its lack of overlap with another row establishes a multiplicity-adjusted comparison. The null control repeats the procedure at mean $.8$ and concentrations $.5, 8, 128$, with seeds $20261024 + i$. This changes the beta shapes to κt and $\kappa(1 - t)$, where $t = (.8 - .625)/.36$; it does not change the gates. Table C.1 reports the corresponding uncertainty estimates and marginal power intervals. Table C.3 summarizes the analytical support frontiers and corresponding cutoffs.

Table C.1: Uncertainty for Table 3. Power intervals are marginal two-sided 95% Clopper–Pearson intervals across 2000 independent paths. SE is the sample standard deviation of capped path cost divided by $\sqrt{2000}$, measured in exact ordering observations. These intervals are not simultaneous across rows or methods.

κ	Direct power interval	Log power interval	Direct cost SE	Log cost SE
0.5	[0.6089, 0.6517]	[0.0057, 0.0148]	60.86	17.72
2	[0.9190, 0.9417]	[0.0114, 0.0231]	51.45	22.72
8	[0.9982, 1.0000]	[0.8865, 0.9133]	21.26	57.37
32	[0.9982, 1.0000]	[0.9982, 1.0000]	4.52	5.89
128	[0.9982, 1.0000]	[0.9982, 1.0000]	1.10	1.21

The finite mismatch grid contains 140 equally spaced means $.801, .802, \ldots, .940$ at each of the five concentrations. For every mean m, set $t = (m - .625)/.36$ and draw $B \sim \mathrm{Beta}(\kappa t, \kappa(1 - t))$. Thus the grid retains the same interval support and success-qualified means while changing shape in a declared manner. Quadrature gives respectively 13, 6, 2, 0, and 0 strict-or-boundary residual mismatches. This is $21/700$ of the declared grid; no distribution over real tasks is inferred from that design fraction. The grid includes every nonmismatch state in its denominator.

The fourth support row in Table C.3 deliberately allows low-probability orderings. Its large excluded mean margin shows why a wide support assumption alone may not be useful. The first three rows

Table C.2: Continuous null controls at mean success and target .8, with the same support, cap, and error allocation as Table 3. Both gates are under their protected nulls; the log null is interior because the Jensen gap is positive. Each row uses 2000 independent paired paths. Intervals are marginal 95% Clopper–Pearson intervals for crossing probability, not simultaneous validity guarantees.

κ	Direct crossings	Direct interval	Log crossings	Log interval
0.5	78/2000	[0.0309, 0.0484]	13/2000	[0.0035, 0.0111]
8	47/2000	[0.0173, 0.0311]	10/2000	[0.0024, 0.0092]
128	54/2000	[0.0203, 0.0351]	17/2000	[0.0050, 0.0136]

Table C.3: Analytical positive counterparts at target .8. For all laws on the displayed interval, the safe cutoff is $-\log .8 = .22314355$, and every mean strictly above the second column passes it. The third column is $m^\star_{a,b} - .8$. When the support is restricted to the two displayed endpoints, the fourth-column cutoff is lossless instead. Applying that larger cutoff to unrestricted interval laws is not valid.

Support	Complete above $m^\star_{a,b}$	Excluded mean margin	Two-level cutoff
[0.625, 0.985]	0.82036510	0.02036510	0.24887655
[0.75, 0.85]	0.80156352	0.00156352	0.22510050
[0.79, 0.81]	0.80006250	0.00006250	0.22322168
[0.05, 0.95]	0.89747209	0.09747209	0.54203312

demonstrate the converse regime without deleting the unfavorable broad-support case. The two-level equality in the proposition is checked separately by direct algebra rather than estimating a probability of agreement from stochastic replications. Figure 3 shows the continuous ordering distributions used in the population experiments.

D FINITE INNER SAMPLING AND OBSERVABLE HETEROGENEITY

A histogram of per-ordering success fractions mixes two sources of variation. Let $Q_1, \dots, Q_K$ be iid ordering probabilities, and, conditional on them, let $S_j \sim \mathrm{Bin}(n, Q_j)$ independently. Write $\widehat{q}_j = S_j/n$, $m = \mathbb{E}Q$, and $s^2 = \mathrm{Var}(Q)$. The law of total variance gives

$$\mathrm{Var}(\widehat{q}_j) = s^2 + \frac{\mathbb{E}[Q(1-Q)]}{n} = s^2 + \frac{m(1-m) - s^2}{n}. \tag{28}$$

The second term is generation noise, not ordering heterogeneity. Increasing K estimates this convolution more accurately without removing the second term. Binomial-mixture recovery is a distinct statistical problem (Tian et al., 2017; Vinayak et al., 2019).

For $K, n \geq 2$, the latent-variance estimator in Eq. (16) is unbiased. Indeed, the first term has expectation $\mathrm{Var}(\widehat{q}_j)$, while the expectation of the second is $\mathbb{E}[Q(1-Q)]/n$. The estimate can be negative at finite sample sizes. Truncating it to zero is acceptable for a display, but changes its expectation and does not create an upper confidence bound. Uncertainty must retain the ordering as a cluster; the Kn outputs are not independent draws of the latent Q.

Nonlinear estimation has an additional bias. For the clipped log score in Eq. (6), define

$$h_{n,\delta}(q) = \sum_{s=0}^{n} \binom{n}{s} q^s (1-q)^{n-s} u(s/n). \tag{29}$$

The plug-in ordering average $K^{-1}\sum_j u(\widehat{q}_j)$ converges to $\mathbb{E}h_{n,\delta}(Q)$ as $K \to \infty$ at fixed n, not generally to $\mathbb{E}u(Q)$. This is an instance of the nonlinear nested-estimation issue studied by Rainforth et al. (2018). Clipping prevents a blanket Jensen argument for the sign of the bias on the entire parameter space; Table D.1 evaluates the finite sum rather than assuming a sign.

Proof of Theorem 3. The count probabilities are expectations of the Bernstein basis polynomials $\binom{n}{s}Q^s(1-Q)^{n-s}$, $s = 0, \dots, n$. This basis spans all polynomials of degree at most n, so the count

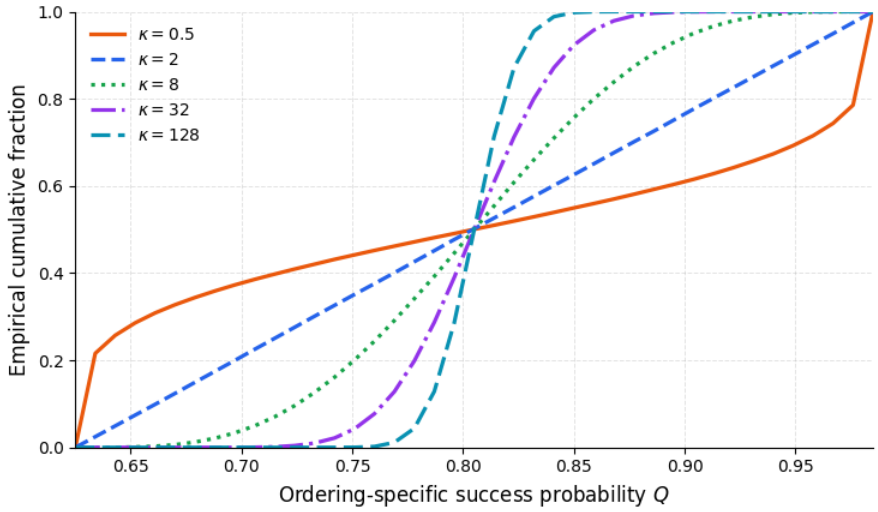

Figure 3: Cumulative ordering distributions used in the continuous-population experiments of Table 3. Each curve is the empirical CDF obtained from 256,000 beta draws per row, aggregated into 40 bins. All five distributions have the same mean ordering-specific success probability, $\mathbb{E}[Q] = .805$, while varying substantially in concentration through $\kappa \in \{.5, 2, 8, 32, 128\}$. No distribution contains a population endpoint atom. The figure therefore isolates the effect of distributional concentration under a fixed mean, providing the controlled population model used in the subsequent analysis.

law determines exactly those moments. To prove the converse directly, interpolate ϕ at any $n+1$ distinct points by a degree-n polynomial. If ϕ is not such a polynomial, an additional point differs from that interpolant. The matrix with rows $1, q, \ldots, q^n, \phi(q)$ on these $n+2$ points is nonsingular. A nonzero vector annihilating the first $n+1$ rows therefore has nonzero inner product with the final row. Its entries sum to zero. Normalizing its positive and negative parts gives two probability laws whose first n moments agree but whose ϕ expectations differ. They produce the same count law and hence the same distribution of every finite or infinite iid count history. Consistency under both laws would require incompatible limits under identical observations. □

The moment limitation is the mechanism behind the binomial-mixture lower bounds of Tian et al. (2017). It does not rule out useful approximate recovery, parametric models, increasing n, or partially identified intervals. It also does not establish that every fixed-depth study misclassifies a given law. It specifies which functional cannot be identified uniformly without additional information.

D.1 Exact finite-depth calculation

We evaluated Eq. (29) for each continuous law in Eq. (21) and the homogeneous control $Q = .805$. The target was .8, clipping was .01, and inner depths were $n \in \{5, 10, 20, 50, 100, 200\}$. Each beta expectation used normalized 256-node Gauss–Jacobi quadrature. The conditional binomial probabilities were summed over every $s = 0, \ldots, n$. The homogeneous control requires only that finite binomial sum. Table D.1 reports $n = 20$; the numbers are analytical expectations, not estimates from generator samples.

For the homogeneous control, expected plug-in losses at depths $10, 20, 50, 100, 200$ are respectively .231386, .223381, .219379, .218135, and .217521. Their limit is $-\log .805 = .216913$. At depth 20, the observed fraction has variance .00784875 despite zero latent variance. This control separates an artifact of estimating probabilities from the genuine Jensen gap in the continuous-ordering study.

Table D.1: Finite-depth measurement effects at 20 verifier draws per ordering, mean success .805, and clipped-log cutoff $-\log .8 = .223144$. Expected plug-in is $\mathbb{E}h_{20,.01}(Q)$. Noise share is the fraction of $\mathrm{Var}(\widehat{q})$ contributed by $\mathbb{E}[Q(1-Q)]/20$. The first two laws fail the population log criterion; the remaining four satisfy it, although all six expected plug-in scores exceed the cutoff. Increasing the number of orderings alone does not remove these biases.

Ordering law	True residual	Expected plug-in	Latent variance	Noise share
$\kappa = .5$	.233947	.243760	.021600	23.86%
$\kappa = 2$	.225374	.233280	.010800	40.36%
$\kappa = 8$	.219710	.226569	.003600	68.05%
$\kappa = 32$	.217672	.224235	.000982	88.82%
$\kappa = 128$	.217107	.223598	.000251	96.89%
Homogeneous	.216913	.223381	0	100%

D.2 A FINITE ORDERING AND GENERATION BUDGET

We sampled 2000 independent datasets per law, each comprising 200 fresh orderings and 20 conditionally independent bits per ordering, for a total of 4000 simulated verifier calls per dataset. The six laws are ordered as the five concentrations above followed by the homogeneous control. Seeds are $202609231 + i$ for zero-based law index i. Each computational batch contains 50 datasets: a 50×200 array of latent probabilities is generated first, followed by a $50 \times 200 \times 20$ array of independent uniforms, thereby fixing the simulation draw order.

To use all bits without overstating independence, we pair the 200 orderings into 100 disjoint pairs. For each pair, we average the ten four-bit A statistics and ten four-bit W statistics formed from consecutive disjoint inner draws. Each pair-average preserves expectations m and s^2 and has range length one. Thus, there are 100 independent blocks, not 1000: inner groups sharing an ordering also share its latent probability. At local error .05, the resulting range-based audit has radius .148021. We apply Proposition 2 with the known simulation support $[.625, .985]$ to each dataset, charging all 4000 calls, including unresolved cases.

Table D.2: Controlled observation-budget study with 200 orderings and 20 verifier bits per ordering in every independent dataset. A raw flag is $\overline{q} > .8$ together with the plug-in mean loss at least $-\log .8$; it is not a certificate. Audit decisions count either certified mismatch or a certified log criterion using the block-valid radius. Every dataset is unresolved by that conservative audit. The final column retains negative unbiased variance estimates rather than truncating away their sampling variation. All counts are across independent simulated datasets, not independent model tasks.

Ordering law	Raw mismatch flags	Audit decisions	Negative $\widehat{s}^2_{\mathrm{unb}}$
$\kappa = .5$	1107/2000	0/2000	0/2000
$\kappa = 2$	984/2000	0/2000	0/2000
$\kappa = 8$	748/2000	0/2000	1/2000
$\kappa = 32$	639/2000	0/2000	260/2000
$\kappa = 128$	587/2000	0/2000	779/2000
Homogeneous	581/2000	0/2000	1006/2000

Raw flags persist even in the homogeneous control, where neither population mismatch nor ordering effects are present. Conversely, an unresolved moment audit does not establish that a task lies outside the mismatch band. These observations distinguish three outputs that should not be conflated: a noisy histogram, a model-based estimate of the latent distribution, and a finite-sample certificate. Applying the analysis to generated programs additionally requires the sampled ordering identities, complete verifier outcomes, and a verifier whose tests are excluded from the generator's context.

Table D.3: Exact finite-binomial expectations for the homogeneous law $Q = .805$ and clipping $\delta = .01$. Leading bias is $(1 - m)/(2mn)$; raw and jackknife expectations use Eqs. (29) and (31). No generated-program data enter this calculation. The correction is not exact at finite depth and can have negative bias.

Depth n	Leading bias	Raw expectation	Jackknife expectation	True residual
10	0.01211180	0.23138571	0.21394889	.21691300
20	0.00605590	0.22338098	0.21612263	.21691300
50	0.00242236	0.21937871	0.21686567	.21691300
100	0.00121118	0.21813483	0.21690210	.21691300
200	0.00060559	0.21752124	0.21691033	.21691300

D.3 Leading bias and count-based correction

Fix $\delta > 0$ and $\delta < a \le q \le b < 1 - \delta$. Put $\Delta = S/n - q$. On a fixed neighborhood of this interval, $u(x) = -\log x$ is smooth. Taylor expansion and binomial central moments give

$$h_{n,\delta}(q) = -\log q + \frac{1-q}{2nq} + \frac{(1-q)(5-q)}{12n^2q^2} + O(n^{-3}), \tag{30}$$

uniformly for $q \in [a, b]$, with constants depending on the interval and clipping level. In detail, $\mathbb{E}\Delta = 0$, $\mathbb{E}\Delta^2 = q(1-q)/n$, $\mathbb{E}\Delta^3 = q(1-q)(1-2q)/n^2$, and $\mathbb{E}\Delta^4 = 3q^2(1-q)^2/n^2 + O(n^{-3})$. Combining the cubic and quartic Taylor terms yields $(1-q)(5-q)/(12q^2)$. Fifth and sixth moments contribute $O(n^{-3})$ after expansion through order five with a sixth-order remainder. Outside the fixed smooth neighborhood, Hoeffding's inequality gives exponentially small probability and the clipped score is bounded. This proves Eq. (30) and its first-order main-text specialization. The constants do not remain bounded as the support approaches a clipping boundary, so the expansion is not a global bias guarantee.

For homogeneous $m = .805$ at $n = 20$, the leading term is .00605590 nats, while the exact bias is .00646798. The difference includes higher-order terms and upper-clipping events. At $m = .8$ and $n = 100$, the leading term is .00125. For a heterogeneous law in the same fixed interior interval, averaging Eq. (17) yields $\mathbb{E}[(1 - Q)/Q]/(2n) + O(n^{-2})$. Jensen gives $\mathbb{E}[1/Q] \ge 1/m$, explaining why a calculation at the pooled mean alone can understate that leading contribution. A leading-order depth criterion for a homogeneous log margin $g = \log(m/p^\star) > 0$ is $n \gtrsim (1 - m)/(2mrg)$ to keep the bias below fraction r of the margin. This is a design approximation, not a finite-sample sufficiency guarantee.

The count-based jackknife in Eq. (18) is the binomial delete-one jackknife. It uses no samples beyond the original ordering counts. A term with zero coefficient is omitted: at $s = 0$ only deleting a failure is possible, and at $s = n$ only deleting a success is possible. Therefore u is never evaluated outside $[0, 1]$. Averaging over which of the n bits is deleted shows exactly that

$$\mathbb{E}[j_{n,\delta}(S) \mid Q = q] = nh_{n,\delta}(q) - (n-1)h_{n-1,\delta}(q). \tag{31}$$

The n^{-1} term cancels, leaving $O(n^{-2})$ bias on a fixed interior interval. This is the classical jackknife mechanism, whose nonuniform behavior in the binomial model is analyzed by Jiao & Han (2020). It does not contradict Theorem 3: for any fixed n, the expectation of the corrected count statistic is still a polynomial in q of degree at most n, rather than the nonpolynomial clipped log function. A bias-corrected point estimate is not automatically a one-sided valid bound. Table D.3 reports the finite-depth bias and jackknife correction behavior. Table D.4 reports the exact finite-sample expectations and jackknife biases.

D.4 Deeper inner sampling with ordering-clustered inference

The second count-level experiment used $K = 50$ iid orderings, $n = 200$ independent verifier draws per ordering, and 2000 independent datasets per condition. The five beta laws of Eq. (21) and the homogeneous control were unchanged. NumPy generators used seeds $314159265 + i$ for zero-based condition index i. For each condition, a complete 2000×50 array of latent beta probabilities was drawn first, followed by a binomial count array of the same shape. In the homogeneous condition,

Table D.4: Exact binomial-mixture expectations at $n = 200$, $m = .805$, and $\delta = .01$. Beta expectations use the same normalized 256-node Gauss–Jacobi rule as the continuous study. The least concentrated laws retain appreciable clipping effects at this depth; the first-order interior approximation is not substituted for their finite-sample calculation.

Ordering law	True residual	Raw expectation	Jackknife expectation	Jackknife bias
$\kappa = 0.5$	0.23394736	0.23488250	0.23370950	−0.00023786
$\kappa = 2$	0.22537407	0.22605647	0.22533340	−0.00004067
$\kappa = 8$	0.21970986	0.22033583	0.21970674	−0.00000312
$\kappa = 32$	0.21767218	0.21828519	0.21766945	−0.00000273
$\kappa = 128$	0.21710690	0.21771636	0.21710421	−0.00000269
Homogeneous	0.21691300	0.21752124	0.21691033	−0.00000268

Table D.5: Count-level simulation with 50 orderings and 200 verifier bits per ordering. Coverage is the fraction of 2000 independent datasets whose ordering-clustered percentile interval contains the true variance or true residual. All intervals are marginal and approximate. The columns are not simultaneous task-wise certification results. Each row uses a specified mathematical law, not an LLM task population.

Law	Raw flags	Corrected flags	Mean $\widehat{s}^2_{\mathrm{unb}}$	Var. coverage	Log coverage
$\kappa = 0.5$	503/2000	465/2000	0.021532	0.9435	0.9395
$\kappa = 2$	389/2000	359/2000	0.010793	0.9255	0.9375
$\kappa = 8$	216/2000	176/2000	0.003590	0.9245	0.9390
$\kappa = 32$	121/2000	61/2000	0.000966	0.9185	0.9370
$\kappa = 128$	68/2000	13/2000	0.000247	0.9100	0.9345
Hom.	46/2000	1/2000	-2.71×10^{-7}	0.9095	0.9455

the latent array was constant and consumed no random draws. Sampling a binomial count simulates the exact sufficient statistic of 200 raw bits; each dataset is nevertheless charged 10,000 verifier calls.

For every dataset, we computed $\overline{q}$, Eq. (16), the raw residual mean, and Eq. (18) from the same counts. The raw and corrected flags require $\overline{q} > .8$ and their respective residual point estimate at least $-\log .8$. These flags are descriptive classifications; neither is a certificate. The homogeneous control has zero true ordering variance and no target mismatch.

We formed nominal 95% percentile intervals by resampling the 50 complete ordering counts with replacement, using 2000 bootstrap draws per dataset. A 2000×50 bootstrap-index matrix was drawn once using seed 271828182 and held fixed across conditions and datasets. Each resample recomputed the mean, the unbiased-variance formula, and the jackknife mean. Inner bits were not independently resampled across orderings. Empirical interval coverage was then checked against the known mathematical population value on all 2000 independently generated datasets. Table D.5 shows the coverage shortfall rather than interpreting nominal bootstrap levels as a theorem.

The true variance is positive in every beta row. Negative unbiased estimates occurred in $0, 0, 0, 0, 212$ datasets respectively, and in 1057 homogeneous datasets. In the homogeneous control the sample standard deviation of this estimate was 1.5951×10^{-4}; negative values were preserved. The corrected residual standard deviations across the six conditions were $.027689, .019496, .011633, .007409, .005719, .004894$ nats. These fluctuations are dominated by the finite number of independent ordering clusters in the heterogeneous rows, even though the inner depth is ten times that of Table D.2. Bias correction addresses the inner nonlinear transformation; it cannot identify an unobserved hard support or eliminate outer sampling uncertainty.

For any fixed target p, an interval-based descriptive classification can require $m_L > p$ and $\mu_L \geq c^\star_u(p)$ for a separated mismatch, $\mu_U < c^\star_u(p)$ for the log criterion, and otherwise return unresolved. A finite-sample certificate requires jointly valid intervals and explicit error allocation, unlike the marginal percentile intervals just evaluated. At the narrow .805 versus .8 margin, no dataset in this experiment had both the bootstrap lower mean above .8 and the bootstrap lower residual above the cutoff. Small correction bias alone is therefore insufficient to establish mismatch from a 50-

ordering pilot. The same definitions apply at targets $.5, .7, .9$ without changing the estimand or silently treating a failed individual output as a false certificate.

E REPLACING HARD SUPPORT BY CONTROLLED TAIL MASS

The support restriction in Proposition 1 is not established by observed extrema. For clipped loss, a weaker requirement is available: account explicitly for the probability mass outside a declared central interval. Choose $\delta \le a < p^\star < b \le 1-\delta$, where $p^\star$ is the success target, and let $S_{a,b}(q)$ be the affine secant in Eq. (24), extended to all of $[0,1]$. Define

$$D_a = \max\{0, B_\delta - S_{a,b}(\delta)\}, \qquad D_b = \max\{0, a_\delta - S_{a,b}(1)\}. \tag{32}$$

For any ordering distribution on $[0,1]$, not necessarily supported on $[a,b]$, write $\tau_a = \mathbb{P}(Q < a)$ and $\tau_b = \mathbb{P}(Q > b)$. Then

$$\mathbb{E}u(Q) \le S_{a,b}(m) + D_a\tau_a + D_b\tau_b. \tag{33}$$

On $[a,b]$, the secant majorizes u. On $[\delta,a]$, convexity makes the maximum of $u - S_{a,b}$ occur at an endpoint; on $[0,\delta]$, that difference is increasing and attains its maximum at δ. This gives D_a. The same endpoint argument on $[b,1-\delta]$, followed by the constant upper-clipping segment, gives D_b. Averaging the resulting pointwise inequality proves Eq. (33). Zero tail mass recovers the interval-support result exactly.

Tail mass can itself be bounded from finite verifier counts. Let K independent orderings each receive n bits, with counts S_j. Fix integer detection thresholds r_a, r_b before observing these counts, and set

$$d_a = \mathbb{P}\{\mathrm{Bin}(n,a) \le r_a\} > 0, \qquad d_b = \mathbb{P}\{\mathrm{Bin}(n,b) \ge r_b\} > 0. \tag{34}$$

Binomial stochastic monotonicity implies

$$\mathbb{P}(S_j \le r_a) \ge d_a\tau_a, \qquad \mathbb{P}(S_j \ge r_b) \ge d_b\tau_b.$$

If U_a, U_b are one-sided binomial upper confidence bounds for these two observable detection probabilities, then $\tau_a \le \min(1, U_a/d_a)$ and $\tau_b \le \min(1, U_b/d_b)$ on their joint coverage event. Exact binomial inversion (Clopper & Pearson, 1934) is one option. The two counts may be dependent within an ordering; a union bound, rather than an independence assumption, combines their error allocations.

Obtain a lower confidence bound m_L from the first verifier bit of each ordering, whose marginal law is iid $\mathrm{Ber}(m)$. Allocate errors $\alpha_m + \alpha_a + \alpha_b \le \alpha$. Because the secant is decreasing, the entirely observable bound

$$U_\mu = \min\left\{B_\delta,\ S_{a,b}(m_L) + D_a \min(1, U_a/d_a) + D_b \min(1, U_b/d_b)\right\} \tag{35}$$

contains $\mathbb{E}u(Q)$ from above with probability at least $1-\alpha$. Thus $U_\mu < c_u^\star(p^\star)$ is a valid certificate without asserting hard population support. The bounds may share the same data under the stated error split. An empty or negative upper bound can occur only outside their simultaneous coverage event and is handled as an unresolved audit, not an automatic acceptance.

This construction trades an uncheckable hard-support claim for statistical uncertainty about rare tails. It need not be informative at small budgets. False detections from central probabilities can make U_a/d_a or U_b/d_b large, and lower clipping can make D_a substantial. The interval and thresholds must be fixed using independent calibration data or protected simultaneously over a declared finite grid. Optimizing them retrospectively without that correction would invalidate the certificate. The construction provides a validity route, not a measured low-cost deployment result.

F IDENTIFIABILITY FROM CONFIDENCE SUMMARIES

Proposition 3 (Indistinguishable score laws)**.** *Suppose the complete observable score history has the same distribution under two generator laws P_- and P_+, with verifier-success probabilities satisfying $\bar{q}_- \le p^\star < \bar{q}_+$. Any randomized acceptance rule based only on that history and independent private randomness has identical acceptance probabilities under the two laws. If it is α-valid under P_-, its power under P_+ is at most α, including at an unbounded stopping time.*

Proof. The acceptance event is measurable with respect to the score history and the private random seed. Both have the same joint law under the two hypotheses, so the event has equal probability under them. Uniform validity at the lower-success law then gives the bound at the higher-success law. □

For a concrete construction with a fixed verifier, let the output alphabet be $\{y_0, y_1\}$, with $g(y_0) = 0$ and $g(y_1) = 1$. Under P_+, set $P_+(y_1) = q$; under P_-, set $P_-(y_1) = 1 - q$, where $q > \max\{p^\star, 1 - p^\star\}$. Entropy and the absolute top-two margin are identical under the two laws. The distribution of the log probability assigned to a sampled output is also identical: it is $\log q$ with probability q and $\log(1 - q)$ with probability $1 - q$ under either law. Hence, entire iid histories of these statistics have the same distribution under both laws, satisfying the proposition. Entropy of empirical frequencies and entropy over semantic classes without correctness labels are likewise invariant to exchanging the two labels.

The restriction to the observation sigma-field is essential. A procedure that observes a verifier bit sees different Bernoulli laws, while one that observes informative text together with a correctness-linked model need not satisfy the premise. A semantic-entropy system also need not claim a universal verifier-success certificate. Its demonstrated ability to detect confabulations, together with its limitation for consistent errors (Farquhar et al., 2024), is compatible with this proposition. The implication is that a score-only wrapper requires an additional identification or calibration assumption before supporting a stronger certificate claim.

G Direct certification and future-output risk

Proposition 4 (Direct Bernoulli evaluation)**.** *Under the frozen-round sampling law of Section 2, the process in Eq.* (2) *is a nonnegative supermartingale under the conditional null $\bar{q}_t \le p^\star$. With error allocation $\alpha_d w_t$, the probability of accepting any round in this null is at most α_d.*

Proof. Marginalizing each fresh ordering gives $\mathbb{P}(X_{t,j} = 1 \mid \mathcal{H}_t) = \bar{q}_t$. Independent ordering and generation draws yield conditional iid Bernoulli outcomes. For a fixed $\theta \in (p^\star, 1]$ and $q \le p^\star$, the expected likelihood-ratio factor is

$$q\frac{\theta}{p^\star} + (1 - q)\frac{1 - \theta}{1 - p^\star} = 1 + \frac{(q - p^\star)(\theta - p^\star)}{p^\star(1 - p^\star)} \le 1. \tag{36}$$

The factors are nonnegative. Integrating over a probability measure fixed before the draws preserves the supermartingale property by Tonelli's theorem. Ville's inequality bounds the conditional probability of a crossing at any inner time by $\alpha_d w_t$. If $\bar{q}_t$ is random through acquisition, apply this bound only on the $\mathcal{H}_t$-measurable event $\{\bar{q}_t \le p^\star\}$, take expectations, and sum over t. Since $\sum_{t \ge 0} w_t = 1$, the result follows. No independence between acquisition rounds is needed. □

A predictable alternative selected separately before each outcome also preserves Eq. (36). A retrospectively selected likelihood ratio does not generally preserve validity. On changing the evidence, the trial distribution changes; a new round with its own error allocation is necessary for the claim about the newly frozen state. The endpoint $\theta = 1$ is defined to have factor $1/p^\star$ on success and zero on failure. It is an exact boundary benchmark, not a robust alternative under occasional failures.

For the uniform prior on $[p^\star, 1]$, after n successes and no failures,

$$T_n = \frac{1}{1 - p^\star}\int_{p^\star}^{1}\left(\frac{\theta}{p^\star}\right)^n d\theta = \frac{1 - (p^\star)^{n+1}}{(n + 1)(1 - p^\star)(p^\star)^n}. \tag{37}$$

At $p^\star = .95$ and $1/\alpha = 40$, its first crossing is $n = 105$, compared with 72 for the endpoint alternative. These calculations do not establish that an unknown generator will produce an all-success pilot.

To prove Eq. (3), let B be the event that an accepted state's success probability is at most $p^\star$. Its probability is at most α_d. Conditional on the accepted state and the full evaluation history, a fresh independently generated output fails with probability $1 - \bar{q}$. Therefore

$$\mathbb{P}(A, \text{failure}) = \mathbb{E}[\mathbf{1}_A(1 - \bar{q})] \le (1 - p^\star)\mathbb{P}(A \setminus B) + \mathbb{P}(B) \le (1 - p^\star)\mathbb{P}(A) + \alpha_d.$$

An already selected pilot output need not have this conditional distribution. Nor does the claim survive an undeclared change to the verifier, sampling temperature, ordering law, or context.

If several acceptance branches are combined, their allocated error probabilities add. Replacing α_d by α_{tot}, the sum of the branchwise budgets, yields the same joint-risk bound. When $\mathbb{P}(A) > 0$, division by $\mathbb{P}(A)$ gives conditional risk at most

$$1 - p^\star + \frac{\alpha_{\mathrm{tot}}}{\mathbb{P}(A)}.$$

H VERIFIER-BIT COMPLEXITY AND FINITE-WIDTH EVALUATION

Exact ordering probabilities are generally unavailable. If each fresh ordering supplies only one verifier bit, the observations are iid $\mathrm{Ber}(\mathbb{E}Q)$. Thus, two ordering distributions with the same mean can yield identical observations but different nonlinear residuals. Repeated evaluation within an ordering is therefore needed to identify that residual, but is unnecessary for direct success certification. Verifier bits are also a garbling of exact Q: an exact-probability oracle can simulate every subsequent bit. The corresponding oracle lower bounds therefore transfer to black-box procedures without correctness-linked side information (Appendix I).

For an implementable residual, define $r_\delta(q) = \delta + (1 - 2\delta)q$ and $\mu^{\mathrm{s}} = \mathbb{E}[-\log r_\delta(Q)]$. Independently flipping each verifier bit with probability δ yields success probability $r_\delta(q)$. This smoothing differs from clipping, and its largest safe cutoff is $c^\star_{\mathrm{s}} = -\log r_\delta(p^\star)$ by convexity. The allocation below estimates the functional itself and applies with either this cutoff or a stricter one.

For one ordering, let K be the first successful smoothed trial. With cap L, observe $\min(K, L)$ trials and return

$$Z_L = \sum_{k=1}^{L} \frac{\mathbf{1}\{K > k\}}{k}, \quad H_L = \sum_{k=1}^{L} \frac{1}{k}, \quad b_L = \frac{(1-\delta)^{L+1}}{(L+1)\delta}. \tag{38}$$

Inverse-binomial sampling gives $\mathbb{E}Z_\infty = -\log r_\delta(q)$ (van Opheusden et al., 2020). Truncation makes $Z_L \in [0, H_L]$ and leaves bias in $[0, b_L]$. At local error α_t, averaging m independent ordering-level scores gives

$$[L^{\mathrm{s}}_t, U^{\mathrm{s}}_t] = [\bar{Z}_m - \rho_m, \bar{Z}_m + \rho_m + b_L], \qquad \rho_m = H_L \sqrt{\frac{\log(2/\alpha_t)}{2m}}. \tag{39}$$

Theorem 4 (Finite width and unavoidable verifier cost)**.** *Choose the smallest L with $b_L \le w/2$ and*

$$m = \left\lceil \frac{8H_L^2 \log(2/\alpha_t)}{w^2} \right\rceil. \tag{40}$$

The interval in Eq. (39) *has conditional coverage at least $1 - \alpha_t$ and width at most w on every realization. Its verifier cost obeys $C_t \le mL$ and $\mathbb{E}C_t \le m/\delta$. Conversely, for $0 < \delta, w \le 1/8$ and $0 < \alpha < 1/4$, every almost surely terminating, bits-only procedure returning an interval of width at most w with coverage at least $1 - \alpha$ uniformly over ordering distributions satisfies*

$$\sup_P \mathbb{E}_P C \;\ge\; \frac{d(1-\alpha, \alpha)}{128\delta w^2} \tag{41}$$

for estimating μ^{s}. The lower bound permits adaptive ordering selection, revisits, and stopping under the bits-only observation model.

The converse lower-bounds width-w residual estimation, not certification at a fixed success target; the latter inherits the oracle-testing lower bound through garbling, which need not be tight in verifier calls. The constructive result combines the harmonic tail bound with concentration across independent orderings. For the converse, homogeneous probabilities $q_0 = \delta$ and $q_1 = \delta(1 + 8w)$ have smoothed residuals separated by more than $2w$, while each raw bit contributes at most $128\delta w^2$ KL divergence. A valid narrow interval must distinguish these two laws, so a stopped change-of-measure argument lower-bounds the required number of calls. Appendix J proves the allocation; Appendix I proves the converse, including adaptive sampling.

Since $L = O(\delta^{-1}\log(1/(\delta w)))$, the upper bound is $O(\delta^{-1}w^{-2}\log^2 L\log(1/\alpha_t))$. Thus the exponents δ^{-1} and w^{-2} are necessary in the worst case, while logarithmic factors and constants remain unresolved. This is not a claim of instance-optimal allocation. Choosing $\alpha_t = \alpha/(H+1)$ covers all $H+1$ adaptive acquisition rounds. On that event, $\mu_t^{\mathrm{s}} \le U_t^{\mathrm{s}} \le \mu_t^{\mathrm{s}} + w$ brackets an implemented crossing between the population cutoffs $c_{\mathrm{s}}^\star$ and $c_{\mathrm{s}}^\star - w$; validity itself requires no positive drift.

H.1 A VERIFIER-ONLY HETEROGENEITY AUDIT

We prove Proposition 2 and record the complete simulation protocol underlying Table 4. The audit estimates two moments rather than the full ordering law or the residual to arbitrary precision; consequently, a sufficient decision need not resolve every mismatch.

H.2 PROOF OF THE FOUR-BIT AUDIT

Proof of Proposition 2. Let Q, Q' be independent ordering probabilities in one block. Conditional on them, the four bits are independent, with the first pair having mean Q and the second pair mean Q'. Therefore $\mathbb{E}A = m$, and

$$\mathbb{E}[W \mid Q, Q'] = \tfrac{1}{2}(Q-Q')^2, \qquad \mathbb{E}W = \tfrac{1}{2}\mathbb{E}(Q-Q')^2 = \mathrm{Var}(Q).$$

The variable A lies in $[0,1]$, while W lies in $[-1/2, 1/2]$. Each sample mean thus has deviation probability at most $2e^{-2Kr_K^2} = \alpha/2$. A union bound proves joint coverage, even though the two statistics within a block are dependent. The known support implies $m \in [a,b]$ and $s^2 \le (b-a)^2/4$, so intersecting with these ranges preserves coverage. In implementation both interval endpoints can instead be projected to the respective known range, which avoids an empty interval outside the coverage event and leaves the guarantee unchanged.

On joint coverage, $-\log m$ lies between $-\log m_U$ and $-\log m_L$. Applying the curvature bounds to $s^2 \in [v_L, v_U]$ proves Eq. (20). Both decisions are deterministic implications on this same event. They therefore share the one error allocation α, rather than spending a new budget for each inequality. Across a finite collection of states, summing their local error allocations gives simultaneous validity; arbitrary inspection over block counts would require additional spending or time-uniform intervals. □

Knowing m and s^2 does not in general identify $\mathbb{E}[-\log Q]$. The bounds can remain on opposite sides of the cutoff even with infinite data. An unresolved audit is consequently neither evidence of alignment nor proof of mismatch. The audit is useful precisely when the moment bounds separate the decision, and it still requires the population support assumption; its sample extrema are not substituted for that assumption.

H.3 FOUR-BIT AUDIT PROTOCOL AND RESULTS

For each of the five continuous laws in Table 4, we drew $K = 10^6$ independent blocks of two orderings and two verifier bits per ordering. The audit saw only these four bits; latent probabilities were used to generate the mathematical environment but were not passed to its decision rule. Local error was .01, so the union bound gives total error .05 over both moment intervals and both possible decisions across all five laws. The resulting common radius was $\sqrt{\log 400/(2\times 10^6)}$. The support $[.625, .985]$ was fixed by the environment definition, not estimated from the observed bits.

We used seeds $20261031 + i$ for zero-based concentration index i and blocks of 10000 for memory-limited execution. Within each block, we drew a 10000×2 array of latent beta probabilities and a $10000 \times 2 \times 2$ array of independent uniforms for the verifier bits. We accumulated the sums of the block average A and variance statistic W and applied the stated interval formulas. All 4×10^6 verifier calls per row are counted.

I BITS-ONLY OBSERVATION AND LOWER BOUNDS

The bits-only experiment permits a procedure to request an iid new ordering, retain its identity, obtain any number of independent verifier bits from that ordering, revisit an earlier ordering, and stop

adaptively. An ordering's latent probability is drawn from an unknown law P; ordering identities and private randomization carry no additional information about P. Each raw verifier draw counts as one call. The model excludes oracles that reveal correctness from an ordering identity or provide an exhaustive table of latent probabilities.

Proposition 5 (One bit per ordering does not identify nonlinear means)**.** *If only one bit is observed from each fresh ordering, the complete sequence is iid* $\mathrm{Ber}(\mathbb{E}_P Q)$*. For any nonaffine continuous* ϕ*, there are ordering laws with the same mean and different values of* $\mathbb{E}_P \phi(Q)$*. No estimator based on this observation scheme can consistently recover that functional uniformly over ordering laws.*

Proof. Marginalizing a fresh ordering gives a Bernoulli bit with success probability $\mathbb{E}_P Q$, independently across draws. If every pair of equal-mean laws also had equal ϕ expectations, each two-point mixture on x, y would have the same expectation as its mean point mass. Thus $\phi(\lambda x + (1-\lambda)y) = \lambda\phi(x) + (1-\lambda)\phi(y)$ for every $\lambda \in [0,1]$, forcing ϕ to be affine. The contrapositive gives distinct functionals under identical observation laws, which rules out uniform consistency. □

For the smoothed residual, a homogeneous law $Q = m$ and the endpoint mixture with $\mathbb{P}(Q = 1) = m$, $0 < m < 1$, already give a strict difference by convexity. Repeated bits within an ordering can distinguish them. Those repetitions are unnecessary if the target is just $\mathbb{E}_P Q$.

Converse part of Theorem 4. Restrict to the homogeneous subfamily $Q = q$ almost surely. Under this subfamily, every raw verifier bit is iid $\mathrm{Ber}(q)$, regardless of how the procedure selects or revisits orderings. Set

$$q_0 = \delta, \qquad q_1 = \delta(1 + 8w), \qquad r_j = \delta + (1 - 2\delta)q_j.$$

For $0 < \delta, w \le 1/8$, both probabilities lie in $(0, 1/4]$. The residual separation is

$$\log \frac{r_1}{r_0} = \log\left(1 + \frac{4(1-2\delta)}{1-\delta} w\right) > 2w. \tag{42}$$

The coefficient of w is at least $24/7 > 3$, and $\log(1 + 3w) \ge 3w/(1 + 3w) > 2w$ for $w \le 1/8$, establishing the inequality.

Let I be the reported interval and C its number of calls. Under q_0, the event $A = \{-\log r_0 \in I\}$ has probability at least $1 - \alpha$. Under q_1, the interval cannot contain both separated means, so coverage at q_1 implies $\mathbb{P}_{q_1}(A) \le \alpha$. Data processing therefore lower-bounds the KL divergence between the stopped transcripts by $d(1-\alpha, \alpha)$. If $\mathbb{E}_{q_0} C = \infty$, the result is immediate. Otherwise the stopped chain rule gives

$$\mathbb{E}_{q_0} C \, d(q_0, q_1) \ge d(1 - \alpha, \alpha). \tag{43}$$

Every action and private randomization has the same conditional kernel under the two hypotheses, so only a newly revealed raw bit contributes KL. This is why adaptive revisits, randomized smoothing, and data-dependent stopping do not alter the information accounting.

For Bernoulli laws, the elementary bound $d(a, b) \le (a-b)^2/[b(1-b)]$ follows from $\log x \le x - 1$. Here $q_1 \ge \delta$ and $1 - q_1 \ge 3/4$, so

$$d(q_0, q_1) \le \frac{64\delta^2 w^2}{(3/4)\delta} = \frac{256}{3}\delta w^2 \le 128\delta w^2.$$

Substitution in Eq. (43) proves the claimed minimax lower bound. The stronger first expression in Eq. (43), rather than its relaxed constant, is used in the numerical lower-bound curves. □

The lower and harmonic upper bounds match in their powers of δ and w, but not in their logarithmic factors or constants. In particular, the homogeneous two-point witness does not establish instance optimality of the rectangular nested design or the displayed harmonic allocation. The numerical study therefore reports the remaining gap explicitly.

I.1 TRANSFER OF EXACT-PROBABILITY LOWER BOUNDS

An exact-probability observer can simulate any bits-only procedure: when the procedure requests a new ordering, draw its latent Q, retain it, and produce all requested bits using independent Bernoulli draws with that probability. Private randomization, smoothing, adaptive revisits, and stopping can be simulated without extra information about P. If N distinct orderings are evaluated, then $N \le C$ whenever each evaluated ordering receives at least one paid call. The simulated procedure has exactly the original acceptance law. Therefore a lower bound on required exact-Q observations also lower-bounds its verifier-call cost. The expectation and all-success quantile bounds remain applicable after such garbling. This argument does not identify one exact probability with the cost of one model call; it uses the more informative oracle only for a lower bound.

J PROOF AND COST OF THE HARMONIC ALLOCATION

The smoothed residual is $\mu_t^{\mathrm{s}} = \mathbb{E}_\pi[-\log r_\delta(q_{t,\pi})]$, with $r_\delta(q) = \delta + (1-2\delta)q$. Write $\mathcal{I}_t^{\mathrm{s}}$ for the interval in Eq. (39), optionally intersected with $[a_\delta, B_\delta]$. Its lower and upper endpoints are $L_t^{\mathrm{s}}, U_t^{\mathrm{s}}$.

Constructive part of Theorem 4. Conditional on an ordering, the smoothed Bernoulli probability is $r \in [\delta, 1-\delta]$. The first-success time satisfies $\mathbb{P}(K > k \mid r) = (1-r)^k$. Hence

$$\mathbb{E}[Z_L \mid r] = \sum_{k=1}^{L} \frac{(1-r)^k}{k}.$$

Using $-\log r = \sum_{k\ge1}(1-r)^k/k$ and summing a geometric upper bound,

$$0 \le -\log r - \mathbb{E}[Z_L \mid r] = \sum_{k>L} \frac{(1-r)^k}{k} \le \frac{(1-r)^{L+1}}{(L+1)r} \le b_L.$$

The identity is the harmonic inverse-binomial construction; the inequality is the deterministic truncation correction used here. Averaging over orderings gives $\nu_L \le \mu_t^{\mathrm{s}} \le \nu_L + b_L$, where $\nu_L = \mathbb{E}[Z_L \mid \mathcal{H}_t]$.

The ordering-level scores are conditionally iid in $[0, H_L]$. Hoeffding therefore gives $|\bar{Z}_m - \nu_L| \le \rho_m$ with conditional probability at least $1-\alpha_t$, proving Eq. (39). Its width is the deterministic quantity $2\rho_m + b_L$. The choices in Eq. (40) make $\rho_m \le w/4$ and $b_L \le w/2$. Appendix J gives the constructive proof and cost analysis.

The evaluator observes at most L trials per ordering. Its conditional expected count is

$$\mathbb{E}[\min(K, L) \mid r] = \sum_{k=0}^{L-1}(1-r)^k = \frac{1-(1-r)^L}{r} \le \frac{1}{\delta}.$$

Thus $C_t \le mL$ pathwise and $\mathbb{E}[C_t \mid \mathcal{H}_t] \le m/\delta$. The interval size is fixed before the current draws; there is no optional inner-mean estimate to correct. Finally, conditional coverage remains valid at an evidence state chosen using earlier histories. A union bound over $H+1$ rounds, each at level $\alpha/(H+1)$, proves simultaneous coverage. □

The fixed-m interval does not authorize inspecting every smaller outer count with the same error allocation. Such inspection requires additional spending over counts or a time-uniform outer confidence sequence. The finite-budget theorem deliberately uses one prescribed outer endpoint at each acquisition round.

The smoothed target admits the slightly sharper implication

$$\mu^{\mathrm{s}} \le -\log[\delta + (1-2\delta)p^\star] \quad \Longrightarrow \quad \bar{q} \ge p^\star,$$

because $\mathbb{E}r_\delta(Q) = \delta + (1-2\delta)\bar{q}$. This is the optimal cutoff $c_{\mathrm{s}}^\star$ in Section H. The acquisition simulation in Appendix L.1 retains the stricter cutoff $\ell^\star$. Neither cutoff is required for the interval-width theorem. A non-strict upper-confidence crossing proves the non-strict success requirement $\bar{q} \ge p^\star$; the testing-complexity results separately protect closed nulls and concern strict alternatives.

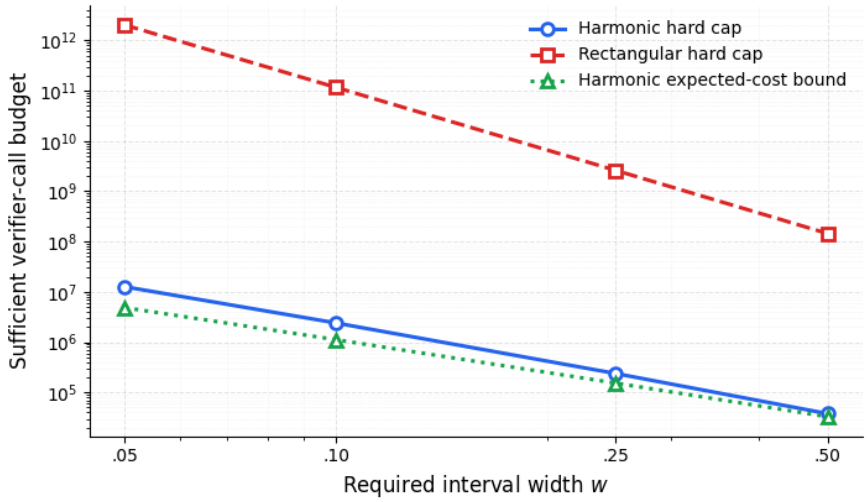

Figure 4: Width versus guaranteed evaluation budget for the same smoothed residual, with $\delta = .05$ and local error .05. Solid and dashed curves are sufficient hard caps, not necessary costs; the dotted curve bounds expected harmonic calls. These are numerical evaluations of Theorem 4 and Eq. (44), not measured speedups. Table J.1 gives the exact sample allocations.

For completeness, the rectangular comparator in Table J.1 evaluates this same smoothed functional. The derivative of $-\log r_\delta(q)$ has magnitude at most $\Lambda_\delta = (1-2\delta)/\delta$, and its range is R_δ. Split the local failure probability equally between all inner intervals and the outer interval. The sufficient counts are

$$\begin{aligned} m_{\mathrm{rect}} &= \left\lceil \frac{8R_\delta^2 \log(4/\alpha_t)}{w^2} \right\rceil, \\ n_{\mathrm{rect}} &= \left\lceil \frac{8\Lambda_\delta^2 \log(4m_{\mathrm{rect}}/\alpha_t)}{w^2} \right\rceil. \end{aligned} \tag{44}$$

All m_{rect} inner sample means have half-width $b = \sqrt{\log(4m_{\mathrm{rect}}/\alpha_t)/(2n_{\mathrm{rect}})}$ with joint failure probability at most $\alpha_t/2$. Their transformed contribution to interval width is at most $2\Lambda_\delta b \le w/2$. The outer half-width is $R_\delta\sqrt{\log(4/\alpha_t)/(2m_{\mathrm{rect}})} \le w/4$. Adding the two contributions gives width at most w and total count $m_{\mathrm{rect}}n_{\mathrm{rect}}$. The large displayed caps describe this explicit construction only. They do not exclude substantially better inner adaptation, variance-dependent boundaries, or alternative unbiased estimators.

Table J.1: Deterministic sufficient verifier-call budgets for estimating the same smoothed residual, with $\delta = .05$ and local error .05. Harmonic counts follow Theorem 4; rectangular counts use simultaneous inner Hoeffding intervals and an outer Hoeffding interval, with their formulas in Appendix J. This comparison does not lower-bound the cost of optimized nested or variance-adaptive procedures.

Width w	Inner cap L	Outer count m	Harmonic hard cap	Rectangular cap
.50	23	1647	37,881	1.448×10^8
.25	31	7659	237,429	2.596×10^9
.10	43	55,843	2,401,249	1.158×10^{11}
.05	52	243,099	12,641,148	2.028×10^{12}

Figure 4 illustrates the relationship between width and the guaranteed evaluation budget under the smoothed residual criterion.

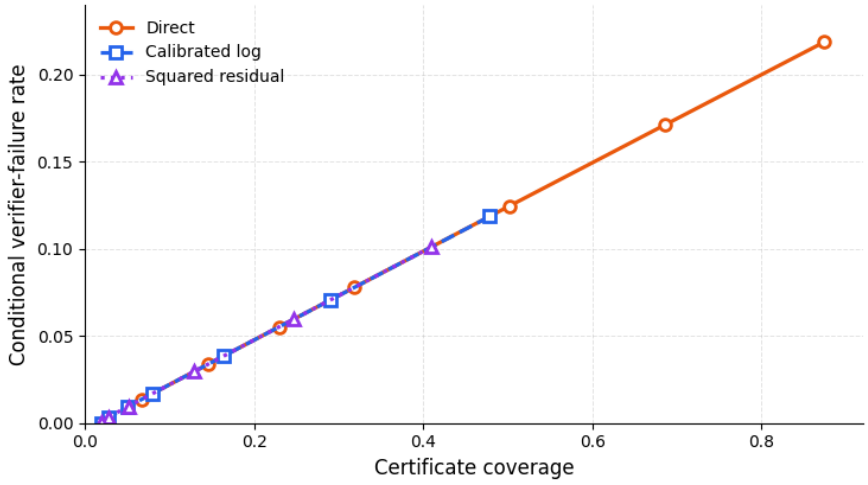

Figure 5: Exact risk–coverage on an equally weighted population of 51 endpoint ordering laws with success probabilities $.50, .51, \ldots, 1$. All methods use local error .05, cap 512, and targets $.50, .60, .70, .80, .85, .90, .95$. Every surrogate uses its optimal safe cutoff; undefined conditional risks at zero coverage are omitted. The smaller residual risk accompanies lower coverage and does not establish a matched-risk advantage.

K Calibrated numerical study and general-alternative controls

Table 2 reports the calibrated all-success counts in the main paper. This appendix checks interior alternatives, finite caps, and risk–coverage behavior under the same declared endpoint experiment.

Figure 5 reports the exact risk–coverage trade-off across the endpoint ordering laws.

The calibrated study uses the same exact endpoint-ordering experiment and recursion as Appendix L, but replaces the conservative residual cutoff by the convex-envelope cutoff. No boundary is fitted to the evaluation outcomes. All finite-budget mixtures use the six fixed relative offsets and equal weights described in Appendix L. No row pools across targets or treats repeated looks as independent observations. Table K.1 reports the calibrated endpoint-family power and call-cost comparison.

Table K.1: Exact endpoint-family comparison at local error .05 and cap 2048, with $\delta = (1 - p^\star)/4$. Both-target rows use bad-order mass $v = \eta_u/4$; mismatch rows use $v = (\eta_u + 1 - p^\star)/2$. Log means optimally calibrated clipped log loss. Calls are restricted means, counting the full cap for nonacceptance. Displayed unit powers are rounded.

$p^\star$	Regime	Direct power	Log power	Direct calls	Log calls
0.50	Both targets	1.000000	1.000000	8.77	24.92
0.50	Target mismatch	1.000000	0.002772	180.04	2042.36
0.80	Both targets	1.000000	1.000000	22.13	144.51
0.80	Target mismatch	1.000000	0.00011	223.51	2047.78
0.90	Both targets	1.000000	0.999982	44.55	396.22
0.90	Target mismatch	0.999901	1.22e-05	395.11	2047.98

The squared-residual control uses normalized cutoff $(1 - p^\star)^2$ and the same observations. At targets $.50, .80, .90$, its powers on the both-target rows are respectively $1.000000, 0.999996, 0.397920$, with restricted mean costs $33.36, 327.18, 1640.18$. The last low power is a finite-budget limitation, since its population target is satisfied. In contrast, every mismatch row lies inside both surrogate nulls; the squared-residual powers there are $0.0009489, 1.298e-06, 1.547e-11$. The recursion defines the corresponding unrounded probabilities and complete crossing trajectories. Table K.2 reports the analytical minimax bounds and information-rate ratios.

Table K.2: Analytical lower and upper bounds from Theorem 2, not uncertainty intervals. The common endpoint alternative has bad-order mass $v = \rho\eta_u$, $\delta = (1-p^\star)/4$, and local error .05. The last column is $d(v, 1-p^\star)/d(v, \eta_u)$, the limiting minimax ratio as $\alpha \downarrow 0$ with the other parameters fixed.

$p^\star$	ρ	Direct minimax interval	Log minimax interval	Information-rate ratio
0.50	0.25	$[6.89, 8.32]$	$[20.55, 22.36]$	2.98
0.50	0.50	$[10.65, 12.56]$	$[51.88, 55.06]$	4.87
0.80	0.25	$[17.92, 19.16]$	$[122.04, 123.89]$	6.81
0.80	0.50	$[22.74, 24.21]$	$[318.62, 321.87]$	14.01
0.90	0.25	$[35.90, 37.09]$	$[334.69, 336.55]$	9.32
0.90	0.50	$[43.32, 44.68]$	$[877.77, 881.03]$	20.26
0.95	0.25	$[71.22, 72.39]$	$[832.91, 834.77]$	11.69
0.95	0.50	$[83.43, 84.73]$	$[2187.81, 2191.06]$	26.22

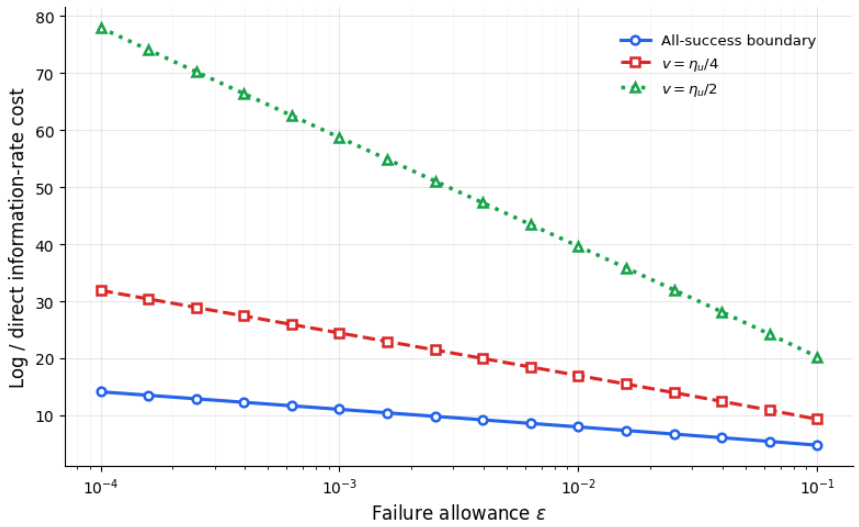

Figure 6: Exact information-rate ratios after optimal clipped-log calibration, with $\delta = \varepsilon/4$. Interior alternatives remain a fixed fraction of the residual-null boundary. Approaching that boundary increases the separation beyond its all-success value, as quantified by Eq. (27). These are analytical rate ratios, not finite-sample measured speedups.

Figure 6 illustrates the analytical information-rate ratios under optimal clipped-log calibration.

K.1 Calibration-program and bit-complexity checks

For each target $.50, .80, .90$, the calibration audit solves a linear program over probabilities $0, .001, \ldots, 1$, with nonnegative distribution weights, total weight one, and mean fixed to the target. The objectives are the linear, clipped-log, and squared residuals. All nine optima agree with the continuum formulas within 10^{-8}. This is a reproducible numerical check on these targets, not a replacement for the continuum proof. Figure 7 compares the necessary lower bound with sufficient harmonic verifier budgets at fixed width and confidence.

The bit-complexity study evaluates Eq. (43) and the harmonic budgets at $\delta \in \{.1, .05, .01, .005\}$ and $w \in \{.1, .05\}$. Every pair satisfies the exact separation and KL-inequality assertions required by the converse proof. At width .05, the exact two-point lower bounds for these four smoothing levels are 364.57, 781.73, 4118.86, and 8290.26, whereas the harmonic expected-cost bounds are 1,718,910, 4,861,980, 44,939,100, and 111,296,800. The gap reflects unmatched logarithmic factors and constants; neither the figure nor the table asserts a tight finite-parameter approximation.

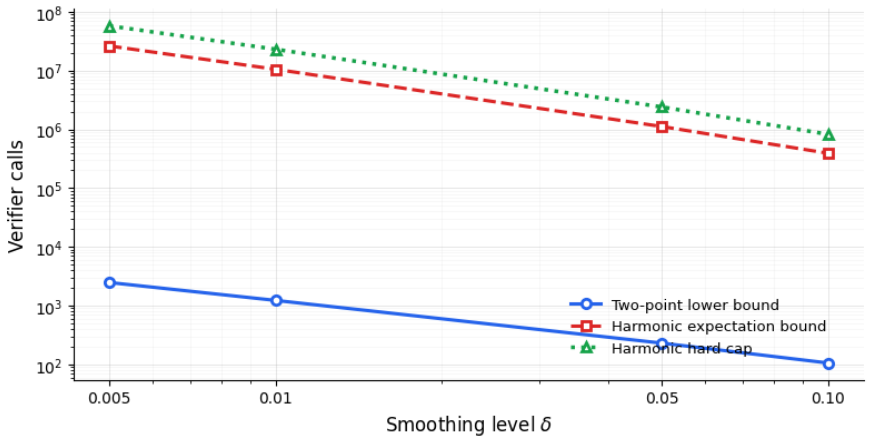

Figure 7: Necessary versus sufficient verifier cost at width $w = .1$ and confidence .95. The solid line is the exact two-point lower bound in Eq. (43); the other curves are sufficient harmonic budgets. All values are analytical evaluations. Agreement in polynomial exponents leaves a substantial finite-parameter gap.

K.2 Risk, coverage, and false-certificate estimands

For the finite population of 51 endpoint environments, let a_i be exact acceptance probability at environment success q_i. Coverage is $51^{-1}\sum_i a_i$, joint served-output failure is $51^{-1}\sum_i a_i(1-q_i)$, and conditional risk is their ratio when coverage is positive. False-certification probability is instead $51^{-1}\sum_i a_i \mathbf{1}\{q_i \le p^\star\}$. A failed fresh output does not by itself identify a false certificate: a generator can satisfy the success target and still fail on an individual draw. Conversely, population risk averaged over accepted instances does not establish a per-instance probability certificate. These two quantities use different events and denominators.

L Exact numerical protocol and complete controls

We record the conservative-cutoff numerical study. Its η_δ uses $\ell^\star = -\log(p^\star + \delta)$ rather than the optimally calibrated η_u used in the main results. Both studies use the same exact recursion while retaining their distinct cutoffs.

The endpoint and finite-state calculations are deterministic. Their parameter choices are specified below, and no random seed is involved. We evaluated all-success counts at $p^\star \in \{.50, .80, .90, .95\}$ with $\delta = (1-p^\star)/4$ and local error .025. At $p^\star = .95$, $\delta = e^{-6}$, and local error .025, direct endpoint and uniform-mixture tests need 72 and 105 all-success draws. The probability and residual index-spent Hoeffding counts are 4216 and 246276; the sharp residual endpoint count for this conservative cutoff is 477. These are exact analytical counts, not observed model runtimes.

For Table L.3, we evaluated $p^\star \in \{.5, .8, .9\}$, $\delta = (1-p^\star)/4$, local error .05, and cap 2048. At each target the four ordering laws have bad-order probabilities

$$\eta = 1 - p^\star, \qquad \eta = \eta_\delta, \qquad \eta = \eta_\delta/4, \qquad \eta = (1-p^\star)/2.$$

The first two are boundary controls for the direct and residual nulls, respectively; the latter two yield the rows in Table L.3. The six alternatives have equal prior weights and offsets $z \in \{.1, .25, .5, .75, .9, 1\}$. Likelihood mixtures are computed in log space. Because the likelihood is increasing in the number of successes, the largest accepted failure count completely specifies the boundary at each time. For a nonexistent crossing, $k_n = -1$ and no mass is absorbed.

$$s_{n,f} = \mathbf{1}\{f > k_n\}\{(1-v)s_{n-1,f} + v s_{n-1,f-1}\}, \tag{45}$$

Starting from $s_{0,0} = 1$ and zero elsewhere, the unabsorbed mass before applying the boundary at time n is $\tilde{s}_{n,f} = q s_{n-1,f} + (1-q)s_{n-1,f-1}$. The first-crossing probability at n is $a_n = \sum_{f \le k_n} \tilde{s}_{n,f}$.

Thus power is $\sum_{n=1}^{N} a_n$, and restricted mean calls are $\sum_{n=0}^{N-1} \sum_f s_{n,f}$. The recursion is exact for the declared finite-state model, evaluated in floating-point arithmetic. It is not an empirical interval for an unknown population. We checked that absorbed plus surviving mass remained within 2×10^{-10} of one. Table L.1 reports the exact boundary-control crossing probabilities.

Table L.1: Exact boundary controls at cap 2048 and local error .05. A residual-boundary state has mixture success $1-\eta_\delta > p^\star$, so high direct acceptance in its row is valid, not a false direct certificate. Values displayed as 1.0000 are rounded.

$p^\star$	Boundary law	Direct crossing probability	Residual crossing probability
.50	Direct boundary	.040445	$< 10^{-6}$
.80	Direct boundary	.041747	$< 10^{-6}$
.90	Direct boundary	.040763	$< 10^{-6}$
.50	Residual boundary	1.0000	.042435
.80	Residual boundary	1.0000	.037104
.90	Residual boundary	1.0000	.032250

For Table L.2, fixed-look procedures inspect the 16 times $32, 64, \ldots, 512$. They use the exact binomial upper-tail probability at each look and reject when it is at most .05 or $.05/16$. The mixture inspects all 512 times. We evaluated null crossing at $q = .8$ and power at $q = .9$ without fitting an alternative to either path. The absorbing recursion charges every method its full cap when it never accepts.

For Figure 8, the environment population places weight $1/51$ on each $q \in \{.50, .51, \ldots, 1.00\}$. If A_j is a method's exact acceptance probability in environment j, its population coverage is $C = \sum_j A_j/51$, joint acceptance and fresh failure is $F = \sum_j A_j(1 - q_j)/51$, and conditional failure is F/C when $C > 0$. Each target defines a separate curve point, not a simultaneous claim over adaptively selected targets. The population mass of boundary-or-bad certifications is $\sum_j A_j \mathbf{1}\{q_j \leq p^\star\}/51$. This quantity is distinct from future-output failure.

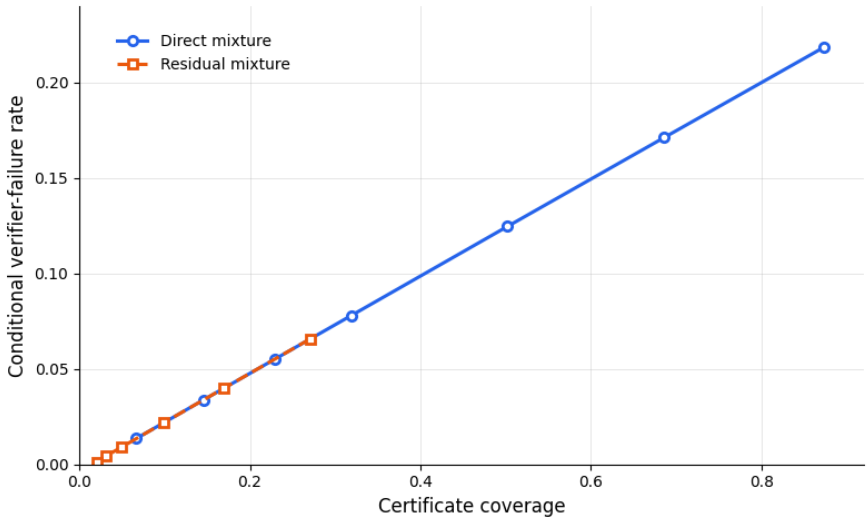

Figure 8: Exact risk–coverage curves on an equally weighted population of 51 two-point ordering environments with mixture success probabilities $.50, .51, \ldots, 1$. Targets are $.50, .60, .70, .80, .85, .90, .95$; cap 512 and local error .05 are fixed. Each point averages exact acceptance probabilities, and risk uses a fresh independent output. The residual rule has zero coverage at target .95, so its conditional risk is undefined and that point is omitted. The curves describe this declared population, not a language-model benchmark.

Figure 8 shows the exact risk–coverage curves across the declared two-point ordering environments.

Table L.2: Exact crossing probabilities for $p^\star = .80$, local error .05, and cap 512. Boundary crossing is evaluated at $q = .80$, including equality in the null. The first two procedures inspect $32, 64, \ldots, 512$ and use exact binomial upper-tail tests, respectively without and with a 16-look correction. The mixture is inspected at every trial. Power and restricted mean calls use the fixed alternative $q = .90$. These probabilities are obtained by recursion, not estimated from repeated streams.

Procedure	Boundary crossing	Power at $q = .90$	$\mathbb{E}[\tau \wedge 512]$
Fixed-look test reused	.170818	.9999998	68.67
Finite-look Bonferroni	.013890	.9999250	134.51
Direct mixture e-process	.036750	.9998626	102.21

Table L.3: Exact finite-budget power and restricted mean verifier calls with cap $N = 2048$, local error .05, and $\delta = (1 - p^\star)/4$. "Both targets" uses bad-order probability $\eta_\delta/4$; "Residual gap" uses $(1 - p^\star)/2$. Calls include all capped nonacceptances. Power displayed as 1.0000 is rounded, not an assertion of probability exactly one.

$p^\star$	Ordering law	Direct power	Residual power	Direct calls	Residual calls
.50	Both targets	1.0000	1.0000	7.51	44.85
.80	Both targets	1.0000	1.0000	20.56	226.38
.90	Both targets	1.0000	.9964	42.18	605.11
.50	Residual gap	1.0000	.00262	29.35	2042.69
.80	Residual gap	1.0000	.0000194	102.22	2047.96
.90	Residual gap	1.0000	$< 10^{-6}$	223.02	2048.00

L.1 Finite-width Monte Carlo acquisition check

We separately simulated the harmonic evaluator, rather than estimating individual ordering probabilities. This study uses $\delta = .05$, width .10, target .50, $H = 7$, and total residual error .05, split equally over the eight rounds. The allocation gives $L = 43$ and $m = 87{,}321$ orderings per round. For a given population residual μ_t^{s}, the two ordering-specific smoothed success probabilities are $\exp[-(\mu_t^{\mathrm{s}} - .02)]$ and $\exp[-(\mu_t^{\mathrm{s}} + .02)]$, each with probability one half. They lie in $[\delta, 1 - \delta]$ for all evaluated states; the underlying verifier probability is $(r - \delta)/(1 - 2\delta)$.

The three population residual profiles are

$$\begin{aligned} \text{positive: } & (1.20, 1.04, .88, .72, .56, .40, .24, .08), \\ \text{saturating: } & (1.20, 1.04, .90, .78, .70, .66, .64, .63), \\ \text{reversing: } & (1.20, .94, 1.00, .72, .80, .60, .48, .56). \end{aligned} \tag{46}$$

The positive profile has deterministic positive drift; the other profiles test saturation and negative increments. For each profile we generated 50 independent paths using NumPy's default random generator initialized with seed 20260923. Each ordering's first-success time was drawn from its exact geometric distribution, which simulates the complete inner Bernoulli stopping law without explicitly allocating an array for every trial. Counts nevertheless charge $\min(K, L)$ simulated verifier trials, not one geometric draw.

All eight rounds were generated on the same continued path to check the bracket, including rounds after acceptance. Operational prefix cost counts only through acceptance or the cap; full-path diagnostic cost additionally counts the continuation. Coverage held at every round in $50/50$ paths for each profile, and all $50/50$ truncated-bracket checks per profile held. These are empirical Monte Carlo checks, not substitutes for the coverage theorem. Table L.4 reports the corresponding stopping and cost observations.

Figure 9 illustrates a simulated harmonic-evaluation trajectory and the corresponding population, implemented, and deeper crossing rounds.

Table L.4: Simulated harmonic evaluation on three fully specified residual profiles. The implemented crossing was identical across all 50 replications within each profile; ∞ means no crossing through $H = 7$. Prefix calls are mean $\pm$ sample standard deviation across 50 independent simulated paths, in millions, and include all evaluation calls up to the displayed stopping decision, with all eight rounds charged on timeout. They are numerical-simulator counts, not language-model measurements.

Profile	τ_{β^s}	$\hat{\tau}_s$	τ_{β^s+w}	Prefix calls (10^6)
Positive	4	5	5	$1.210350 \pm .001119$
Saturating	∞	∞	∞	$1.616963 \pm .001266$
Reversing	6	6	6	$1.424932 \pm .001231$

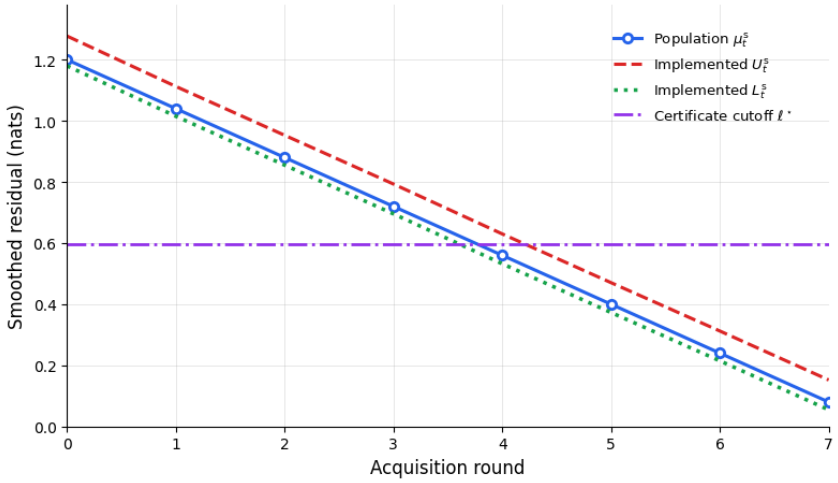

Figure 9: A simulated harmonic-evaluation trajectory, using the first replication of the positive profile in Eq. (46). The population crossing is at round 4; the implemented and deeper crossings are both at round 5. The width target is .10, and simultaneous residual error is .05 across all eight rounds. All rounds are simulated on the same continued path, including the continuation after acceptance.

M SECONDARY FINITE-LAW STRESS PANEL

The executed panel consists of fifty designed probability laws D01–D50 and their population-mean-matched homogeneous controls. It is a stress suite, not a random sample of benchmark tasks. Every law uses $K = 100$ independent ordering probabilities and $n = 200$ conditionally independent Bernoulli outcomes per ordering. The panel and control arms therefore contain 10^6 simulated verifier bits each.

The generation seed is 2026092307. A law first samples its mixture component, then its beta probability within that component, and finally its binomial count. The matched homogeneous counts follow before the next law. Population moments and log means use normalized 256-node Gauss–Jacobi quadrature within each beta component. The random-number generator is NumPy's default generator. Clipping is .01, and negative values of Eq. (16) are retained.

The 2000 percentile-bootstrap replicates resample whole ordering counts. One 2000×100 index array is generated with seed 2026092308 and reused across laws and controls. This preserves each law's marginal procedure but induces dependence across summaries; the classifications are descriptive and do not have simultaneous or anytime-valid coverage. Table M.1 reproduces the complete classification counts. Per-law point estimates, intervals, variance estimates, and decisions are in the accompanying machine-readable output rather than printed as thousands of table cells.

Across the D laws, the lower variance endpoint is positive for $32/50$ laws, the median variance estimate is .000462, and the median raw-to-jackknife correction is .002682 nats. No homogeneous control has a positive lower variance endpoint or a separated mismatch classification. The two count

Table M.1: Complete classifications for the fifty-law panel and matched homogeneous controls. All 50 laws remain in every denominator.

Arm / p	Mismatch	Aligned	Below	Unresolved	Unresolved fraction
D laws / .5	2	28	8	12	24%
D laws / .7	0	17	17	16	32%
D laws / .9	0	2	32	16	32%
Homogeneous / .5	0	37	8	5	10%
Homogeneous / .7	0	25	19	6	12%
Homogeneous / .9	0	2	35	13	26%

arrays contain 695,797 and 695,126 successes, respectively, out of 10^6 simulated trials each. These aggregates check the ledger; they do not replace per-law analysis.

M.1 Complete specification of the fifty laws

Write $\mathcal{B}(a, b; \alpha, \beta)$ for the law of $a + (b - a)V$ with $V \sim \mathrm{Beta}(\alpha, \beta)$. For $a < m < b$ and $c > 0$, define

$$F(m; a, b, c) = \mathcal{B}\left(a, b; c\frac{m-a}{b-a}, c\frac{b-m}{b-a}\right). \tag{47}$$

This law has mean m and variance $(m - a)(b - m)/(c + 1)$. Table M.2 uses

$$\begin{aligned} C(m) &= F(m; \max\{.025, m - .07\}, \min\{.989, m + .07\}, 55)\,, \\ T(w) &= w\,\mathcal{B}(.025, .095; 2, 4) + (1 - w)\,\mathcal{B}(.900, .989; 9, 2), \\ B(m, c) &= F(m; .025, .989, c), \qquad L(m) = F(m; m - .10, m + .10, 18), \\ N(m) &= F(m; \max\{.025, m - .08\}, \min\{.989, m + .08\}, 18)\,. \end{aligned} \tag{48}$$

A mixture in the second line selects its component independently for every ordering. No parameter is inferred from the observed fractions. For each law d, the paired control is a point mass at $\mathbb{E}_{P_d} Q$.

Table M.2: All fifty probability laws. The five columns contain concentrated, two-band, broad, lower-mean, and near-target environments.

Concentrated	Two-band	Broad	Lower mean	Near target
D01 $C(0.57)$	D02 $T(0.12)$	D03 $B(0.65, 0.65)$	D04 $L(0.24)$	D05 $N(0.505)$
D06 $C(0.76)$	D07 $T(0.2)$	D08 $B(0.79, 1.2)$	D09 $L(0.43)$	D10 $N(0.715)$
D11 $C(0.96)$	D12 $T(0.32)$	D13 $B(0.9, 0.85)$	D14 $L(0.36)$	D15 $N(0.905)$
D16 $C(0.86)$	D17 $T(0.06)$	D18 $B(0.55, 2.2)$	D19 $L(0.47)$	D20 $N(0.53)$
D21 $C(0.66)$	D22 $T(0.26)$	D23 $B(0.83, 1.7)$	D24 $L(0.31)$	D25 $N(0.745)$
D26 $C(0.93)$	D27 $T(0.15)$	D28 $B(0.7, 0.5)$	D29 $L(0.4)$	D30 $N(0.925)$
D31 $C(0.79)$	D32 $T(0.35)$	D33 $B(0.88, 2.8)$	D34 $L(0.28)$	D35 $N(0.51)$
D36 $C(0.975)$	D37 $T(0.09)$	D38 $B(0.61, 0.9)$	D39 $L(0.46)$	D40 $N(0.72)$
D41 $C(0.63)$	D42 $T(0.22)$	D43 $B(0.93, 1.1)$	D44 $L(0.38)$	D45 $N(0.945)$
D46 $C(0.88)$	D47 $T(0.045)$	D48 $B(0.76, 3.5)$	D49 $L(0.49)$	D50 $N(0.775)$

For $F(m; a, b, c)$, the second moment is $m^2 + (m - a)(b - m)/(c + 1)$. Mixture means and variances follow by weighting first and second moments; log means are integrated componentwise. Equations (47)–(48) therefore determine every population value independently of the finite samples.

M.2 MACHINE-READABLE LEDGER AND REPLAY

The companion file `P41_probability_law_counts.csv` has SHA-256 `f6e3c6f27658c0c0e1927ad48e3e386c3874dc6ddbdee164a7edd0ed0f594212`. It contains 100 ordered counts for every D-law and homogeneous arm, with 200 trials behind each count. The script `reproduce_probability_panel.py` has SHA-256 `3229e424d2947b4f851c2fe90414d655b9be75e9787b58586bd29959bed7a318`. Running it with NumPy reproduces the point estimates, the shared-seed ordering bootstrap, Table M.1, and the companion per-law estimates file. The ledger replaces the former pages of typeset raw counts; it is probability-model data, not generated-program output.

N FROZEN-MODEL AUDIT PROTOCOL AND EXECUTED TABLES

The code-generation audit uses `Qwen/Qwen2.5-Coder-1.5B-Instruct` (Hui et al., 2024), with model and tokenizer revision `d18ecad881a290f0c792f1c2b5dbee1820d4a424`, and EvalPlus `v0.3.1` (Liu et al., 2023). It comprises 25 HumanEval+ tasks and 25 MBPP+ tasks, 100 independently sampled evidence orderings per task, and 200 independent generations per ordering, for a total budget of 10^6 outputs.

Within each benchmark, task identifiers were ranked by the SHA-256 digest of the UTF-8 string `20260923|benchmark|task_id`, using `HumanEval+` and `MBPP+` as the benchmark names. The first 25 identifiers were selected as target tasks, and the subsequent six identifiers were assigned as a disjoint demonstration pool. Each target was presented with the same six public problem-and-reference-solution cards, while its own problem statement and function signature remained fixed. Evidence orderings were sampled independently with replacement from the uniform distribution over the $6!$ permutations. Repeated permutations were retained as distinct ordering draws. The target reference solution and hidden verifier assertions were excluded from the model context.

Decoding uses temperature .7, nucleus mass .95, a 512-token completion cap, the checkpoint chat template, and its end-of-sequence tokens. Randomness is indexed by task, ordering draw, and inner draw. The isolated, network-free verifier assigns one exactly when a completion passes the expanded suite; syntax errors, wrong signatures, assertion failures, and timeouts receive zero. Infrastructure failures are rerun before analysis. The released audit retains task and demonstration manifests, prompts, permutations, generation indices, raw outputs, verifier outcomes, and stopping reasons.

Per-task analysis uses the 100 ordering counts to compute $\overline{q}$, the raw clipped log mean, Eq. (18), and Eq. (16), with clipping at .01 and 2000 complete-ordering bootstrap resamples. The fixed targets are $.5, .7, .9$ with a .05-nat separation margin. Every task remains in every denominator. Each mean-matched homogeneous control uses the fitted task mean and the same 100×200 simulated count budget; it serves as a diagnostic rather than additional model evidence.

Checkpoint audit: task-level decisions, aggregate statistics, controls, and ordering sensitivity. The four tables report the checkpoint audit, comprising task-level decisions (Table N.1), aggregate checkpoint statistics (Table N.2), mean-matched homogeneous controls (Table N.3), and ordering-sensitivity diagnostics (Table N.4).

Table N.1: Qwen2.5-Coder-1.5B-Instruct task-level decisions under the frozen EvalPlus protocol. Each benchmark contains 25 tasks, 100 independently sampled evidence orderings per task, and 200 generations per ordering. Decisions use the fixed 0.05-nat separation margin and ordering-clustered percentile intervals.

Benchmark	p	Mismatch	Aligned	Below	Unresolved	Unresolved fraction
HumanEval+	.5	2	13	4	6	24%
HumanEval+	.7	1	5	11	8	32%
HumanEval+	.9	0	1	19	5	20%
MBPP+	.5	3	11	5	6	24%
MBPP+	.7	1	3	13	8	32%
MBPP+	.9	0	0	21	4	16%
Combined	.5	5	24	9	12	24%
Combined	.7	2	8	24	16	32%
Combined	.9	0	1	40	9	18%

Table N.2: Aggregate checkpoint statistics for the frozen-model audit. Pooled success is computed over all generations; remaining summaries are across tasks. Variance is the unbiased ordering-level estimate in Eq. (16).

Benchmark	Pooled success	Median $\overline{q}$	Median $\widehat{\mu}_{\mathrm{JK}}$	Median $\widehat{s}^2_{\mathrm{unb}}$	Positive variance LB	Median correction
HumanEval+	0.6558	0.676	0.444	0.0049	15/25	0.0019
MBPP+	0.5974	0.612	0.536	0.0068	16/25	0.0021
Combined	0.6266	0.641	0.489	0.0057	31/50	0.0020

Table N.3: Mean-matched homogeneous controls for the checkpoint audit. Each control uses the fitted task mean and the same 100×200 count budget. All 50 controls remain in every denominator.

p	Mismatch	Aligned	Below	Unresolved	Unresolved fraction
.5	0	31	9	10	20%
.7	0	10	29	11	22%
.9	0	1	42	7	14%

Table N.4: Ordering-sensitivity diagnostics for the checkpoint study. Jensen gap is $\widehat{\mu}_{\mathrm{JK}} + \log \overline{q}$. Statistics summarize the 50 model tasks and their mean-matched controls.

Statistic	Checkpoint tasks	Homogeneous controls
Median $\widehat{s}^2_{\mathrm{unb}}$	0.0057	0.000004
Median Jensen gap (nats)	0.0186	0.00021
Positive variance lower endpoint	31/50	1/50
$J > 0.05$ nats	14/50	0/50
$J > 0.10$ nats	7/50	0/50
Median raw–jackknife difference	0.0020	0.0018