CALIBRATION UNDER BEST-OF-N SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

How far does calibration at small Best-of-N budgets constrain calibration at larger ones? For a fixed predictor under Best-of-N selection, we derive a bound $d_{K,N}$ on population calibration error when calibration holds at budgets $1, \dots, K$. The error at the next budget is $(K+1)4^{-K}$, and the error is order one when $N \asymp K^2$. A single model whose predictor is the raw score has calibration error of this order at every budget $N > K$. We extend the analysis to finite samples: an extrapolation bound recovers $d_{K,N}$ when moment uncertainty is zero and yields bounds on residual moments at larger budgets from finite-sample calibration estimates. A minimax result characterizes what additional passive labels reveal about the mean residual under selection, given calibration constraints and within-prompt ranks. We validate the theory on synthetic data and on language-model generations.

1 INTRODUCTION

A verifier reports a correctness probability (Lightman et al., 2024; Park et al., 2025). That probability may be well calibrated for an ordinary candidate yet miscalibrated for the winner of a large pool. The same score can lie in the upper tail for one prompt and near the median for another. Increasing the pool size changes the distribution of prompts at each score level, so the reported probability can drift even when the prompt distribution is fixed. This raises the question: how much does calibration at small inference budgets constrain calibration at larger ones?

The issue persists under perfect calibration at the population level. In the construction with two prompt types in App. B.1, ordinary scores are perfectly calibrated. As the budget grows, selected accuracy increases from 0.50 to 0.55, while mean confidence increases to 0.75. This calibration gap occurs while answer quality improves; reward hacking that degrades quality at large budgets is a separate phenomenon (Huang et al., 2025; Gao et al., 2023). Calibration must account for how selection changes the distribution within each reported confidence level. Suppose a fixed predictor is known to be calibrated at budgets $1, \dots, K$. We ask how large its calibration error can be at a target budget $N > K$, over all distributions consistent with these constraints. These calibration conditions constrain the population moments at the known budgets. This setting arises when historical calibration information at small budgets is available and generating or labeling new large pools is costly. We give an explicit formula for the extrapolation bound $d_{K,N}$ and show that it is tight within a universal factor. For example, full calibration at budgets $1, \dots, 16$ guarantees population expected calibration error (ECE) at most 0.05 through budget 63. Fig. 1 illustrates how the bound follows from polynomial approximation of the selection weight.

Theorem (Informal; see Thm. 1)**.** *Suppose a fixed predictor is fully calibrated at budgets $1, \dots, K$. Then its population calibration error at any budget $N > K$ is at most $d_{K,N}$, a function of the Chebyshev nodes. In particular, at budget $K+1$, $d_{K,K+1} = (K+1)4^{-K}$. A single model whose predictor is the raw score has calibration error within a universal factor of $d_{K,N}$ at every budget $N > K$.*

The error $d_{K,N}$ is order one when $N \asymp K^2$: calibration at K budgets yields weaker error bounds beyond roughly K^2 (Cor. 1). We extend this to finite samples, where each calibration moment is known only through an error bound. The resulting extrapolation bound (Thm. 2) balances weight approximation against amplification of estimation error and reduces to $d_{K,N}$ when the moment bounds are zero. A minimax characterization quantifies what additional passive labels reveal about the mean residual under selection and the calibration constraints (App. D).

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.

The key technical tool is the connection between calibration extrapolation and best-L^1 polynomial approximation. Each budget's calibration constraint becomes a polynomial moment condition via each score's relative rank U within its prompt, and the distance of the target weight from the known polynomial span determines the remaining error. Interpolation at the roots of a Chebyshev polynomial of the second kind solves this approximation problem.

Related Work. Dorner et al. (2026) show that a finite prefix of Best-of-N accuracies can be arbitrarily close to a curve with an interior limiting accuracy while the alternative limits are zero or one. Our target is the worst-case calibration error at a finite budget given calibration constraints. Under assumptions on the reward tail, Li et al. (2026a) extrapolate Best-of-N policy gradients from smaller rollout groups through moment cancellation. Their moments correct bias in the training estimator; ours express calibration constraints on a fixed predictor. Related scaling predictions use tail structure of the reward distribution (Li et al., 2026b) or a beta-binomial model for pass@k (Kazdan et al., 2025). Pass@k concerns whether any candidate is correct (Chen et al., 2021), whereas Best-of-N returns the scorer's choice (Shyamal et al., 2026).

More broadly, calibration under selection has been studied in auctions (McMahan and Muralidharan, 2012). Multicalibration supports transfer across populations (Hébert-Johnson et al., 2018; Kim et al., 2022), and Wu et al. (2024) characterize a linear structure of density ratios under distribution shift. Dong et al. (2026) give a condition for expectation consistency preserving reliability under covariate shift. We quantify the information in the Best-of-N weight family under conditionally i.i.d. sampling. Best-of-N policy analysis provides the change of measure for order statistics (Beirami et al., 2025). The resulting calibration shift is related to selection bias and the winner's curse (Zrnic and Fithian, 2025); our estimand is correctness conditional on a probability report.

Organization. The rest of the paper is organized as follows. Sec. 2 introduces the problem setting and the role of relative rank in selection. Sec. 3 presents the main extrapolation bound and its quadratic budget scale. Sec. 4 extends the analysis to finite samples and additional labels. Sec. 5 presents experiments validating the theoretical predictions, and Sec. 6 concludes. Proofs, further theoretical results, and experimental details can be found in the Appendix.

2 Calibration under Selection

Draw a prompt $X \sim \mu$ and, conditional on $X = x$, draw candidate observations Z_i i.i.d. from a fixed law Q_x. A candidate response Z to prompt X has ranking score $S \in [0, 1]$ from a verifier and binary correctness Y. A fixed probability predictor maps each candidate's available features to $C \in [0, 1]$, without observing the correctness label. Selection uses S, and the predicted probability does not depend on competing candidates. The prediction may be the verifier's score itself or a transformation of candidate features. All fitting is performed on independent data. All spaces are standard Borel. Attach an independent $V \sim \mathrm{Unif}[0, 1]$ to each candidate for tie-breaking. At sampling budget N, select $I_N = \arg\max_{i \le N}(S_i, V_i)$ in lexicographic order, giving uniform tie-breaking. Reports may use a candidate's own tie variable and remain fixed as competitors are added. Write $P = P_1$ for the ordinary candidate law augmented by the prompt and tie variable, and P_N for the selected law. Each prompt contributes one selected answer. Expectations are denoted by $\mathbb{E}$ and $\mathbb{E}_N$.

Definition 1 (Calibration targets)**.** Define the residual $R = Y - C$. Mean calibration at budget N means $\Delta_N := \mathbb{E}_N[R] = 0$. Full calibration is the condition

$$\mathbb{E}_N[Y \mid C] = C \quad P_N\text{-almost surely.} \tag{1}$$

For measurable tests $h : [0, 1] \to [-1, 1]$, define the residual moments and population expected calibration error (Gupta and Ramdas, 2021; Błasiok and Nakkiran, 2024) by

$$\theta_{N,h} = \mathbb{E}_N[Rh(C)], \tag{2}$$

$$\mathrm{ECE}_N(C) = \mathbb{E}_N|\mathbb{E}_N[R \mid C]| = \sup_{|h| \le 1} |\theta_{N,h}|. \tag{3}$$

Full calibration controls every bounded measurable confidence test. For a fixed partition $(I_b)_{b=1}^J$, use the tests $h_b(C) = \mathbf{1}\{C \in I_b\}$. The population counterpart of binned ECE (Pakdaman Naeini et al., 2015; Guo et al., 2017) is $\sum_b |\theta_{N,h_b}| \le \mathrm{ECE}_N(C)$. Conditional bin reliability additionally depends on selected bin mass $q_{N,b} = P_N(C \in I_b)$.

Selection can change reliability even for perfectly calibrated ordinary scores. In a construction with two prompt types, the same high score is common on one prompt type and rare on the other. Larger pools increasingly return that score from both types, changing their proportions within the reported confidence level. Accuracy rises from 0.50 toward 0.55, while mean confidence rises from 0.50 toward 0.75. App. B.1 gives the complete construction and the resulting curves.

Selection Weights. Let $F_x(s) = \mathbb{P}(S \le s \mid X = x)$ be the conditional cumulative distribution function (CDF) of the score and define the randomized probability integral transform (Rüschendorf, 2009)

$$U = F_X(S^-) + V\{F_X(S) - F_X(S^-)\}. \tag{4}$$

Then $U \mid X \sim \mathrm{Unif}[0,1]$, and maximizing the augmented score (S, V) is equivalent almost surely to maximizing U. For continuous conditional scores, $U = F_X(S)$. This transform gives a candidate's relative rank in the conditional score distribution for its prompt.

Lemma 1 (Selected law)**.** *For every P-integrable function f of an individual candidate,*

$$\mathbb{E}_N f = \mathbb{E}[w_N(U) f], \qquad w_N(u) = N u^{N-1}. \tag{5}$$

The prompt marginal under P_N remains μ.

Proof sketch. A candidate with relative rank u beats all $N-1$ competitors with probability u^{N-1}. Exchangeability supplies the factor N, and $\mathbb{E}[w_N(U) \mid X] = 1$ preserves prompt mass. The full argument, including point masses in the score distribution, is in App. A. □

For each test, set $r_h(u) = \mathbb{E}[Rh(C) \mid U = u]$. Since U is uniform, the selected residual is

$$\theta_{N,h} = \int_0^1 w_N(u) r_h(u)\, du. \tag{6}$$

Thus different budgets weight the same residual differently. Calibration at every budget is equivalent to $\mathbb{E}[Y - C \mid C, U] = 0$: conditional on C, relative rank then carries no additional correctness information (Thm. 4, proved in App. A). Our question is quantitative: how much do the first K budgets alone constrain the remaining weighted residuals?

3 Calibration at Scale

Suppose full calibration holds at budgets $1, \dots, K$. We first bound the population ECE at every target budget consistent with those facts. The function $d_{K,N}$ below gives a universal bound; a single model that reports the raw score matches this bound within a fixed factor.

3.1 An Error Bound

The weights at the known budgets $1, \dots, K$ span the space

$$\mathcal{V}_K = \mathrm{span}\{w_1, \dots, w_K\} = \mathrm{span}\{1, u, \dots, u^{K-1}\}.$$

All norms below are with respect to Lebesgue measure on $[0,1]$. The approximation error of w_N by $\mathcal{V}_K$ determines the extrapolation bound; a construction with calibrated raw scores establishes its sharpness up to a universal factor.

Theorem 1 (Calibration extrapolation bound)**.** *Let $K \ge 1$ be an integer. Every predictor fully calibrated at budgets $1, \dots, K$ satisfies*

$$\mathrm{ECE}_N(C) \le d_{K,N} \qquad (N \ge 1). \tag{7}$$

For each K, there is a single model with continuous conditional score distributions and $C = S$, fully calibrated at those budgets, such that

$$\mathrm{ECE}_N(S) = |\Delta_N| = \frac{3}{16} d_{K,N} \qquad (N > K). \tag{8}$$

Here $d_{K,N} = 0$ for $N \le K$. For $N > K$, it is given by

$$d_{K,N} = 1 + 2\sum_{j=1}^{K} (-1)^j u_j^N, \qquad u_j = \frac{1 + \cos(j\pi/(K+1))}{2}. \tag{9}$$

The upper bound controls all confidence tests; the model attaining the lower bound satisfies full calibration at every known budget. The value $d_{K,N}$ is the optimal scalar approximation error; the calibrated construction matches it within the stated universal factor. Taking $h = 1$ also yields $|\Delta_N| \le d_{K,N}$ under mean calibration alone. App. C.2 records scalar duality and the identity for the next budget; Rem. 1 discusses general predictors.

Proof. Reduction to weight approximation. For every confidence test h, the function $r_h(u) = \mathbb{E}[(Y - C)h(C) \mid U = u]$ satisfies $\|r_h\|_\infty \le 1$. Full calibration at $1, \dots, K$ gives $\int p r_h = 0$ for every $p \in \mathcal{V}_K$, so

$$|\theta_{N,h}| = \left| \int_0^1 (w_N - p) r_h \right| \le \|w_N - p\|_1.$$

Taking the supremum over h and the infimum over p reduces the ECE bound to

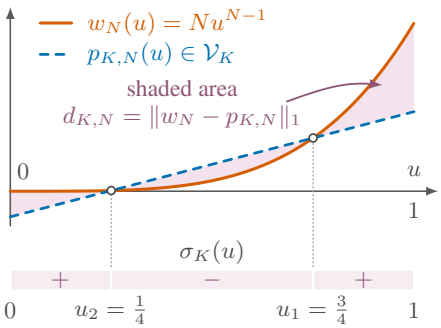
$$d_{K,N} = \inf_{p \in \mathcal{V}_K} \|w_N - p\|_1. \tag{10}$$

For $N \le K$, this is immediate from $w_N \in \mathcal{V}_K$; assume henceforth that $N > K$.

Optimal approximation. Let $p_{K,N} \in \mathcal{V}_K$ interpolate w_N at the nodes in Eq. (9), and set $\sigma_K(u) = \operatorname{sign} \mathsf{U}_K(2u - 1)$, where U_K is the Chebyshev polynomial of the second kind. The key properties are that σ_K annihilates $\mathcal{V}_K$ and that $w_N - p_{K,N}$ has sign σ_K. The latter follows from $w_N^{(K)} > 0$ on $(0, 1)$; App. C.2 verifies both properties. Thus, for every $p \in \mathcal{V}_K$,

$$\|w_N - p\|_1 \ge \int_0^1 (w_N - p)\sigma_K = \int_0^1 w_N \sigma_K = \|w_N - p_{K,N}\|_1.$$

The interpolant is therefore optimal; Fig. 1 illustrates its L^1 error as the area between the two curves. Integrating $w_N \sigma_K$ over its sign intervals gives $d_{K,N}$ in Eq. (9), proving Eq. (10). The single construction with calibrated raw scores and calibration error $\frac{3}{16} d_{K,N}$ at every $N > K$ is given in App. C.1. □



Alternating signs at the Chebyshev nodes

Figure 1: Calibration extrapolation as weight approximation (Thm. 1), illustrated for $K = 2$, $N = 5$. The target weight w_N and its optimal interpolant $p_{K,N}$ enclose total area $d_{K,N}$. Circles mark the Chebyshev nodes; the strip shows the alternating error signs.

3.2 A Quadratic Extrapolation Scale

The error curve quantifies the range over which known calibration extends. The error at the next budget is exponentially small, but substantial ambiguity can persist at much larger budgets.

Corollary 1 (Quadratic extrapolation scale)**.** *For every $K \ge 1$, $d_{K,K+1} = (K + 1)4^{-K}$. Put $M = K + 1$. As $K \to \infty$,*

$$d_{K,N} \longrightarrow \begin{cases} 0, & N/M^2 \to 0, \\ 1, & N/M^2 \to \infty. \end{cases} \tag{11}$$

If $N/M^2 \to c \in (0, \infty)$, then

$$d_{K,N} \longrightarrow \Phi(c) = 1 + 2 \sum_{j=1}^{\infty} (-1)^j e^{-\pi^2 c j^2/4}. \tag{12}$$

Proof sketch. At $N = K + 1$, the scaled Chebyshev interpolation error has the stated L^1 norm. A positive binomial representation controls $N/M^2 \to 0$; the bounds $u_j \le e^{-\pi^2 j^2/(4M^2)}$ control larger budgets. When $N/M^2 \to c \in (0, \infty)$, dominated convergence in Eq. (9) yields Φ. See App. C.3 for full proofs and nonasymptotic bounds. □

The largest node has $1 - u_1$ of order M^{-2}, while the selected weight concentrates within distance of order N^{-1} from one. Their comparison explains the quadratic scale. Below that scale, calibration at

small budgets tightly controls future error. At that scale, the calibrated construction in Thm. 1 permits ECE of constant order. The transition concerns the worst-case error bound, rather than predicting a failure budget for each verifier.

For a given error tolerance τ, budgets satisfying $d_{K,N} \le \tau$ have calibration error within that tolerance. At $K = 16$, for example, $d_{16,63} \approx 0.0481$ and $d_{16,64} \approx 0.0514$. Thus the bound at budget 63 is below tolerance 0.05, whereas at budget 64 it exceeds this tolerance. This decision uses population equalities. The next section determines what remains when these moments are estimated from finite data.

4 Extrapolation from Finite Samples

Thm. 1 characterizes the case of full calibration. In practice, calibration is verified from finite data and known only within confidence intervals. We now ask: how does this uncertainty propagate to the target budget? The resulting bound extends the same problem of weight approximation as Eq. (10). When the moment bounds are obtained from data, the resulting bounds apply only to the finite test set used for estimation.

4.1 Extrapolation with Uncertain Calibration Moments

For a finite budget set $\mathcal{B}$, moment bounds $\epsilon_n \ge 0$, and a residual bound $B_\infty > 0$, define

$$D_{\mathcal{B},N}(B_\infty, \epsilon) = \sup_{\substack{\|r\|_\infty \le B_\infty \\ |\int w_n r| \le \epsilon_n \ (n \in \mathcal{B})}} \left| \int w_N r \right|.$$

For a confidence test, take $r = r_h$ and $B_\infty = 1$. This is the largest absolute target moment consistent with these moment bounds.

Theorem 2 (Extrapolation under moment bounds)**.** *For every finite $\mathcal{B}$ and nonnegative ϵ,*

$$D_{\mathcal{B},N}(B_\infty, \epsilon) = \inf_{\alpha \in \mathbb{R}^{\mathcal{B}}} \left\{ \sum_{n \in \mathcal{B}} |\alpha_n| \epsilon_n + B_\infty \left\| w_N - \sum_{n \in \mathcal{B}} \alpha_n w_n \right\|_1 \right\}. \tag{13}$$

In particular, $D_{\{1,\dots,K\},N}(1, 0) = d_{K,N}$.

Proof sketch. For any coefficients α, split the target moment into a linear combination of the constrained moments and the remainder weighted by $w_N - \sum_{n \in \mathcal{B}} \alpha_n w_n$. The moment constraints bound the first part; the remainder is bounded by B_∞ times the L^1 weight error. This gives the upper bound in Eq. (13). Conversely, Hahn–Banach applied to the seminorm defined by the infimum, together with L^1–L^∞ duality, yields a feasible residual attaining this bound, including when the moment bounds are zero. Zero moment bounds at budgets $1, \dots, K$ with $B_\infty = 1$ recover Eq. (10). The full proof is in App. C.4. □

The first term measures amplification of moment uncertainty; the second measures unresolved weight. An accurate approximation may require large coefficients, so the optimum balances both terms. The interpolant for the next budget under zero moment bounds illustrates this sensitivity (App. C.5).

Effect on The Budget Range. Suppose only uniform moment bounds $\epsilon_n = \epsilon$ are known at budgets $n = 1, \dots, K$, with fixed $0 < \epsilon < \tau < 1$. A residual concentrated near $u = 1$ can satisfy every moment bound while having a large target moment at a larger budget. Specifically, Cor. 3 gives

$$D_{\{1,\dots,K\},N}(1, \epsilon \mathbf{1}) \ge 1 - (1 - \epsilon)^{N/K}.$$

Achieving tolerance τ therefore requires $N \le K \log(1-\tau) / \log(1-\epsilon)$. An upper bound of the same order shows that the largest budget satisfying the tolerance is $\Theta_{\epsilon,\tau}(K)$. This comparison concerns bounded residuals with only noisy moment constraints; its constants depend on the moment bound and target tolerance. The quadratic scale in Cor. 1 relies on zero calibration moments.

4.2 Bounds from Labeled Data

Fix finite nonempty budget and test sets $\mathcal{B}, \mathcal{H}$, a predictor fitted on independent data, and a failure probability $\delta \in (0,1)$. Suppose estimates $\widehat{\theta}_{n,h}$ have simultaneous error bounds $\rho_{n,h}$ with probability at least $1-\delta$. Set

$$\epsilon_{n,h} = |\widehat{\theta}_{n,h}| + \rho_{n,h}. \tag{14}$$

Applying Thm. 2 on this event gives

$$|\theta_{N,h}| \le D_{\mathcal{B},N}(1, \epsilon_{\cdot,h}) \quad (N \ge 1,\ h \in \mathcal{H}). \tag{15}$$

The same event covers every target budget: extrapolation is a deterministic consequence of the confidence intervals. Each moment meets the chosen tolerance τ whenever its extrapolation bound is at most τ. Summing the bounds for fixed bins controls binned ECE; conditional bin reliability also requires a positive lower bound on selected bin mass (Cor. 5). Full continuous ECE requires control beyond a finite test set (Lee et al., 2023; Hu et al., 2024; Rossellini et al., 2025).

Complete Pools. Generate and label n_{pool} candidates on each of m independent prompts. For each $n \le n_{\text{pool}}$, average the winning residual test over all n-subsets within a prompt, then average over prompts (Nakano et al., 2021). Every subset is distributed as n conditionally i.i.d. candidates, so the estimate is unbiased. Each average within a prompt lies in $[-1,1]$; Hoeffding's inequality and a union bound give the common confidence radius $\sqrt{2\log(2|\mathcal{B}||\mathcal{H}|/\delta)/m}$ for $\mathcal{B} \subseteq \{1,\dots,n_{\text{pool}}\}$. This supplies the confidence radius used to form the moment bounds in Eq. (14) without observing U, at a cost of mn_{pool} labels. The efficient subset formula and full proof appear in App. F.

Passive Labels with Known Population Ranks. With one ordinary labeled candidate per independent prompt and observed true U, use

$$\widehat{\theta}_{n,h} = \frac{1}{m}\sum_{i=1}^{m} w_n(U_i)(Y_i - C_i)h(C_i). \tag{16}$$

The identity for the selected law gives unbiasedness. Since $|w_n R h(C)| \le n$ and $\mathbb{E}w_n^2 = n^2/(2n-1)$, Bernstein's inequality supplies a radius of order $\sqrt{n\tau_\delta/m} + n\tau_\delta/m$, where $\tau_\delta = \log(2|\mathcal{B}||\mathcal{H}|/\delta)$. The radius and proof are in App. E. Independent reference ranks provide an alternative when population ranks are unavailable (App. H.1); App. K.5 records generation and label costs.

Additional Labels under Calibration Constraints. A complementary experiment assumes exact calibration at $1,\dots,K$ and supplies m independent observations of ordinary candidates (X,S,U,Y) with known population ranks, continuous conditional score CDFs, and $C = S$. For estimating the mean residual Δ_N, Thm. 8 characterizes the minimax error up to universal constants by $\omega_{K,N}(m^{-1/2}) \le d_{K,N}$ using a modulus-based approach related to optimal recovery. Thus labels can reduce the ambiguity left by these constraints. With $K=1$ and $N \ge 2$, the rate is $\min\{1, \sqrt{N/m}\}$, reflecting the scarcity of passive labels at high relative ranks. The full characterization, estimator, and matching lower bound are given in App. D.

5 Experiments

Two synthetic experiments validate the quantitative predictions of Thm. 1 and 8; two further experiments on language-model generations test whether rank-aware reporting mitigates calibration drift in practice. App. K details the experimental protocol and cost breakdown.

5.1 The Quadratic Extrapolation Scale

Setup. We use the calibrated construction from the lower bound of Thm. 1. Draw $X \sim \text{Unif}[1/4, 3/4]$ and, conditional on the prompt, candidates with $U \sim \text{Unif}[0,1]$, score $S = X + U/8$, and report $C = S$. Correctness is

$$\mathbb{P}(Y=1 \mid X,U) = S - \tfrac{1}{4}\mathbf{1}\{S \in [3/8, 3/4]\}\,\sigma_K(U), \qquad \sigma_K(u) = \operatorname{sign} \mathsf{U}_K(2u-1),$$

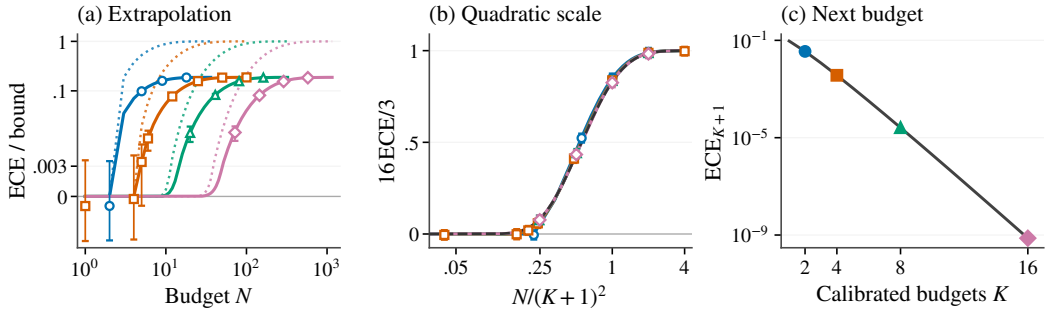


Figure 2: Exact calibration and its extrapolation scale. Colors/markers: ○ $K = 2$, □ $K = 4$, △ $K = 8$, ◇ $K = 16$. **(a)** Population ECE $3d_{K,N}/16$ (solid) and universal bound $d_{K,N}$ (dotted); the vertical axis is linear within ± 0.003 and logarithmic outside. **(b)** Rescaled ECE and limit Φ (black dashed). Markers in (a,b) show all 20 sampled settings; error bars show one standard deviation. **(c)** Next-budget ECE $\frac{3}{16}(K+1)4^{-K}$.

with candidates conditionally independent within a prompt. This model is fully calibrated at budgets $1, \ldots, K$ and has population $\mathrm{ECE}_N = |\Delta_N| = 3d_{K,N}/16$ at every $N > K$, so the population ECE is available in closed form. We set $K \in \{2, 4, 8, 16\}$ and evaluate budgets from $K+1$ through the quadratic scale. Each of the 20 sampled settings uses 100 repetitions of 10^4 prompts; the next-budget panel evaluates the exact formula.

Results. Fig. 2 illustrates both aspects of Cor. 1 in a single probability model. At the next budget the error is exponentially small, dropping by roughly a factor of four per additional calibrated budget (Fig. 2c). Yet the same model reaches constant-order error at the quadratic scale: exact calibration at 16 budgets already yields ECE of order 10^{-1} near budget 578 (Fig. 2a). Under the rescaled coordinate $N/(K+1)^2$, all four curves collapse onto the analytic limit Φ (Fig. 2b), consistent with $(K+1)^2$ being the natural extrapolation scale. Finite-sample estimates agree with the population values throughout. Even at moderate K, the next-budget error is negligible: $d_{8,9} \approx 1.4 \times 10^{-4}$, so calibration through budget 8 already provides a strong guarantee at budget 9. The gap between the construction's ECE and the upper bound is a fixed factor of $3/16$ at every budget $N > K$, showing that the bound cannot be improved by more than this universal constant.

5.2 Label Information and the Minimax Rate

Setup. We evaluate the minimax rate $\omega_{K,N}(m^{-1/2})$ from Thm. 8 and the dual estimator it prescribes. We use the same raw-score embedding with four generating models, all calibrated at budgets $1, \ldots, K$:

$$P_0 : \mathbb{P}(Y = 1 \mid X, U) = S, \qquad P_1 : \mathbb{P}(Y = 1 \mid X, U) = S + \tfrac{1}{8}\mathbf{1}\{S \in [3/8, 3/4]\}\, r^\star(U),$$

where $r^\star$ is a feasible near-optimal residual for the modulus problem satisfying $\|r^\star\|_\infty \le 1$, $\|r^\star\|_2 \le m^{-1/2}$, and orthogonality to $w_1, \ldots, w_K$. Two additional models P_2, P_3 use perturbations based on the Legendre polynomial and the Chebyshev sign function from Sec. 5.1; these are calibrated at the same budgets but are not constrained by the local L^2 bound. The local hard direction changes with (K, N, m); the fixed perturbations depend only on K. Each label comes from an independent prompt with the population rank observed, matching the observation model of App. D.1. We vary $K \in \{1, 4\}$, $N \in \{8, 32, 128\}$, and $m \in \{64, \ldots, 10^4\}$, and report the maximum MAE over the four models. The estimator uses the dual construction from Thm. 8; the weight function (p, ψ) depends only on (K, N, m) and does not use the evaluation labels. MAE is computed over 1000 repetitions per model and setting. We compute the modulus numerically; details appear in App. I.

Results. The modulus follows the $m^{-1/2}$ rate predicted by Thm. 8 across the entire grid (Fig. 3a; full grid in Fig. 6): away from saturation, multiplying m by 16 divides the modulus by 4. Labels therefore substantially reduce the ambiguity that remains from calibration constraints alone; at $K = 1$

and $N = 8$, for instance, the constraints leave $d_{1,8} = 0.99$ whereas 4096 labels reduce the achievable error to 0.028.

Additional calibration budgets are most informative near the calibrated prefix; their contribution diminishes beyond the quadratic scale (Fig. 3b): past $N \asymp (K+1)^2$, the constraints contribute little and the labels carry the information. At $m = 10^4$, increasing K from 1 to 4 reduces MAE by a factor of 8.23 at $N = 8$, but only 1.08 at $N = 128$. The dual estimator is consistent with this picture: its MAE follows the $m^{-1/2}$ rate and remains below the feasible dual bound across the full grid (Fig. 7). These results suggest that even a moderate number of labels can substantially narrow the uncertainty about calibration at new budgets, particularly when the extrapolation bound $d_{K,N}$ is large.

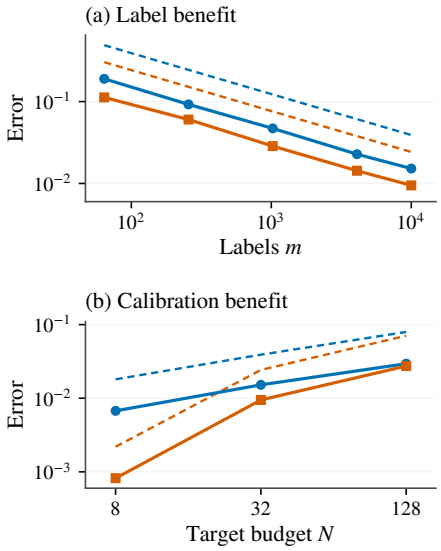


Figure 3: Labels improve residual estimation. • $K = 1$; ■ $K = 4$. Solid: four-model maximum MAE; dashed: same-color dual bound. **(a)** At $N = 32$, $16\times$ more labels gives about $4\times$ lower MAE. **(b)** At $m = 10^4$, calibration benefits diminish as N grows. Shading: approximate 95% Monte Carlo envelopes. Full grid and interval definitions: App. I.

5.3 Calibration Drift in Program Synthesis

Setup. We use eight configurations of program-synthesis generations from the CALIBRI release (Campos et al., 2026), spanning three models (Qwen3-Coder-30B (Qwen Team, 2025), GPT-OSS-20B (OpenAI, 2025), R1-Distill-Qwen-32B (DeepSeek-AI, 2025)) and three benchmarks (LiveCodeBench (Jain et al., 2025), McEval (Chai et al., 2024), MultiPL-E (Cassano et al., 2023)). LiveCodeBench contains recent competitive programming problems; McEval spans multiple programming languages; MultiPL-E provides translations of HumanEval across languages. Each prompt has ten candidate programs with token log probabilities and binary test outcomes. We compare a score-only report $g(S)$ with a rank-aware report $g(S, \widehat{U})$, both trained at $N = 1$. Selection always maximizes the original score, so both reports produce the same winners at every budget. In a second comparison, we split each pool into five selection candidates and five disjoint references. Both reports use the same labels and are selected by three-fold grouped cross-validation on $N = 1$ Brier score. Winner weights average over all size-N subsets; ECE pools the resulting bin residuals across test prompts in five equal-width bins. App. K gives the complete fitting, binning, and rank definitions.

Results. Score-only ECE increases with N in all eight configurations (Fig. 4), showing that fitting at $N = 1$ does not prevent calibration drift under stronger selection. Including the rank feature consistently lowers ECE at the largest budget; these improvements appear at budgets not used during fitting and affect only the reported probabilities, not selection accuracy. The score-only drift is consistent with the growing concentration of the selection weight $w_N(u) = Nu^{N-1}$ at high ranks: a verifier calibrated only at $N = 1$ reports increasingly miscalibrated probabilities as the selection budget grows. The independent $5 + 5$ design shows the same pattern in all eight configurations (Fig. 8): the benefit persists when rank references are disjoint from the selection candidates, so the improvement cannot be attributed to information leakage between selection and rank estimation. The drift pattern is qualitatively similar across the three model architectures, suggesting that rank dependence is a general property of the selection mechanism rather than of any particular generator. At $N = 10$, the ECE reductions range from 1.34 to 13.30 percentage points. LiveCodeBench/GPT-OSS has the largest reduction, from 20.34% to 7.05% ECE; even this configuration still shows substantial calibration error with the rank feature. The pointwise 95% paired bootstrap intervals include zero in seven of the eight configurations; Tab. 1 gives endpoint comparisons for both designs.

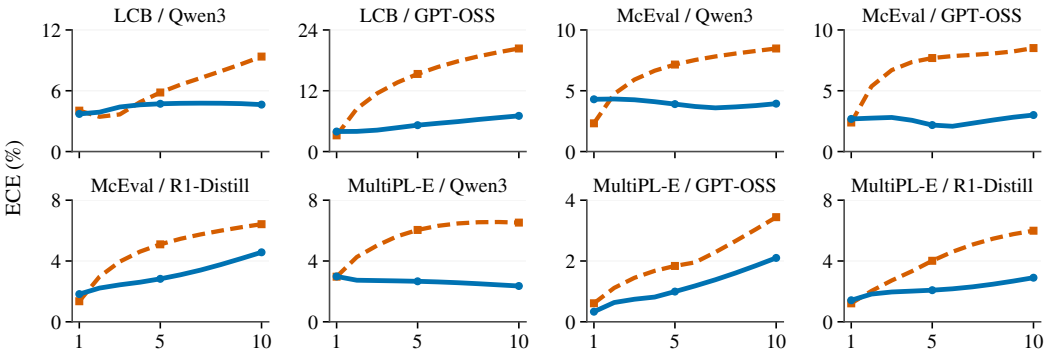


Figure 4: Rank-aware reports reduce high-budget calibration error across all eight configurations. Each panel shows ECE against selection budget N in the fixed ten-candidate pool (LCB: LiveCodeBench): ■ dashed squares denote score-only $g(S)$; ● solid circles denote score+rank $g(S, \widehat{U})$. Both reports are fitted only at $N = 1$ and select the same winners. Lines show point estimates; all budgets are shown. Each vertical axis starts at zero, with a panel-specific range. Paired uncertainty and the independent $5 + 5$ comparison appear in App. J.

5.4 A Rank Permutation Control

Setup. Thm. 4 identifies residual dependence on relative rank, conditional on reported confidence, as the obstruction to calibration across budgets. We test whether the improvement in Sec. 5.3 relies on each candidate being paired with its own rank, using Qwen3 on LiveCodeBench with the independent $5+5$ design. The fixed split contains 527 training prompts and 264 test prompts; all runs use the same cross-validation folds. The negative control retains the rank feature but permutes which rank is assigned to which candidate across prompts within score deciles, so that the marginal rank distribution is preserved but individual candidates no longer carry their own rank. Each permuted report is trained with the same model class and procedure. App. K.6 records the details.

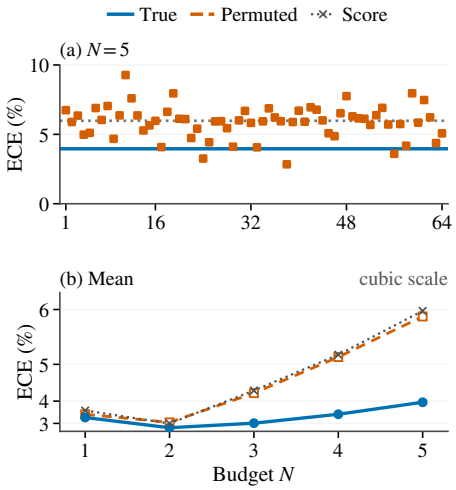


Figure 5: Rank permutation control on LiveCodeBench/Qwen3 with disjoint $5 + 5$ pools. **(a)** ECE at $N = 5$ for each of 64 displayed permutations. **(b)** Mean ECE over the same 64 runs. Blue: true rank; vermilion: permuted rank; dotted gray: score only.

Results. The mean permuted-rank curve closely follows the score-only baseline at all budgets (Fig. 5b). The gap between true-rank and permuted-rank ECE is negligible at $N = 1$ and grows with the selection budget as w_N concentrates on high-rank candidates, in line with Thm. 4. Individual runs vary considerably (Fig. 5a), but their average tracks the score-only curve. At $N = 5$, mean ECE is 5.89%, compared with 5.98% for score only and 3.96% for true rank.

6 Conclusion and Discussion

We derived a tight bound $d_{K,N}$ on calibration error under Best-of-N selection: the next-budget error is $(K + 1)4^{-K}$, and the bound reaches order one at scale $(K + 1)^2$. For finite samples, passive labels reduce the remaining ambiguity at minimax rate $\min 1, \sqrt{N/m}$. Experiments confirm both rates and show that rank-aware reports mitigate calibration drift in practice. Extending the framework to pool-dependent predictors (Kim et al., 2026) and prompt-dependent budgets (Snell et al., 2025; Huang et al., 2026) remains open.

AI USE STATEMENT

Generative AI tools were used for language polishing, routine coding and debugging during experiments, and assisting with proof checking. The research, analysis, interpretation of results, and preparation of the final content were done by the authors.

ETHICS STATEMENT

This work develops theoretical bounds and empirical diagnostics for calibration under Best-of-N selection. The experiments use publicly available datasets and pre-trained models. We are not aware of specific ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

Full proofs are given in Sec. A and C to F. Experimental details, including dataset splits, model configurations, and metrics, are described in Sec. K. Additional results appear in Sec. I and J.

REFERENCES

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017. https://doi.org/10.1214/17-STS611.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 3580–3602. PMLR, 2025. URL https://proceedings.mlr.press/v267/beirami25a.html.

Jarosław Błasiok and Preetum Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *International Conference on Learning Representations*, volume 2024, pages 1679–1707, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/06cf4bae7ccb6ea37b968a394edc2e33-Paper-Conference.pdf.

Viola Campos, Robin Kuschnereit, and Adrian Ulges. Multicalibration for LLM-based code generation. In *Proceedings of the 1st International Workshop on AI for Software Quality Evaluation - Judgment, Metrics, Benchmarks, and Beyond*, pages 61–71. ACM, 2026. https://doi.org/10.1145/3786175.3788347. URL https://doi.org/10.1145/3786175.3788347.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. MultiPL-E: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 49(7):3675–3691, 2023. https://doi.org/10.1109/TSE.2023.3267446. URL https://doi.org/10.1109/TSE.2023.3267446.

Linzheng Chai, Shukai Liu, Jian Yang, Yuwei Yin, Ke Jin, Jiaheng Liu, Tao Sun, Ge Zhang, Changyu Ren, Hongcheng Guo, Zekun Wang, Boyang Wang, Xianjie Wu, Bing Wang, Tongliang Li, Liqun Yang, Sufeng Duan, and Zhoujun Li. McEval: Massively multilingual code evaluation. arXiv preprint arXiv:2406.07436, 2024. URL https://arxiv.org/abs/2406.07436. Preprint version; also published at ICLR 2025.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025. URL https://arxiv.org/abs/2501.12948. Journal version published in Nature, 645:633–638, 2025.

Gianluca Detommaso, Martin Andres Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10624–10641. PMLR, 2024. URL https://proceedings.mlr.press/v235/detommaso24a.html.

Jinzong Dong, Zhaohui Jiang, and Bo Yang. Expectation consistency loss: Rethink confidence calibration under covariate shift. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.21552. arXiv:2605.21552.

Florian E. Dorner, Yatong Chen, André F. Cruz, and Fanny Yang. ROC-n-reroll: How verifier imperfection affects test-time scaling. arXiv preprint arXiv:2507.12399, 2026. URL https://arxiv.org/abs/2507.12399v3. Version 3; earlier version published at ICLR 2026.

Tamás Erdélyi. The "Full Müntz Theorem" revisited. *Constructive Approximation*, 21(3):319–335, 2005. https://doi.org/10.1007/s00365-004-0573-6.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. https://doi.org/10.1214/aos/1013203451.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3942–3952. PMLR, 2021. URL https://proceedings.mlr.press/v139/gupta21b.html.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Lunjia Hu, Arun Jambulapati, Kevin Tian, and Chutong Yang. Testing calibration in nearly-linear time. In *Advances in Neural Information Processing Systems*, volume 37, pages 116022–116059. Curran Associates, Inc., 2024. https://doi.org/10.52202/079017-3683. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d20e3c35bedb17a9f6f01fc434a30fa3-Paper-Conference.pdf.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is Best-of-N the best of them? Coverage, scaling, and optimality in inference-time alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 25075–25126. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25c.html.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. CaTS: Calibrated test-time scaling for efficient LLM reasoning. In *International Conference on Learning Representations*, pages 98256–98278, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9f9ecbf4062842df17ec3f4ea3ad7f54-Abstract-Conference.html.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, pages 58791–58831, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94074dd5a072d28ff75a76dabed43767-Abstract-Conference.html.

Joshua Kazdan, Rylan Schaeffer, Youssef Allouah, Colin Sullivan, Kyssen Yu, Noam Levi, and Sanmi Koyejo. Efficient prediction of Pass@k scaling in large language models. arXiv preprint arXiv:2510.05197, 2025. URL https://arxiv.org/abs/2510.05197.

Michael P. Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4):e2108097119, 2022. https://doi.org/10.1073/pnas.2108097119.

Yegon Kim, Seungyoo Lee, Chaeyun Jang, Hyungi Lee, and Juho Lee. Parallel test-time scaling with multi-sequence verifiers. arXiv preprint arXiv:2603.03417, 2026. URL https://arxiv.org/abs/2603.03417.

Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. T-Cal: An optimal test for the calibration of predictive models. *Journal of Machine Learning Research*, 24(335):1–72, 2023. URL https://www.jmlr.org/papers/v24/22-0320.html.

Muheng Li, Jian Qian, and Wenlong Mou. What should post-training optimize? A test-time scaling law perspective. arXiv preprint arXiv:2605.10716, 2026a. URL https://arxiv.org/abs/2605.10716.

Muheng Li, Jian Qian, and Wenlong Mou. Predicting and improving test-time scaling laws via reward tail-guided search. arXiv preprint arXiv:2602.01485, 2026b. URL https://arxiv.org/abs/2602.01485.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

H. Brendan McMahan and Omkar Muralidharan. On calibrated predictions for auction selection mechanisms. arXiv preprint arXiv:1211.3955, 2012. URL https://arxiv.org/abs/1211.3955.

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

Yuji Nakatsukasa and Alex Townsend. Error localization of best L_1 polynomial approximants. *SIAM Journal on Numerical Analysis*, 59(1):314–333, 2021. https://doi.org/10.1137/19M1242860.

Ansong Ni, Pengcheng Yin, Yilun Zhao, Martin Riddell, Troy Feng, Rui Shen, Stephen Yin, Ye Liu, Semih Yavuz, Caiming Xiong, Shafiq Joty, Yingbo Zhou, Dragomir Radev, and Arman Cohan. L2CEval: Evaluating language-to-code generation capabilities of large language models. *Transactions of the Association for Computational Linguistics*, 12:1311–1329, 2024. https://doi.org/10.1162/tacl_a_00705. URL https://aclanthology.org/2024.tacl-1.72/.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. arXiv preprint arXiv:2508.10925, 2025. URL https://arxiv.org/abs/2508.10925.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1):2901–2907, 2015. https://doi.org/10.1609/aaai.v29i1.9602.

Young-Jin Park, Kristjan Greenewald, Kaveh Alimohammadi, Hao Wang, and Navid Azizan. Know what you don't know: Uncertainty calibration of process reward models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 38852–38895. Curran Associates, Inc., 2025. https://doi.org/10.52202/085713-1297. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/37664246a1e07e212ddacea6e5a523f2-Paper-Conference.pdf.

Qwen Team. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating bias in calibration error estimation. In *Proceedings of The 25th International Conference on Artificial*

Intelligence and Statistics, volume 151 of *Proceedings of Machine Learning Research*, pages 4036–4054. PMLR, 2022. URL https://proceedings.mlr.press/v151/roelofs22a.html.

Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. Can a calibration metric be both testable and actionable? In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4937–4972. PMLR, 2025. URL https://proceedings.mlr.press/v291/rossellini25a.html.

Ludger Rüschendorf. On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927, 2009. https://doi.org/10.1016/j.jspi.2009.05.030.

Divya Shyamal, Marta Knežević, Lan Tran, Chanakya Ekbote, Vijay Lingam, and Paul Pu Liang. SCATR: Simple calibrated test-time ranking. arXiv preprint arXiv:2604.16535, 2026. URL https://arxiv.org/abs/2604.16535.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, pages 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Claudio Spiess, David Gros, Kunal Suresh Pai, Michael Pradel, Md Rafiqul Islam Rabin, Amin Alipour, Susmit Jha, Prem Devanbu, and Toufique Ahmed. Calibration and correctness of language models for code. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*, pages 540–552. IEEE, 2025. https://doi.org/10.1109/ICSE55347.2025.00040. URL https://doi.org/10.1109/ICSE55347.2025.00040.

Jiayun Wu, Jiashuo Liu, Peng Cui, and Zhiwei Steven Wu. Bridging multicalibration and out-of-distribution generalization beyond covariate shift. In *Advances in Neural Information Processing Systems*, volume 37, pages 73036–73078. Curran Associates, Inc., 2024. https://doi.org/10.52202/079017-2325. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/859b6564b04959833fdf52ae6f726f84-Abstract-Conference.html.

Tijana Zrnic and William Fithian. A flexible defense against the winner's curse. *The Annals of Statistics*, 53(6):2516–2535, 2025. https://doi.org/10.1214/25-AOS2553.

A Proofs for the Selected Law and Calibration Characterization

This section proves the representation in terms of relative rank and the calibration condition stated in Sec. 2.

Notation.

Symbol	Meaning
P, P_N	Ordinary and budget-N selected laws
$S, C, Y, R = Y - C$	Ranking score, report, correctness, residual
U, V	Relative rank within a prompt and independent tie variable
$w_N(u) = Nu^{N-1}$	Selection density ratio
$\mathcal{V}_K,\ M = K + 1$	Calibration weight span and scale parameter
$d_{K,N},\ \omega_{K,N}(t)$	Extrapolation error and minimax error scale
$D_{\mathcal{B},N}(B_\infty, \epsilon)$	Largest absolute target moment under the given bounds
$v_N = N^2/(2N-1)$	Second moment of the selection weight
$m,\ n_{\text{pool}},\ k$	Independent prompts, pool size, reference count
$R_{\text{ref}},\ \widehat{U}_k = R_{\text{ref}}/k$	Reference rank count and estimated relative rank
$\eta(s,u),\ g(s,u)$	Oracle conditional correctness and fitted predictor
$\Delta_U,\ \tau_\delta$	Width of a rank bin and log confidence factor

A.1 The Randomized Probability Integral Transform

Fix a prompt x and $t \in (0,1)$. Let s be a generalized t-quantile of its score law. If t lies inside a jump interval $[F_x(s^-), F_x(s)]$, scores strictly below s contribute mass $F_x(s^-)$ to $P(U \le t \mid x)$. A fraction $(t - F_x(s^-))/(F_x(s) - F_x(s^-))$ of the point mass at s contributes the remaining mass. Thus $P(U \le t \mid x) = t$. The argument using a generalized inverse gives the same identity on the continuous part. Hence $U \mid X$ is uniform. Across different score levels, the associated intervals of CDF values are ordered. At a score with positive conditional probability, the relative rank increases with V. The lexicographic ordering of (S, V) and U therefore give the same maximizer almost surely.

A.2 Proof of the Identity for the Selected Law

Proof. For bounded f, conditional independence gives $P(I_N = 1 \mid X, Z_1, V_1) = U_1^{N-1}$. By exchangeability and iterated conditioning,

$$\begin{aligned}\mathbb{E}f(X, Z_{I_N}, V_{I_N}) &= N\mathbb{E}\big[f(X, Z_1, V_1)\mathbf{1}\{I_N = 1\}\big] \\ &= N\mathbb{E}\big[f(X, Z_1, V_1)U_1^{N-1}\big],\end{aligned}$$

which is Eq. (5). Monotone convergence extends the identity to nonnegative functions, and decomposition into positive and negative parts gives the integrable case. Finally, $\mathbb{E}[w_N(U) \mid X] = \int_0^1 Nu^{N-1}\,du = 1$, so the prompt marginal is unchanged. □

A.3 Weights at Point Masses of the Score Distribution

Corollary 2 (Weights after averaging over the tie variable)**.** *Set $u_- = F_X(S^-)$ and $u_+ = F_X(S)$. At a point mass, where $u_+ > u_-$, the weight for a function independent of V is*

$$\overline{w}_N(X, S) = \frac{u_+^N - u_-^N}{u_+ - u_-}. \tag{17}$$

Away from point masses it is $NF_X(S)^{N-1}$.

Proof. Conditional on (X, S) at a point mass, U is uniform on $[u_-, u_+]$, and is independent of the remaining candidate variables given (X, S) because V was independent. Average Nu^{N-1} over this interval. The case without a point mass has $U = F_X(S)$. □

A.4 Mean Calibration across Budgets

Theorem 3 (Mean calibration at every budget)**.** *Under the setting of Sec. 2, the following are equivalent:*

(i) $\mathbb{E}_N[Y - C] = 0$ *for every integer* $N \geq 1$*;*

(ii) $\mathbb{E}[Y - C \mid U] = 0$ *almost surely.*

More generally, for any fixed bounded measurable h*, the equalities* $\theta_{N,h} = 0$ *for all* $N \geq 1$ *hold if and only if* $\mathbb{E}[(Y - C)h(C) \mid U] = 0$ *almost surely.*

Proof. For $r_h(u) = \mathbb{E}[Rh(C) \mid U = u]$, Lem. 1 gives

$$\theta_{N,h} = N \int_0^1 u^{N-1} r_h(u)\, du. \tag{18}$$

Thus $r_h = 0$ almost everywhere implies calibration at every budget. Conversely, if all these moments are zero, the finite signed measure $r_h(u)\, du$ annihilates every polynomial. By uniform polynomial approximation on $[0, 1]$, it annihilates every continuous function and is therefore zero. This proves the equivalence for bounded h; taking $h = 1$ gives mean calibration. □

A.5 Full Calibration across Budgets

Theorem 4 (Calibration at every budget)**.** *In the setting of Sec. 2,* C *is fully calibrated under every* P_N*,* $N \geq 1$*, if and only if*

$$\mathbb{E}[Y - C \mid C, U] = 0 \quad P\text{-almost surely}. \tag{19}$$

For binary Y*, this is equivalent to ordinary calibration together with* $Y \perp U \mid C$*.*

Proof. If Eq. (19) holds, then for every bounded measurable h and every N,

$$\theta_{N,h} = \mathbb{E}\big[w_N(U)h(C)\mathbb{E}[R \mid C, U]\big] = 0,$$

which is the testing characterization of full calibration.

Conversely, full calibration at every budget gives $\mathbb{E}[w_N(U)R\mathbf{1}\{C \in B\}] = 0$ for every Borel set $B \subseteq [0, 1]$ and every N. Applying Thm. 3 with $h = \mathbf{1}_B$ yields $\mathbb{E}[R\mathbf{1}\{C \in B\}\mathbf{1}\{U \in A\}] = 0$ for every Borel set A. Hence the finite signed measure $\mu_R(D) = \mathbb{E}[R\mathbf{1}\{(C, U) \in D\}]$ is zero on measurable rectangles and, by uniqueness, on their generated sigma-field. This is equivalent to Eq. (19) and avoids an uncountable intersection of conditional statements that hold almost surely.

For binary Y, the conditional mean determines the conditional distribution. Thus Eq. (19) and the tower property give $P(Y = 1 \mid C, U) = P(Y = 1 \mid C) = C$, which is ordinary calibration and $Y \perp U \mid C$. The converse follows by the same identity for the conditional mean. □

B Supporting Constructions

The constructions in this section connect bounded residual functions to valid Bernoulli models, supporting the sharpness claim in Thm. 1 and the lower bound in Thm. 8. Whenever an L^∞ residual is used in a probability kernel, we take a Borel representative that is zero on any exceptional null set, so its essential bound holds pointwise without changing moments or norms.

B.1 The Example with Two Prompt Types

Example 1 (Two prompt types)**.** Let $P(X = A) = P(X = B) = 1/2$ and set $C = S$. Scores take values $1/4$ and $3/4$, with conditional laws

	$P(S = \frac{3}{4} \mid X)$	$\mathbb{E}[Y \mid X, S = \frac{3}{4}]$	$\mathbb{E}[Y \mid X, S = \frac{1}{4}]$
A	0.9	0.8	0.7
B	0.1	0.3	0.2

At either score, averaging over prompts gives $\mathbb{E}[Y \mid S] = S$. At $N = 1$, both accuracy and mean confidence are 0.50. As $N \to \infty$, every prompt produces a winner at the high score with probability tending to one. Mean confidence approaches 0.75, while accuracy approaches $(0.8 + 0.3)/2 = 0.55$.

The pooled correctness probabilities are $0.9(0.8) + 0.1(0.3) = 3/4$ at the high score and $0.1(0.7) + 0.9(0.2) = 1/4$ at the low score. Selection returns the low score with probabilities 10^{-N} and $(9/10)^N$ on the two prompt types. With $q_N^{\text{low}} = 10^{-N} + (9/10)^N$, averaging over prompts gives

$$\mathbb{E}_N C = \tfrac{3}{4} - \tfrac{1}{4} q_N^{\text{low}}, \tag{20}$$

$$\mathbb{E}_N Y = \tfrac{11}{20} - \tfrac{1}{20} q_N^{\text{low}}, \tag{21}$$

$$\Delta_N = \tfrac{1}{5}(q_N^{\text{low}} - 1). \tag{22}$$

B.2 Calibrated Raw Scores

Lemma 2 (Realizing bounded residual functions). *Let $X \sim \mathrm{Unif}[1/4, 3/4]$, let $U \sim \mathrm{Unif}[0, 1]$ be independent of X, and define $S = X + U/8$. Put $A = [3/8, 3/4]$. For a measurable $r : [0, 1] \to [-1, 1]$ and $0 < \lambda \le 1/4$, define*

$$P(Y = 1 \mid X, U) = S + \lambda \mathbf{1}\{S \in A\} r(U), \qquad C = S. \tag{23}$$

This is a valid conditionally i.i.d. candidate model with continuous score CDFs and relative rank U within each prompt. If $\int_0^1 r(u)\, du = 0$, the score is calibrated under P. For each N,

$$\Delta_N = \frac{3\lambda}{4} \int_0^1 w_N(u) r(u)\, du. \tag{24}$$

Moreover, full calibration under P_N holds if and only if $\int_0^1 w_N r = 0$.

In this model, the selected mass of A is $3/4$, and

$$\mathrm{ECE}_N(S) = |\Delta_N| = \frac{3\lambda}{4} \left| \int_0^1 w_N r \right|. \tag{25}$$

Proof. Conditional on the shared prompt X, generate independent $U_i \sim \mathrm{Unif}[0, 1]$, set $S_i = X + U_i/8$, and draw independent Bernoulli labels with probabilities Eq. (23). These probabilities lie in $[0, 1]$: the perturbation is at most $1/4$ on $A = [3/8, 3/4]$, and outside A the probability is $S \in [1/4, 7/8]$. The conditional score law is uniform on $[X, X + 1/8]$, with CDF value $F_X(S) = U$.

For $(s, u) \in A \times [0, 1]$, the value $x = s - u/8$ belongs to $[1/4, 3/4]$, so the joint density of (S, U) is constant on this rectangle. Consequently, $U \mid S = s$ is uniform for almost every $s \in A$ and $P(S \in A \mid U = u) = (3/8)/(1/2) = 3/4$. Using Lem. 1, we obtain

$$P_N(S \in A) = \tfrac{3}{4},$$

$$\mathbb{E}_N[Y - S \mid S] = \lambda \mathbf{1}\{S \in A\} \int_0^1 w_N(u) r(u)\, du.$$

The second identity follows by weighting the conditional uniform law on A; outside A the residual is zero. It proves the equivalence for full calibration and, upon taking signed or absolute expectations, Eqs. (24) and (25). At $N = 1$, the condition that the integral is zero gives ordinary calibration. □

B.3 A Smooth Legendre Construction

Theorem 5 (Calibration at finitely many budgets and later failure). *For every integer $K \ge 1$, there is a model with continuous conditional score distributions such that the raw score $C = S$ is fully calibrated at every budget $N \le K$, but not at $N = K + 1$. The same model has a nonzero limiting mean residual as $N \to \infty$. More precisely, in Lem. 2, take*

$$r(u) = \ell_K(u) := \frac{1}{K!} \frac{d^K}{du^K} \{u^K (u - 1)^K\}, \tag{26}$$

the shifted degree-K Legendre polynomial. Then, for $N > K$,

$$\Delta_N = \frac{3\lambda}{4} N \frac{\{(N-1)!\}^2}{(N-1-K)!(N+K)!}, \tag{27}$$

and $\Delta_N = 0$ for $N \le K$. In particular, $\Delta_{K+1} = \frac{3\lambda}{4}(K+1)(K!)^2/(2K+1)! > 0$ and $\Delta_N \to 3\lambda/4$.

Proof. In the Rodrigues formula above, the polynomial inside braces and its first $K-1$ derivatives vanish at both endpoints. Integrating by parts K times shows that $\int_0^1 u^{N-1}\ell_K(u)\,du = 0$ for $N-1 < K$. For $N > K$, the same calculation gives

$$\begin{aligned}\int_0^1 u^{N-1}\ell_K(u)\,du &= \frac{(N-1)!}{K!(N-1-K)!}\int_0^1 u^{N-1}(1-u)^K\,du \\ &= \frac{\{(N-1)!\}^2}{(N-1-K)!(N+K)!},\end{aligned}$$

using the integer beta integral. The standard Legendre bound $|\ell_K(u)| \le 1$ on $[0,1]$ makes the construction valid, and $\ell_K(1) = 1$.

Apply Lem. 2 to obtain full calibration at all $N \le K$ and formula Eq. (27). Finally, under density w_N, $U \to 1$ in probability because $P_N(U \le 1-\epsilon) = (1-\epsilon)^N$. Boundedness and continuity of ℓ_K imply $\int w_N \ell_K \to \ell_K(1) = 1$. □

The first failure can be numerically small when K is large. This smooth construction establishes nonidentification, while the Chebyshev construction solves the extremal problem for bounded residuals. Thm. 1 and 7 give the sharper magnitude and range conclusions. Negating either construction produces overconfidence instead of underconfidence.

B.4 Sparse Budget Sets

Theorem 6 (Determining sets of integer budgets)**.** *Let $\mathcal{B} \subseteq \mathbb{N}$ be an infinite set of budgets containing 1. The following are equivalent:*

- *(i) For every model in Sec. 2, mean calibration at every budget in $\mathcal{B}$ implies mean calibration at all integer budgets.*
- *(ii) For every such model, full calibration at every budget in $\mathcal{B}$ implies full calibration at all integer budgets.*
- *(iii)* $\displaystyle\sum_{N\in\mathcal{B},\,N>1} \frac{1}{N-1} = \infty.$

If the series converges, failure of both implications can occur with continuous conditional scores and a raw score $C = S$ with pooled calibration.

Proof. We invoke the classical full Müntz theorem in $L^1[0,1]$ (Erdélyi, 2005, Theorem 1.1). For distinct nonnegative integer exponents λ, the span of u^λ is dense in $L^1[0,1]$ if and only if the associated reciprocal series diverges; the general criterion is $\sum(\lambda+1)/\{(\lambda+1)^2+1\} = \infty$. For $\lambda = N-1$, this is equivalent to (iii). Inclusion of the constant exponent causes no change to divergence.

If (iii) holds and a bounded residual function annihilates all u^{N-1} for $N \in \mathcal{B}$, continuity of its integration functional on L^1 implies that it annihilates all of L^1. It is therefore zero almost everywhere. This proves (i). Applying the same argument to the signed residual measures indexed by confidence events, as in Thm. 4, proves (ii).

If (iii) fails, the closed span is a proper closed subspace of $L^1[0,1]$. By Hahn–Banach separation and the identification $(L^1)^* = L^\infty$, there exists a nonzero bounded measurable r annihilating this span. Normalize it so $|r| \le 1$. Since $1 \in \mathcal{B}$, $\int r = 0$. Some integer moment must be nonzero; otherwise ordinary polynomial completeness would force $r = 0$. Realize this r using Lem. 2. The raw score is fully calibrated for every budget in $\mathcal{B}$, but has a nonzero mean residual at an untested integer budget. Both implications fail. □

C Proofs for the Extrapolation Bound

C.1 Calibrated Realization for Thm. 1

The proof of Thm. 1 in the main text establishes the upper bound $\mathrm{ECE}_N(C) \le d_{K,N}$ via the approximation identity Eq. (10). The following construction completes the proof of the lower bound Eq. (8). Recall that $\sigma_K(u) = \operatorname{sign} \mathsf{U}_K(2u-1)$ is the sign function from that proof.

Draw $X \sim \mathrm{Unif}[1/4, 3/4]$ and, conditional on X, draw independent candidates with $U \sim \mathrm{Unif}[0,1]$, $S = X + U/8$, $C = S$, and

$$P(Y = 1 \mid X, U) = S + \tfrac{1}{4}\mathbf{1}\{S \in A\}\sigma_K(U), \qquad A = [3/8, 3/4].$$

On A, the perturbation has magnitude at most $1/4$; outside A, the success probability is S. These probabilities therefore lie in $[0,1]$, and the model has continuous conditional scores with relative rank $U = F_X(S)$.

The joint density of (S,U) is constant on $A \times [0,1]$, so $U \mid S = s$ is uniform for almost every $s \in A$, while $P(S \in A \mid U = u) = 3/4$. Under the selected law,

$$P_N(S \in A) = \tfrac{3}{4},$$
$$\mathbb{E}_N[Y - S \mid S] = \tfrac{1}{4}\mathbf{1}\{S \in A\}\int_0^1 w_N \sigma_K = \tfrac{1}{4}\mathbf{1}\{S \in A\} d_{K,N}.$$

The conditional residual is zero at every $N \le K$, and for $N > K$ its signed and absolute means both equal $3d_{K,N}/16$. This proves Eq. (8); the construction depends only on K and therefore works at every future budget simultaneously.

C.2 Chebyshev Interpolation and Scalar Duality

We supply the Chebyshev calculations used in the proof of Thm. 1, then record scalar duality and the identity for the next budget.

Annihilation of the polynomial span. Put $M = K+1$ and recall that $\mathsf{U}_K(\cos\theta) = \sin(M\theta)/\sin\theta$. Its shifted roots are the nodes u_j in Eq. (9). Under $u = (1+\cos\theta)/2$, the factor $p((1+\cos\theta)/2)\sin\theta$ for $p \in \mathcal{V}_K$ lies in the span of $\sin(j\theta)$, $1 \le j \le K$. Integration over the M sign intervals gives

$$j\int_0^\pi \sin(j\theta)\operatorname{sign}\sin(M\theta)\,d\theta = \mathrm{Re}\sum_{\ell=0}^{2M-1} e^{i\pi\ell(1+j/M)} = 0.$$

The geometric sum is zero because its ratio is a nontrivial $2M$th root of unity. Consequently, $\sigma_K(u) = \operatorname{sign}\mathsf{U}_K(2u-1)$ satisfies $\int_0^1 p\sigma_K = 0$ for every $p \in \mathcal{V}_K$.

Sign of the interpolation remainder. For $N > K$, let $p_{K,N} \in \mathcal{V}_K$ interpolate w_N at $u_1, \dots, u_K$. Away from these nodes, the interpolation remainder is

$$w_N(u) - p_{K,N}(u) = \frac{w_N^{(K)}(\xi_u)}{K!}\prod_{j=1}^K (u - u_j), \qquad \xi_u \in (0,1).$$

Since $w_N^{(K)} > 0$ on $(0,1)$ and U_K has positive leading coefficient, the remainder has sign σ_K. Together with annihilation, this proves the optimality of $p_{K,N}$ as shown in the main text.

Evaluation of the optimal error. The sign function σ_K is positive just below one and alternates at the u_j. With $u_0 = 1$ and $u_{K+1} = 0$, integration gives

$$\int_0^1 w_N(u)\sigma_K(u)\,du = \sum_{j=0}^K (-1)^j (u_j^N - u_{j+1}^N) = 1 + 2\sum_{j=1}^K (-1)^j u_j^N = d_{K,N}$$

for $N > K$. For $N \le K$, the same integral is zero by annihilation.

Scalar duality. The optimality certificate in the main text also gives the dual for bounded residuals:

$$d_{K,N} = \sup_{\substack{\|r\|_\infty \le 1 \\ \int pr = 0 \ (p \in \mathcal{V}_K)}} \left| \int w_N r \right|. \tag{28}$$

Indeed, every feasible r obeys $|\int w_N r| = |\int (w_N - p) r| \le \|w_N - p\|_1$ for each $p \in \mathcal{V}_K$. For $N > K$, the sign function σ_K constructed in that proof attains the value $d_{K,N}$; for $N \le K$, the target weight itself belongs to $\mathcal{V}_K$, so the value is zero. This is the interpolation argument using Chebyshev polynomials of the second kind applied to the Best-of-N weights (Nakatsukasa and Townsend, 2021).

The identity for the next budget. At the next budget, the interpolation remainder is $(K+1)4^{-K}\mathsf{U}_K(2u-1)$ by its leading coefficient and roots. Hence, with $M = K+1$ and $u = (1+\cos\theta)/2$,

$$d_{K,K+1} = (K+1)4^{-K}\int_0^1 |\mathsf{U}_K(2u-1)|\,du = \frac{K+1}{2\,4^K}\int_0^\pi |\sin(M\theta)|\,d\theta = \frac{K+1}{4^K},$$

as asserted in Cor. 1. For fixed K, each $u_j < 1$ in Eq. (9), so $d_{K,N} \to 1$ as $N \to \infty$.

Remark 1 (General predictors and the constant in the upper bound)**.** If C need not equal S, take one prompt, $S = U \sim \mathrm{Unif}[0,1]$, $C \equiv 1/2$, and $Y \mid U \sim \mathrm{Bern}((1+\sigma_K(U))/2)$. The same model is fully calibrated at $1,\dots,K$ and satisfies $\mathrm{ECE}_N(C) = d_{K,N}/2$ for every $N > K$. This supplements the construction with raw scores in Eq. (8).

For each fully calibrated model at $1,\dots,K$ and $N > K$, the upper bound is strict. Choose $h(C) = \operatorname{sign}\mathbb{E}_N[Y - C \mid C]$, which attains the ECE supremum. Equality in the residual bound would require $|\mathbb{E}[(Y-C)h(C) \mid U]| = 1$ almost everywhere, since the interpolation error is nonzero off its finitely many roots. This forces $|Y - C| = 1$, contradicting ordinary full calibration, which requires $Y = C$ on $C \in \{0,1\}$. Nevertheless no constant smaller than one improves Eq. (7) uniformly over all K, N and general predictors. For $K = 1$, take $S = U$, $C \equiv a$, and $Y = \mathbf{1}\{U > 1-a\}$. Then ordinary full calibration holds and $\mathrm{ECE}_N(C) = 1 - (1-a)^N - a$. With $a = (\log N)/N$, its ratio to $d_{1,N} = 1 - 2^{1-N}$ tends to one.

C.3 The Quadratic Scale

Theorem 7 (Transition on the scale K^2)**.** *Set $M = K+1$. For every $N \ge 1$,*

$$d_{K,N} = \frac{2M}{4^N} \sum_{\substack{q\in\mathbb{Z} \\ |(2q+1)M| \le N}} \binom{2N}{N + (2q+1)M}. \tag{29}$$

In particular,

$$d_{K,N} \le \frac{4M}{\sqrt{\pi N}} \frac{\exp\{-M^2/(N+1/2)\}}{1 - \exp\{-8M^2/(N+1/2)\}}, \tag{30}$$

$$d_{K,N} \ge 1 - 2\exp\{-\pi^2 N/(4M^2)\}. \tag{31}$$

For sequences with $K \to \infty$, $d_{K,N} \to 0$ if $N/M^2 \to 0$, and $d_{K,N} \to 1$ if $N/M^2 \to \infty$. If $N/M^2 \to c \in (0,\infty)$, then

$$d_{K,N} \longrightarrow \Phi(c) := 1 + 2\sum_{j=1}^{\infty} (-1)^j e^{-\pi^2 c j^2/4}. \tag{32}$$

Proof. Positive representation. Using $u_j = \cos^2(j\pi/(2M))$ and pairing the indices j and $2M-j$ in Eq. (9), we obtain the following identity for every $N \ge 1$ (including the zero value for $N \le K$ established by orthogonality):

$$\begin{aligned} d_{K,N} &= \sum_{j=0}^{2M-1} (-1)^j \cos^{2N}(j\pi/(2M)) \\ &= 4^{-N} \sum_{r=0}^{2N} \binom{2N}{r} \sum_{j=0}^{2M-1} \exp\{i\pi j(1 + (r-N)/M)\}. \end{aligned}$$

The $j = M$ term in the first line is zero. The inner geometric sum in the second line equals $2M$ when $r - N$ is an odd multiple of M and is zero otherwise, proving Eq. (29).

Bounds for finite budgets. Define $p_N(s) = 4^{-N}\binom{2N}{N+s}$ for $|s| \le N$. The Gaussian bound $\cos x \le e^{-x^2/2}$ gives $p_N(0) = \frac{2}{\pi}\int_0^{\pi/2} \cos^{2N} x\, dx \le 1/\sqrt{\pi N}$. For $1 \le s \le N$, set $x_j = (j-1/2)/(N+1/2)$. Then

$$\log \frac{p_N(s)}{p_N(0)} = \sum_{j=1}^{s} \log \frac{1-x_j}{1+x_j} \le -2\sum_{j=1}^{s} x_j = -\frac{s^2}{N+1/2}.$$

By symmetry, Eq. (29), and $(2q+1)^2 \ge 1+8q$,

$$\begin{aligned} d_{K,N} &\le \frac{4M}{\sqrt{\pi N}} \sum_{q=0}^{\infty} e^{-(2q+1)^2 M^2/(N+1/2)} \\ &\le \frac{4M}{\sqrt{\pi N}} \frac{e^{-M^2/(N+1/2)}}{1-e^{-8M^2/(N+1/2)}}, \end{aligned}$$

which proves Eq. (30). For the complementary bound, the alternating sum in Eq. (9) and the same cosine inequality yield

$$d_{K,N} \ge 1 - 2u_1^N = 1 - 2\cos^{2N}(\pi/(2M)) \ge 1 - 2e^{-\pi^2 N/(4M^2)}.$$

The cosine inequality follows by integrating $(\log\cos x)' = -\tan x \le -x$ on $[0, \pi/2)$. Together with $d_{K,N} \le \|w_N\|_1 = 1$, these bounds prove the limits for $N/M^2 \to 0$ and $N/M^2 \to \infty$; in the former case, $N + 1/2 \le 3N/2$ ensures that exponential decay dominates the prefactor.

Transition profile. If $N/M^2 \to c > 0$, the expansion of $\log\cos x$ at zero gives $u_j^N \to e^{-\pi^2 c j^2/4}$ for each fixed j. Extend the summands by zero for $j \ge M$. The cosine bound dominates them eventually by the summable sequence $e^{-\pi^2 c j^2/8}$, so dominated convergence in Eq. (9) proves Eq. (12). The bound for finite budgets above also supplies uniform control in the regime $N/M^2 \to 0$. □

Remark 2 (Theta representation of the transition)**.** The profile is the Jacobi theta function $\Phi(c) = \theta_4(0, e^{-\pi^2 c/4})$. Poisson summation applied to the Gaussian in Eq. (32), equivalently the classical theta transformation (NIST DLMF, §20.7(viii)), gives

$$\Phi(c) = \frac{4}{\sqrt{\pi c}} \sum_{q=0}^{\infty} e^{-(2q+1)^2/c} = \frac{4}{\sqrt{\pi c}} e^{-1/c}\{1 + O(e^{-8/c})\} \quad (c \downarrow 0). \tag{33}$$

This describes the limiting profile; Eq. (30) separately controls finite budgets and joint limits.

For a tolerance of 0.05, solving $\Phi(c) = 0.05$ gives $c \approx 0.2188496602$. Evaluating Eq. (29) gives the following largest budgets satisfying $d_{K,N} \le 0.05$:

Calibrated budgets K	8	16	32	128
Largest N within tolerance	18	63	238	3642

These are thresholds of the universal upper bound under full calibration at the first K budgets, not the critical budgets for the largest achievable ECE in any probability model. Under mean calibration alone they bound only $|\Delta_N|$.

C.4 Extrapolation under Moment Bounds

Proof of Thm. 2. For every feasible r and every $\alpha \in \mathbb{R}^{\mathcal{B}}$,

$$\left|\int w_N r\right| \le \sum_{n\in\mathcal{B}} |\alpha_n|\epsilon_n + B_\infty \left\| w_N - \sum_{n\in\mathcal{B}} \alpha_n w_n \right\|_1,$$

which proves the upper bound. To obtain equality, consider the continuous seminorm on real $L^1[0,1]$

$$\mathcal{J}(f) = \inf_{\alpha\in\mathbb{R}^{\mathcal{B}}} \left\{ \sum_{n\in\mathcal{B}} \epsilon_n|\alpha_n| + B_\infty \left\| f - \sum_{n\in\mathcal{B}} \alpha_n w_n \right\|_1 \right\}.$$

It satisfies $\mathcal{J}(f) \le B_\infty \|f\|_1$ and $\mathcal{J}(w_n) \le \epsilon_n$ for $n \in \mathcal{B}$. By Hahn–Banach, the functional $\ell(aw_N) = a\mathcal{J}(w_N)$ extends to L^1 with $|\ell(f)| \le \mathcal{J}(f)$. The bound by the L^1 norm and L^1–L^∞ duality give a representation $\ell(f) = \int fr$ with $\|r\|_\infty \le B_\infty$. In particular,

$$\left|\int w_n r\right| \le \mathcal{J}(w_n) \le \epsilon_n \quad (n \in \mathcal{B}), \qquad \int w_N r = \mathcal{J}(w_N).$$

Thus r is feasible and gives equality. This argument includes zero moment bounds. With $\mathcal{B} = \{1, \dots, K\}$, $B_\infty = 1$, and $\epsilon = 0$, the right-hand side is the approximation error in Eq. (10). □

The following corollary quantifies the budget range under uniform positive moment bounds.

Corollary 3 (Budget range with a fixed noise floor)**.** *Let $\mathcal{B} = \{1, \dots, K\}$ and $0 < \epsilon < \tau < 1$. For uniform moment bounds,*

$$D_{\mathcal{B},N}(1, \epsilon\mathbf{1}) \ge 1 - (1-\epsilon)^{N/K}. \tag{34}$$

Consequently $D_{\mathcal{B},N}(1, \epsilon\mathbf{1}) \le \tau$ requires

$$N \le K \frac{\log(1-\tau)}{\log(1-\epsilon)}. \tag{35}$$

Every $N \le K$ has $D_{\mathcal{B},N}(1, \epsilon\mathbf{1}) \le \epsilon$. For $N > K$, setting $\rho = N/K$ also gives the stable upper bound

$$D_{\mathcal{B},N}(1, \epsilon\mathbf{1}) \le \epsilon + 2(\rho - 1)\rho^{-\rho/(\rho-1)}. \tag{36}$$

Thus the largest budget satisfying the tolerance has order $\Theta_{\epsilon,\tau}(K)$ for fixed ϵ, τ.

Proof. Put $u_* = (1-\epsilon)^{1/K}$ and $r(u) = \mathbf{1}\{u > u_*\}$. For every $n \le K$, $\int w_n r = 1 - u_*^n \le \epsilon$, so r is feasible and its target moment proves Eq. (34). Rearranging gives Eq. (35); a target among the known budgets is already bounded by ϵ.

For $N > K$, use only the Kth weight in Eq. (13). This gives $D \le \epsilon + \|w_N - w_K\|_1$. The densities cross at $u_0 = (K/N)^{1/(N-K)}$; integration on either side gives $\|w_N - w_K\|_1 = 2(u_0^K - u_0^N) = 2(\rho-1)\rho^{-\rho/(\rho-1)}$. □

The constants depend on the fixed noise floor; the conclusion is not uniform as $\epsilon \to 0$. This corollary concerns bounded residuals under noisy moment constraints only. The construction used for the lower bound need not satisfy ordinary or full calibration, so it is not a lower bound in the smaller model class of Thm. 8.

C.5 Sensitivity at the Next Budget

For the next budget $N = K + 1$, the optimal approximant under zero moment bounds in Thm. 1 is

$$p_K(u) = (K+1)\{u^K - 4^{-K}\mathsf{U}_K(2u-1)\} = \sum_{n=1}^{K} \alpha_n w_n(u). \tag{37}$$

Its coefficient sum is

$$A_K = \sum_{n=1}^{K} |\alpha_n| = \frac{\mathsf{T}_{K+1}(3) - 1}{2\,4^K} - 1, \tag{38}$$

where T_j is the Chebyshev polynomial of the first kind. This quantity grows exponentially and measures the sensitivity of the optimum under zero moment bounds. Hence uniform moment bounds $\epsilon_n \le \epsilon$ imply

$$\left|\int w_{K+1} r\right| \le B_\infty \frac{K+1}{4^K} + A_K \epsilon. \tag{39}$$

The polynomial $\mathsf{U}_K(2u-1) = 4^K \prod_{j=1}^{K}(u - u_j)$ has positive roots. Its coefficients therefore alternate in sign. After its leading term is cancelled in Eq. (37), the coefficient γ_j of u^j in p_K has

sign $(-1)^{K-j+1}$ for $0 \le j < K$. Since $\alpha_{j+1} = \gamma_j/(j+1)$,

$$A_K = \int_0^1 \sum_{j=0}^{K-1} |\gamma_j| u^j \, du = (-1)^{K+1} \int_0^1 p_K(-u) \, du$$

$$= (K+1)4^{-K} \int_0^1 \mathsf{U}_K(2u+1) \, du - 1.$$

The last identity uses $\mathsf{U}_K(-x) = (-1)^K \mathsf{U}_K(x)$. Integrating with $\mathsf{T}'_{K+1} = (K+1)\mathsf{U}_K$ and using $\mathsf{T}_{K+1}(1) = 1$ yields Eq. (38). Substituting the resulting coefficients into the upper bound of Thm. 2 proves Eq. (39).

D Additional Labels under Calibration Constraints

D.1 Observation Model and Minimax Characterization

We supplement known calibration at small budgets with m labels of ordinary candidates. Let $\mathcal{P}_K$ denote the models of Sec. 2 with continuous conditional score CDFs, $C = S$, and full calibration at $1, \dots, K$. These population equalities are given, and the data consist of m independent tuples (X, S, U, Y), each from a fresh prompt. Unlike the setting with calibration constraints alone, this experiment assumes observed true U. The estimand is the selected mean residual Δ_N.

The constraints restrict which residuals are possible; the labels restrict which are statistically indistinguishable. Both restrictions are captured by

$$\omega_{K,N}(t) = \sup_{\substack{\|r\|_\infty \le 1,\ \|r\|_2 \le t \\ \int pr = 0 \ (p \in \mathcal{V}_K)}} \left| \int w_N r \right|. \tag{40}$$

Boundedness comes from binary outcomes; the zero moments encode the known calibration conditions. The L^2 radius controls statistical distinguishability.

Theorem 8 (Joint minimax error). *For all $K, N, m \ge 1$, let $\mathfrak{R}_{K,N,m} = \inf_{\widehat{\Delta}} \sup_{P \in \mathcal{P}_K} \mathbb{E}_{P^{\otimes m}} |\widehat{\Delta} - \Delta_N(P)|$. Then*

$$\frac{1}{64} \omega_{K,N}(m^{-1/2}) \le \mathfrak{R}_{K,N,m} \le \omega_{K,N}(m^{-1/2}). \tag{41}$$

The lower bound continues to hold with unlimited independent unlabeled observations. For $t > 0$,

$$\omega_{K,N}(t) = \inf_{\substack{p \in \mathcal{V}_K \\ \psi \in L^2[0,1]}} \{\|w_N - p - \psi\|_1 + t\|\psi\|_2\}. \tag{42}$$

For any such (p, ψ), the estimator

$$\widehat{\Delta}_{N,\psi} = \frac{1}{m} \sum_{i=1}^{m} \psi(U_i)(Y_i - S_i) \tag{43}$$

has worst-case expected absolute error at most $\|w_N - p - \psi\|_1 + \|\psi\|_2/\sqrt{m}$.

The same bounds hold if $\mathcal{P}_K$ is enlarged to require only mean calibration at $1, \dots, K$, with the same observations and $C = S$. The upper bound uses only the zero moments, whereas both models used for the lower bound satisfy full calibration, so both classes have the same minimax error up to the stated constants.

The dual form decomposes the target weight into a polynomial p eliminated by known calibration constraints, a component ψ estimated from labels, and an approximation residual. The objective balances the resulting bounds on bias and standard deviation. The dual identity and minimax bounds are proved below; App. D.3 gives a formula for the orthogonal projection error and the regime with small noise.

With only ordinary calibration known, $K = 1$ gives, for $N \ge 2$ and $0 < t \le 1$,

$$\tfrac{1}{4} \min\{1, t\sqrt{N}\} \le \omega_{1,N}(t) \le \min\{1, t\sqrt{N}\}. \tag{44}$$

Even for ordinarily calibrated scores, the minimax rate is $\min\{1, \sqrt{N/m}\}$, reflecting the scarcity of labels at high relative ranks. Thm. 10 gives a lower bound at a fixed confidence level in the same class.

D.2 Proof of the Minimax Characterization

Proof of Thm. 8. Duality. For any feasible r, $p \in \mathcal{V}_K$, and $\psi \in L^2$,

$$\left| \int w_N r \right| \le \|w_N - p - \psi\|_1 \|r\|_\infty + \|\psi\|_2 \|r\|_2 \le \|w_N - p - \psi\|_1 + t\|\psi\|_2.$$

For the reverse inequality, consider the continuous seminorm

$$\mathcal{J}_t(f) = \inf_{p \in \mathcal{V}_K,\ \psi \in L^2} \{\|f - p - \psi\|_1 + t\|\psi\|_2\}, \qquad f \in L^1.$$

It is bounded by $\|f\|_1$, vanishes on $\mathcal{V}_K$, and satisfies $\mathcal{J}_t(\psi) \le t\|\psi\|_2$ on L^2. The Hahn–Banach argument from Thm. 2 gives $r \in L^\infty$ with $\|r\|_\infty \le 1$, $\int w_N r = \mathcal{J}_t(w_N)$, and $|\int fr| \le \mathcal{J}_t(f)$ for every $f \in L^1$. It follows that r annihilates $\mathcal{V}_K$ and $|\int \psi r| \le t\|\psi\|_2$ on L^2. Taking $\psi = r$ gives $\|r\|_2 \le t$. Thus r is feasible and attains the dual value, proving Eq. (42) and existence of a maximizing residual.

Upper bound. For $P \in \mathcal{P}_K$, let $r_P(u) = \mathbb{E}[Y - S \mid U = u]$. Then $\|r_P\|_\infty \le 1$ and $\int p r_P = 0$ for $p \in \mathcal{V}_K$. The estimator in Eq. (43) therefore satisfies

$$|\mathbb{E}_P \widehat{\Delta}_{N,\psi} - \Delta_N(P)| = \left| \int (w_N - p - \psi) r_P \right| \le \|w_N - p - \psi\|_1,$$
$$\mathrm{Var}_P(\widehat{\Delta}_{N,\psi}) \le \frac{1}{m} \mathbb{E}_P[\psi(U)^2 (Y - S)^2] \le \frac{\|\psi\|_2^2}{m}.$$

The triangle inequality and Cauchy–Schwarz give the claimed bound on expected absolute error. Taking deterministic pairs (p, ψ) approaching the infimum in Eq. (42) proves the upper bound in Eq. (41); the infimum need not be attained.

Calibrated two-point lower bound. The claim is immediate if $\omega_{K,N}(m^{-1/2}) = 0$. Otherwise, choose a maximizing residual r with positive target moment. It satisfies $\|r\|_\infty \le 1$, $\|r\|_2 \le m^{-1/2}$, and $\int pr = 0$ for $p \in \mathcal{V}_K$. Use the common score law of Lem. 2 and define the two label kernels by

$$P_0(Y = 1 \mid X, U) = S, \qquad P_1(Y = 1 \mid X, U) = S + \tfrac{1}{8} \mathbf{1}\{S \in A\} r(U).$$

The construction ensures that both models belong to $\mathcal{P}_K$, have identical prompt, score, and rank distributions, and satisfy

$$\delta_{\mathrm{sep}} = \Delta_N(P_1) - \Delta_N(P_0) = \frac{3}{32} \omega_{K,N}(m^{-1/2}). \tag{45}$$

The Bernoulli inequality $D_{\mathrm{KL}}(\mathrm{Bern}(p) \| \mathrm{Bern}(q)) \le (p - q)^2 / [q(1 - q)]$, together with $S(1 - S) \ge 3/16$ on A and $P(S \in A \mid U) = 3/4$, gives

$$D_{\mathrm{KL}}(P_1 \| P_0) \le 4(1/8)^2 \|r\|_2^2, \qquad D_{\mathrm{KL}}(P_1^{\otimes m} \| P_0^{\otimes m}) \le 1/16.$$

Independent unlabeled observations add zero KL, and independent estimator randomization leaves these bounds unchanged. Pinsker's inequality bounds the total variation distance of the entire experiment by $1/\sqrt{32}$.

Choosing the nearer target value converts any estimator into a test. A testing error entails estimation error at least $\delta_{\mathrm{sep}}/2$, while the sum of the two testing error probabilities is at least one minus total variation. Hence

$$\max_{j=0,1} \mathbb{E}_{P_j} |\widehat{\Delta}_N - \Delta_N(P_j)| \ge \frac{\delta_{\mathrm{sep}}}{4} (1 - 1/\sqrt{32})$$
$$= \frac{3}{128} (1 - 1/\sqrt{32})\, \omega_{K,N}(m^{-1/2}) \ge \frac{1}{64} \omega_{K,N}(m^{-1/2}),$$

proving the lower bound. The same reduction yields error at least $\delta_{\mathrm{sep}}/2$ with probability at least $(1 - 1/\sqrt{32})/2 > 1/3$ under one of the two models. □

D.3 Consequences

At $t = 1$, the L^2 constraint is redundant and $\omega_{K,N}(1) = d_{K,N}$. For any $t > 0$,

$$\omega_{K,N}(t) \le \min\{d_{K,N},\, t\nu_{K,N}\}, \tag{46}$$

where $\nu_{K,N} = \mathrm{dist}_2(w_N, \mathcal{V}_K)$. For $N > K$,

$$\nu_{K,N}^2 = \frac{N^2}{2N-1} \prod_{j=1}^{K} \left(\frac{N-j}{N+j-1} \right)^2. \tag{47}$$

Corollary 4 (Behavior for small noise). *Eq. (46), Eq. (47), and Eq. (44) hold, and $\omega_{K,N}(1) = d_{K,N}$. If $e = w_N - \Pi_{\mathcal{V}_K}^{L^2} w_N$ is nonzero and $t \le \|e\|_2/\|e\|_\infty$, then $\omega_{K,N}(t) = t\nu_{K,N}$. The projection here is orthogonal with respect to Lebesgue measure.*

Proof of Cor. 4. Since the interval has measure one, every r with $\|r\|_\infty \le 1$ has $\|r\|_2 \le 1$. Thus the additional L^2 constraint at $t = 1$ is redundant and Eq. (28) gives $\omega_{K,N}(1) = d_{K,N}$. Dropping the L^2 constraint always gives the bound $d_{K,N}$. Orthogonality to $\mathcal{V}_K$ and Cauchy–Schwarz give the bound $t\nu_{K,N}$.

For $K = 1$, let $r_0(u) = \mathbf{1}\{u > 1 - 1/N\} - 1/N$. It has integral zero, sup norm at most one, and $\|r_0\|_2^2 = N^{-1}(1 - N^{-1}) \le N^{-1}$. Its selected moment equals

$$1 - (1 - 1/N)^N - 1/N \ge 1/4 \quad (N \ge 2).$$

For $N = 2$ this is equality; for $N \ge 3$, use $(1 - 1/N)^N \le e^{-1}$ and $1/N \le 1/3$. The residual $\min\{1, t\sqrt{N}\}r_0$ is feasible and proves the lower bound in Eq. (44). The upper bound follows from $d_{1,N} \le 1$ and $\nu_{1,N}^2 \le \|w_N\|_2^2 = N^2/(2N-1) \le N$.

For completeness, we compute this distance. The squared distance of u^λ, $\lambda = N - 1 \ge K$, from the span of $u^0, \dots, u^{K-1}$ is the Schur complement of their Gram matrix, equivalently the ratio of the augmented Gram determinant to the original one. For distinct exponents λ_i, the Gram entries are $G_{ij} = 1/(\lambda_i + \lambda_j + 1)$. The Cauchy determinant identity gives

$$\det G = \frac{\prod_{i<j}(\lambda_j - \lambda_i)^2}{\prod_{i,j}(\lambda_i + \lambda_j + 1)}.$$

Taking the ratio before and after adjoining λ gives

$$\mathrm{dist}_2(u^\lambda, \mathcal{V}_K)^2 = \frac{1}{2\lambda + 1} \prod_{i=0}^{K-1} \left(\frac{\lambda - i}{\lambda + i + 1} \right)^2.$$

Multiplication by N^2 and the substitution $j = i + 1$ yield Eq. (47). For $N \le K$ the distance is zero directly.

Finally let $e = w_N - \Pi_{\mathcal{V}_K}^{L^2} w_N$ be nonzero. The residual $r = te/\|e\|_2$ is orthogonal to $\mathcal{V}_K$, has L^2 norm t, and has sup norm at most one under the stated condition. Since $\int w_N e = \|e\|_2^2$, its target moment equals $t\nu_{K,N}$. This proves the small-t equality. □

When $K^2/N \to 0$, the product in Eq. (47) tends to one: each factor is $1 - (2j-1)/(N+j-1)$, and the sum of logarithms of the magnitudes is $O(K^2/N)$. This explains why variance reduction from the calibration constraints is negligible far beyond the quadratic scale.

E Passive Data and Lower Bounds

E.1 Simultaneous Confidence Bounds

For a fixed predictor and a fixed finite test class $\mathcal{H}$, observe m independent ordinary candidates with known U and use the estimator Eq. (16). Let $v_N = N^2/(2N-1)$, $\tau_\delta = \log(2|\mathcal{B}||\mathcal{H}|/\delta)$, and define

$$r_N(\tau_\delta) = \sqrt{\frac{2v_N\tau_\delta}{m}} + \frac{4N\tau_\delta}{3m}. \tag{48}$$

Theorem 9 (Confidence bounds from passive labels). *Let $\mathcal{B} \subset \mathbb{N}$ and $\mathcal{H}$ be fixed finite nonempty sets, and let $0 < \delta < 1$. For a predictor fitted independently of m i.i.d. observations of ordinary candidates with known population ranks, the estimator Eq. (16) satisfies, with probability at least $1 - \delta$,*

$$|\widehat{\theta}_{N,h} - \theta_{N,h}| \le r_N(\tau_\delta) \qquad ((N,h) \in \mathcal{B} \times \mathcal{H}). \tag{49}$$

Proof. By Lem. 1, $W = w_N(U)Rh(C)$ has mean $\theta_{N,h}$. Moreover, $|W - \mathbb{E}W| \le 2N$ and

$$\mathrm{Var}(W) \le \mathbb{E}W^2 \le \mathbb{E}w_N(U)^2 = N^2 \int_0^1 u^{2N-2}\,du = v_N.$$

Bernstein's inequality with these bounds gives failure probability at most $2e^{-\tau_\delta}$ at the radius in Eq. (48). A union bound over $\mathcal{B} \times \mathcal{H}$ completes the proof; no independence between budget estimates is required. □

The leading confidence radius matches the $K = 1$ minimax scale. The effective sample size (Agapiou et al., 2017) is $m/v_N \approx 2m/N$; particular residuals or bins may have smaller variance than this worst case.

E.2 A Lower Bound within the Calibrated Class

Theorem 10 (Lower bound for passive labels in the calibrated class). *Let $m \ge 1$ and $N \ge 2$. Consider the class of conditionally i.i.d. candidate models with continuous conditional score distributions and ordinary score calibration $\mathbb{E}[Y \mid S] = S$, with $C = S$. Even if X, S, the true U, and arbitrarily much unlabeled score information are available, every estimator $\widehat{\Delta}_N$ based on m passive i.i.d. correctness labels satisfies*

$$\sup_{P \text{ in this class}} P^{\otimes m}\left(|\widehat{\Delta}_N - \Delta_N(P)| \ge \frac{3}{256}\min\left\{1, \sqrt{\frac{N}{m}}\right\}\right) > \frac{1}{3}. \tag{50}$$

Here the unlabeled information has the same distribution in the two models used in the proof, so it cannot distinguish them.

Proof. Use the score construction in Lem. 2. Under P_0, let $Y \mid X, U \sim \mathrm{Bern}(S)$; under P_1, let

$$P_1(Y = 1 \mid X, U) = S + \lambda_m \mathbf{1}\{S \in A\}\left(\mathbf{1}\{U > 1 - 1/N\} - \frac{1}{N}\right), \tag{51}$$

where $\lambda_m = \frac{1}{8}\min\{1, \sqrt{N/m}\}$. The perturbation has zero integral over U, so both models using raw scores are ordinarily calibrated. Their prompt, score, and rank distributions coincide; P_0 is calibrated at every budget. By Eq. (24), their target separation is

$$\delta_{\mathrm{sep}} := \Delta_N(P_1) - \Delta_N(P_0) = \frac{3\lambda_m}{4}\left(1 - (1 - 1/N)^N - \frac{1}{N}\right) \ge \frac{3\lambda_m}{16}. \tag{52}$$

The parenthesized term equals $1/4$ at $N = 2$ and exceeds $1/4$ for $N \ge 3$, since $(1 - 1/N)^N \le e^{-1}$ and $1/N \le 1/3$.

Put $\tau = 1/N$ and $r(U) = \mathbf{1}\{U > 1 - \tau\} - \tau$. Then $\mathbb{E}[\mathbf{1}\{S \in A\}r(U)^2] = \frac{3}{4}\tau(1 - \tau)$. Applying the Bernoulli KL bound used in the proof of Thm. 8, with $S(1 - S) \ge 3/16$ on A, yields

$$D_{\mathrm{KL}}(P_1 \| P_0) \le 4\lambda_m^2/N, \qquad D_{\mathrm{KL}}(P_1^{\otimes m} \| P_0^{\otimes m}) \le \frac{1}{16}. \tag{53}$$

Independent unlabeled data add zero KL, so Pinsker's inequality bounds the total variation distance of the laws of all observed data by $1/\sqrt{32}$. The same testing reduction based on the nearer target as in Thm. 8 gives estimation error at least $\delta_{\mathrm{sep}}/2$ with probability at least $(1 - 1/\sqrt{32})/2 > 1/3$ under one of the models. Since $\delta_{\mathrm{sep}}/2 \ge (3/256)\min\{1, \sqrt{N/m}\}$, this proves Eq. (50). □

E.3 Conditional Statements for Finitely Many Bins

Corollary 5 (Conditional reliability for finitely many bins)**.** *For fixed bins $I_1, \dots, I_J$, estimate*

$$\widehat{q}_{N,b} = \frac{1}{m}\sum_i w_N(U_i)\mathbf{1}\{C_i \in I_b\}, \qquad \widehat{\theta}_{N,b} = \frac{1}{m}\sum_i w_N(U_i)R_i\mathbf{1}\{C_i \in I_b\}.$$

Use Eq. (48) with $\tau_\delta = \log(4|\mathcal{B}|J/\delta)$. With probability at least $1-\delta$, for every pair with $\widehat{q}_{N,b} > r_N(\tau_\delta)$,

$$|\mathbb{E}_N[Y - C \mid C \in I_b]| \le \frac{|\widehat{\theta}_{N,b}| + r_N(\tau_\delta)}{\widehat{q}_{N,b} - r_N(\tau_\delta)}. \tag{54}$$

Also $\sum_b |\theta_{N,b}| \le \sum_b (|\widehat{\theta}_{N,b}| + r_N(\tau_\delta))$ simultaneously.

Proof. The same Bernstein calculation controls each mass estimate, since $0 \le w_N\mathbf{1}\{C \in I_b\} \le N$ and its second moment is at most v_N. Union bound over residuals and masses. On that event, $q_{N,b} \ge \widehat{q}_{N,b} - r_N(\tau_\delta) > 0$ and $|\theta_{N,b}| \le |\widehat{\theta}_{N,b}| + r_N(\tau_\delta)$. Divide the inequalities and use the definition of a residual within a bin. □

If the empirical mass lower bound is not positive, this conditional reliability bound is unavailable for that bin.

E.4 Honest Confidence Intervals

Corollary 6 (An honest interval must sometimes be wide)**.** *For the two models above, let an interval $\mathcal{C}$ cover $\Delta_N(P_j)$ with probability at least $1-\alpha$ under each P_j. Then under P_0,*

$$P_0\{\mathrm{length}(\mathcal{C}) \ge \delta_{\mathrm{sep}}\} \ge 1 - 2\alpha - \mathrm{TV}(P_0^{\otimes m}, P_1^{\otimes m}). \tag{55}$$

Proof. Coverage of the alternative target under P_0 is at least $1 - \alpha - \mathrm{TV}$ by the definition of total variation. Intersect that event with null coverage under P_0. On the intersection the interval contains both values, which are separated by δ_{sep}. □

F Complete Pool Estimation

This section develops the estimator using complete pools introduced in Sec. 4. Suppose each of m independent prompts has a fully labeled pool of n_{pool} i.i.d. candidates. Sort its keys in ascending order and define $H_{j,(r)} = (Y_{j,(r)} - C_{j,(r)})h(C_{j,(r)})$. For $N \le n_{\mathrm{pool}}$, average the selected residual over every N-subset (Nakano et al., 2021):

$$T_{j,N,h} = \sum_{r=N}^{n_{\mathrm{pool}}} \frac{\binom{r-1}{N-1}}{\binom{n_{\mathrm{pool}}}{N}} H_{j,(r)}, \qquad \overline{T}_{N,h} = \frac{1}{m}\sum_j T_{j,N,h}. \tag{56}$$

Sorting costs $O(n_{\mathrm{pool}} \log n_{\mathrm{pool}})$ per prompt; each budget–test sum costs $O(n_{\mathrm{pool}})$. The estimator uses mn_{pool} labels, with uncertainty controlled across the m independent prompts.

Theorem 11 (Confidence bounds from complete pools)**.** *For m independent prompts with fully labeled pools of n_{pool} conditionally i.i.d. candidates, the statistic Eq. (56) is unbiased for $\theta_{N,h}$ for every $N \le n_{\mathrm{pool}}$. For a fixed finite $\mathcal{B} \subseteq \{1, \dots, n_{\mathrm{pool}}\}$ and fixed finite $\mathcal{H}$, with probability at least $1-\delta$,*

$$|\overline{T}_{N,h} - \theta_{N,h}| \le \sqrt{\frac{2\log(2|\mathcal{B}||\mathcal{H}|/\delta)}{m}} \tag{57}$$

holds simultaneously over $\mathcal{B} \times \mathcal{H}$. Reports depend only on individual candidates and are fixed across subsets.

Proof. There are $\binom{r-1}{N-1}$ subsets of size N whose largest rank is r. Thus Eq. (56) is the average selected residual over all N-subsets. Each subset has the conditional law of N fresh i.i.d. candidates, so

$$\mathbb{E}[T_{j,N,h} \mid X_j] = \theta_{N,h}(X_j) := \mathbb{E}[R_{I_N}h(C_{I_N}) \mid X_j].$$

Averaging over prompts proves unbiasedness. The subset weights sum to one, giving $T_{j,N,h} \in [-1, 1]$. Hoeffding's inequality for the m independent prompt averages gives a tail bound $2\exp(-m\epsilon^2/2)$ for each budget–test pair; a union bound yields Eq. (57). □

Proposition 1 (Separating variance within and across prompts)**.** *Let H_N be the residual test of the candidate with the largest score from N fresh candidates at a fixed prompt. Then*

$$\mathrm{Var}(T_{j,N,h}) \le \mathrm{Var}_X\{\theta_{N,h}(X)\} + \frac{N}{n_{\mathrm{pool}}}\mathbb{E}_X\,\mathrm{Var}(H_N \mid X) \le \mathrm{Var}_X\{\theta_{N,h}(X)\} + \frac{N}{n_{\mathrm{pool}}}. \quad (58)$$

Consequently $\mathrm{Var}(\overline{T}_{N,h})$ *is bounded by the right-hand side divided by* m*.*

Proof. Conditional on $X = x$, the statistic is the complete order-N U-statistic with kernel H_N. Let $\zeta_l(x) \ge 0$ denote the degree-l canonical projection variance in its Hoeffding decomposition. Orthogonality and a coefficient comparison give

$$\mathrm{Var}(H_N \mid x) = \sum_{l=1}^{N} \binom{N}{l} \zeta_l(x),$$

$$\mathrm{Var}(T_{j,N,h} \mid x) = \sum_{l=1}^{N} \frac{\binom{N}{l}^2}{\binom{n_{\mathrm{pool}}}{l}} \zeta_l(x) \le \frac{N}{n_{\mathrm{pool}}}\,\mathrm{Var}(H_N \mid x),$$

since $\binom{N}{l}/\binom{n_{\mathrm{pool}}}{l} = \prod_{q=0}^{l-1}(N-q)/(n_{\mathrm{pool}}-q) \le N/n_{\mathrm{pool}}$ for $1 \le l \le N$. The law of total variance and $\mathbb{E}[T \mid X] = \theta(X)$ prove the first inequality in Eq. (58); $|H_N| \le 1$ proves the second. Independence of the m pools gives the bound for their average. □

More candidates at the same prompt reduce only the component within a prompt. For example, let $X \sim \mathrm{Bern}(p)$, $Y = X$ deterministically, scores i.i.d. uniform, and $C = 1/2$. Every pool statistic equals $X - 1/2$ regardless of n_{pool}, so an infinite pool still leaves only m independent observations of the mixture of prompt types. Pool size is not a substitute for prompt diversity.

F.1 The Conditional Hoeffding Decomposition

We record the variance calculation underlying Prop. 1. Fix a prompt x and suppress this conditioning. For a symmetric kernel H on N candidates with mean θ, define

$$G_l(z_1, \ldots, z_l) = \mathbb{E}[H(z_1, \ldots, z_l, Z_{l+1}, \ldots, Z_N) \mid x],$$

$$h_l(z_1, \ldots, z_l) = G_l(z_1, \ldots, z_l) - \sum_{A \subsetneq \{1,\ldots,l\}} h_{|A|}(z_A), \qquad G_0 = h_0 = \theta.$$

Each h_l, $l \ge 1$, has conditional mean zero in every argument. Consequently, distinct canonical terms are orthogonal: integrate an argument belonging to only one of their index sets. The expansion

$$H(Z_1, \ldots, Z_N) = \theta + \sum_{l=1}^{N} \sum_{\substack{A \subseteq [N] \\ |A| = l}} h_l(Z_A) \quad (59)$$

therefore gives $\mathrm{Var}(H) = \sum_l \binom{N}{l}\zeta_l$, where $\zeta_l = \mathbb{E}h_l^2$.

Averaging over all N-subsets of an n_{pool}-candidate pool assigns each l-subset the coefficient $\binom{n_{\mathrm{pool}}-l}{N-l}/\binom{n_{\mathrm{pool}}}{N} = \binom{N}{l}/\binom{n_{\mathrm{pool}}}{l}$. The same orthogonality gives

$$\mathrm{Var}(T \mid x) = \sum_{l=1}^{N} \binom{n_{\mathrm{pool}}}{l} \left(\frac{\binom{N}{l}}{\binom{n_{\mathrm{pool}}}{l}}\right)^2 \zeta_l,$$

which is the required formula. All calculations are conditional on x; variance across prompts enters only through the final application of total variance.

G Reporting and Conditional Reliability

This section provides the conditional identity used in the construction with raw scores and the reporting target for the empirical comparison.

Proposition 2 (Conditional reliability after selection). *Let* $m_N(C) = \mathbb{E}[w_N(U) \mid C]$. *Then* $m_N(C) > 0$ *almost surely and*

$$\mathbb{E}_N[R \mid C] = \frac{\mathbb{E}[w_N(U)R \mid C]}{m_N(C)}. \tag{60}$$

If $\mathbb{E}[R \mid C] = 0$, *the numerator equals* $\mathrm{Cov}(w_N(U), R \mid C)$, *or equivalently* $\mathrm{Cov}(w_N(U), Y \mid C)$.

Proof. The weight is positive for $U \in (0,1)$, an event of probability one, so its conditional expectation is positive almost surely. For any bounded h, the right-hand side of Eq. (60), integrated against $h(C)$ under P_N, becomes

$$\mathbb{E}\left[m_N(C)h(C)\frac{\mathbb{E}[w_N R \mid C]}{m_N(C)}\right] = \mathbb{E}[w_N R h(C)],$$

which is the defining property of $\mathbb{E}_N[R \mid C]$. If the unweighted conditional residual is zero, subtracting $m_N(C)\mathbb{E}[R \mid C]$ yields the covariance identity. Since C is fixed under conditioning, replacing R by Y does not alter the covariance. □

Corollary 7 (Two sufficient conditions for the raw score). *Take* $C = S$. *Each of the following implies full calibration at every budget:*

(i) $\mathbb{E}[Y \mid X, S] = S$ *almost surely;*

(ii) $\mathbb{E}[Y \mid S] = S$ *and* S *has the same conditional CDF for almost every prompt.*

For continuous conditional scores, (ii) can be weakened to: U *is a measurable function of* S *and* $\mathbb{E}[Y \mid S] = S$.

Proof. For (i), independent tie randomization gives $\mathbb{E}[Y - S \mid X, S, V] = 0$. Since (S, U) is measurable with respect to (X, S, V), the tower property gives the condition in Thm. 4. Under (ii), U is a function of (S, V) with a common CDF. Because V is independent of (X, S, Y), conditioning additionally on U does not change $\mathbb{E}[Y \mid S]$. The assertion for continuous scores is immediate because conditioning on (S, U) is then the same as conditioning on S. □

Proposition 3 (Oracle reporting without changing selection). *Define*

$$\eta(s,u) = \mathbb{E}[Y \mid S = s, U = u], \qquad C^\star = \eta(S,U). \tag{61}$$

Then $C^\star$ *is fully calibrated under every* P_N. *If* g *is an injective measurable map of the score with a measurable inverse on its range,* $C = g(S)$ *is fully calibrated for every* N *if and only if* $\eta(S,U) = g(S)$ *almost surely.*

Proof. By iterated conditioning,

$$\mathbb{E}[Y - C^\star \mid C^\star, U] = \mathbb{E}[\mathbb{E}[Y - C^\star \mid S, U] \mid C^\star, U] = 0.$$

Apply Thm. 4. If g is injective with measurable inverse, $(g(S), U)$ and (S, U) generate the same sigma-field. Thus the necessary and sufficient condition in Thm. 4 becomes $\mathbb{E}[Y \mid S, U] = g(S)$. □

G.1 Error of an Approximate Predictor

Proposition 4 (Regression error versus selected calibration). *Let* $C = g(S,U)$ *for a fixed measurable* $g : [0,1]^2 \to [0,1]$, *and let* η *be defined by Eq. (61). Then*

$$\mathrm{ECE}_N(C) \le \mathbb{E}_N|\eta(S,U) - g(S,U)| \le \sqrt{v_N}\,\|\eta - g\|_{L^2(P)}. \tag{62}$$

If $\|\eta - g\|_\infty \le \epsilon$ *in the essential sense,* $\mathrm{ECE}_N(C) \le \epsilon$ *for every budget, with no factor depending on* N. *Also,*

$$\mathbb{E}_N[(Y-g)^2] - \mathbb{E}_N[(Y-\eta)^2] = \mathbb{E}_N[(g-\eta)^2] \le N\mathbb{E}[(g-\eta)^2]. \tag{63}$$

Proof. Since the density ratio depends only on U, conditional correctness given (S, U) is unchanged by selection: $\mathbb{E}_N[Y \mid S, U] = \eta(S, U)$. Iterating to C and applying conditional Jensen gives the first inequality in Eq. (62). Cauchy–Schwarz under P and $\mathbb{E}w_N^2 = v_N$ give the second. The uniform bound follows directly by integrating against a probability law. For Eq. (63), expand the squares; the cross term has conditional expectation zero given (S, U). Finally $w_N \le N$. □

Proposition 5 (Prediction using independent reference samples)**.** *For each prompt, generate $k \ge 1$ conditionally i.i.d. reference candidates from Q_X, independently of the entire selection pool, with independent tie variables as in Sec. 2. Define*

$$R_{\mathrm{ref}} = \sum_{j=1}^{k} \mathbf{1}\{(S_j^{\mathrm{ref}}, V_j^{\mathrm{ref}}) <_{\mathrm{lex}} (S, V)\}, \qquad \widehat{U}_k = R_{\mathrm{ref}}/k,$$

where $<_{\mathrm{lex}}$ is the lexicographic order and V is the same candidate tie variable used in selection. The reference keys may be shared across candidates but remain fixed as competitors are added. Selection uses the original lexicographic key (S, V) and does not use the reference pool or the reported probability. Let $g : [0,1]^2 \to [0,1]$ be fitted on independent data and held fixed.

Suppose η in Eq. (61) has a measurable $[0,1]$-valued extension to $[0,1]^2$, uniformly Λ-Lipschitz in its second coordinate. Define $b_{N,k}(g) = \mathbb{E}_N|g(S, \widehat{U}_k) - \eta(S, \widehat{U}_k)|$. Then

$$\mathrm{ECE}_N(g(S, \widehat{U}_k)) \le b_{N,k}(g) + \Lambda\sqrt{\frac{N}{k(N+1)(N+2)}}. \tag{64}$$

If $|g - \eta| \le \delta_{\mathrm{fit}}$ on the deployed inputs, this is at most $\delta_{\mathrm{fit}} + \Lambda/\sqrt{6k}$ for every integer $N \ge 1$. The expectations include reference randomness; this is not a conditional guarantee for each realized reference pool.

Proof. Conditional on (X, S, V, Y), the reference comparisons are independent Bernoulli variables with success probability $F_X(S^-) + V\{F_X(S) - F_X(S^-)\} = U$. Consequently, $R_{\mathrm{ref}} \mid X, S, U, Y \sim \mathrm{Bin}(k, U)$. Let $\mathcal{R}$ denote the reference pool, including its tie variables. Since $\mathcal{R}$ is independent of the selection pool conditional on X, and selection ignores $\mathcal{R}$, $P(I_N = i \mid X, Z_i, V_i, \mathcal{R}) = U_i^{N-1}$. Thus the augmented selected law still has density ratio $w_N(U)$. The conditional kernel of R_{ref} depends only on U and is unchanged by this weighting, so $\mathbb{E}_N[Y \mid S, U, R_{\mathrm{ref}}] = \eta(S, U)$. Conditional Jensen and the Lipschitz assumption give

$$\mathrm{ECE}_N(g(S, \widehat{U}_k)) \le \mathbb{E}_N|\eta(S, U) - g(S, \widehat{U}_k)| \le b_{N,k}(g) + \Lambda\sqrt{\mathbb{E}_N(U - \widehat{U}_k)^2}.$$

The binomial conditional mean is U and its normalized variance is $U(1-U)/k$. Therefore

$$\mathbb{E}_N(U - \widehat{U}_k)^2 = \frac{1}{k}\int_0^1 N u^{N-1} u(1-u)\, du = \frac{N}{k(N+1)(N+2)} \le \frac{1}{6k}. \tag{65}$$

Substitution proves the claims. □

The extension assumption specifies η at inputs on the empirical rank grid outside the true (S, U) support. Learning error must also be controlled at the deployed inputs; under the ordinary augmented law, Cauchy–Schwarz gives $b_{N,k}(g) \le \sqrt{v_N}\|g(S, \widehat{U}_k) - \eta(S, \widehat{U}_k)\|_2$. A bound on $L^2(P_{S,U})$ error at the true inputs alone does not control predictions on the discrete grid.

Alternatively, if $g(s, \cdot)$ itself is uniformly Λ_g-Lipschitz, no Lipschitz extension of η is needed. Adding and subtracting $g(S, U)$ in the same proof gives

$$\mathrm{ECE}_N(g(S, \widehat{U}_k)) \le \sqrt{v_N}\|\eta - g\|_{L^2(P_{S,U})} + \Lambda_g\sqrt{\frac{N}{k(N+1)(N+2)}}. \tag{66}$$

This separates ordinary regression error from error due to finitely many references; it does not apply to an arbitrary discontinuous histogram map.

Under the uniform fitting assumption in Prop. 5, a sufficient rule for tolerance $\tau > \delta_{\mathrm{fit}}$ at every budget is $k \ge \max\{1, \lceil \Lambda^2/[6(\tau - \delta_{\mathrm{fit}})^2] \rceil\}$. Thus error due to finitely many references alone does

not require k to grow with N. If the fitting bound holds on a high-probability training event, the conclusion holds on that same event.

For the smoothed estimate $\widetilde{U}_k = (R_{\text{ref}} + 1)/(k + 2)$, the corresponding term for rank estimation is $\Lambda\sqrt{[kN + N^2 - N + 2]/[(N + 1)(N + 2)(k + 2)^2]}$, with fitting error evaluated at $\widetilde{U}_k$. This follows by integrating the conditional squared error $\{ku(1 - u) + (1 - 2u)^2\}/(k + 2)^2$; it retains the smoothing bias at the endpoints.

G.2 Histogram Construction

Partition the score space into finitely many measurable cells J_a, and partition $[0, 1]$ in the U coordinate into intervals $I_l^U = [u_{l-1}, u_l)$, with the final interval closed at one. Let the maximal U-width be $\Delta_U = \max_l(u_l - u_{l-1})$. Define the predictor to be the population correctness mean under the base law in each nonempty product cell:

$$C_0(s, u) = c_{a,l} := \mathbb{E}[Y \mid S \in J_a, U \in I_l^U]. \tag{67}$$

Cells of zero base probability also have zero probability under the selected law at each finite budget and may be assigned any report.

Theorem 12 (Budget dependence of binning by relative rank)**.** *The predictor C_0 in Eq. (67) is fully calibrated at $N = 1$. For $N \geq 2$,*

$$\mathrm{ECE}_N(C_0) \leq \frac{N\Delta_U}{2}. \tag{68}$$

More generally, if a fitted predictor $\widehat{C}$ is constant on these same cells and satisfies $|\widehat{c}_{a,l} - c_{a,l}| \leq e_{a,l}$, then

$$\mathrm{ECE}_N(\widehat{C}) \leq \frac{N\Delta_U}{2} + \sum_{a,l} P_N(S \in J_a, U \in I_l^U) e_{a,l}, \tag{69}$$

with the first term replaced by zero at $N = 1$.

Proof. Each cell has zero base residual under C_0. Grouping cells with the same output value therefore preserves zero conditional residual at $N = 1$.

Fix a product cell $G_{a,l}$. Let β_l be the midpoint of the range of w_N on I_l^U, and set $R_0 = Y - C_0$. Since $\mathbb{E}[R_0 \mathbf{1}_G] = 0$,

$$|\mathbb{E}[w_N R_0 \mathbf{1}_G]| = |\mathbb{E}[(w_N - \beta_l) R_0 \mathbf{1}_G]| \leq \tfrac{1}{2}\{w_N(u_l) - w_N(u_{l-1})\} P(G).$$

The total ECE of a piecewise constant report is at most the sum of the absolute weighted residuals over its cells: if several cells share a report, the triangle inequality only reduces the resulting conditional residual mass. Summing first over a uses $P(U \in I_l^U) = u_l - u_{l-1}$. Summing over l gives

$$\mathrm{ECE}_N(C_0) \leq \frac{1}{2}\sum_l (u_l - u_{l-1})\{w_N(u_l) - w_N(u_{l-1})\} \leq \frac{\Delta_U}{2}\{w_N(1) - w_N(0)\} = \frac{N\Delta_U}{2}.$$

For a fitted cell constant, write $Y - \widehat{C} = (Y - C_0) + (C_0 - \widehat{C})$. Apply the same cellwise triangle inequality to the first part and bound the second part by selected cell mass times $e_{a,l}$. This yields Eq. (69). At $N = 1$, the weight is constant, so the first part is zero. □

The upper bound may be truncated at one. The key point is the dependence on the width of the rank bins: fixed coarse U bins do not give calibration at arbitrarily large budgets. For a finite maximum budget B, a sufficient worst-case resolution is Δ_U of order ϵ/B; distributional smoothness can improve this conservative requirement.

For a fully finite-sample version, fix the cells independently of m_0 training candidates and let $n_{a,l}$ be their sample counts. The empirical mean of Y in a nonempty cell obeys, simultaneously over L_0 cells, with probability at least $1 - \delta_0$,

$$|\widehat{c}_{a,l} - c_{a,l}| \leq \min\left\{1, \sqrt{\frac{\log(2L_0/\delta_0)}{2n_{a,l}}}\right\}. \tag{70}$$

For an empty cell use the bound one. Conditional on the cell counts, the labels assigned to each cell are i.i.d. draws from that cell's conditional law, so Hoeffding and a union bound prove Eq. (70). Substitution into Eq. (69) bounds the calibration error of the fitted histogram. Histogram methods that analyze data reuse offer an alternative fitting procedure (Gupta and Ramdas, 2021).

H Independent Reference Ranks

H.1 Rank Weights from Independent References

For each independently sampled prompt, suppose one ordinary candidate is labeled and k additional candidates are generated without labels. These references are independent of the labeled candidate conditional on the prompt. Let R_{ref} be the number of reference keys below its key. All reference comparisons use the same candidate tie variable V as the selection rule. Conditional on the labeled candidate and its prompt,

$$R_{\mathrm{ref}} \mid U \sim \mathrm{Bin}(k, U). \tag{71}$$

This is stronger than a marginal binomial law: conditional on the complete labeled observation, the distribution of R_{ref} depends only on U.

Theorem 13 (Unbiased rank weights)**.** *For $1 \le N \le k+1$, define*

$$\widehat{w}_{N,k}(R_{\mathrm{ref}}) = N \frac{\binom{R_{\mathrm{ref}}}{N-1}}{\binom{k}{N-1}}, \tag{72}$$

with $\binom{R_{\mathrm{ref}}}{r} = 0$ for $R_{\mathrm{ref}} < r$. Then

$$\mathbb{E}[\widehat{w}_{N,k}(R_{\mathrm{ref}}) \mid U] = w_N(U), \qquad 0 \le \widehat{w}_{N,k}(R_{\mathrm{ref}}) \le N. \tag{73}$$

If C and the residual test do not use the same reference randomness, replacing $w_N(U)$ by $\widehat{w}_{N,k}(R_{\mathrm{ref}})$ gives an unbiased estimator of $\theta_{N,h}$. Moreover,

$$v_N \le \mathbb{E}\widehat{w}_{N,k}(R_{\mathrm{ref}})^2 \le N. \tag{74}$$

Thus the confidence bounds for passive labels apply with v_N replaced by the conservative variance proxy N.

Proof. Let $r = N-1$. Among the k references, $\binom{R_{\mathrm{ref}}}{r}$ counts the r-subsets all of whose keys lie below the labeled key. Each such subset has conditional probability U^r, so $\mathbb{E}[\binom{R_{\mathrm{ref}}}{r} \mid U] = \binom{k}{r}U^r$. This proves conditional unbiasedness and, using Eq. (71) conditional on the full labeled observation, unbiasedness of the product of the residual and weight. The pointwise bound is immediate. Since $\mathbb{E}\widehat{w} = \int w_N = 1$, the upper bound on the second moment follows from $\widehat{w}^2 \le N\widehat{w}$. Conditional Jensen gives the lower bound $\mathbb{E}\widehat{w}^2 \ge \mathbb{E}w_N^2 = v_N$. The centered summands are bounded by $2N$, as before. □

A formula for the second moment is also available. Draw two independent uniformly chosen r-subsets of $\{1, \dots, k\}$ and let J be their intersection size. Then

$$\mathbb{E}\widehat{w}_{N,k}^2 = N^2 \mathbb{E}\left[\frac{1}{2r - J + 1}\right], \qquad P(J = j) = \frac{\binom{r}{j}\binom{k-r}{r-j}}{\binom{k}{r}}. \tag{75}$$

Indeed, the product of the two subset indicators requires their union of $2r - J$ keys to fall below the labeled key; integrating U^{2r-J} over uniform U gives the reciprocal. At $k = N-1$, the weight is $N\mathbf{1}\{R_{\mathrm{ref}} = k\}$ and its second moment is N. As $k \to \infty$ with fixed N, the overlap tends to zero in probability and the second moment tends to v_N.

Proposition 6 (Conditional unbiasedness using only the rank of one candidate)**.** *If $N > k+1$, there is no function $\psi : \{0, \dots, k\} \to \mathbb{R}$ satisfying $\mathbb{E}[\psi(R_{\mathrm{ref}}) \mid U = u] = Nu^{N-1}$ for every $u \in [0,1]$ when $R_{\mathrm{ref}} \mid U = u \sim \mathrm{Bin}(k, u)$.*

Proof. For any such function,

$$\mathbb{E}[\psi(R_{\mathrm{ref}}) \mid U = u] = \sum_{l=0}^{k} \psi(l) \binom{k}{l} u^l (1-u)^{k-l},$$

a polynomial of degree at most k. It cannot equal the nonzero degree-$(N-1)$ polynomial Nu^{N-1} on an interval when $N - 1 > k$. □

I Additional Passive-Label Results

The passive-label experiment in Sec. 5.2 tests how accurately the selected mean residual can be estimated when both low-budget calibration constraints and ordinary labeled candidates are available. This complements the extrapolation experiment, which studies the calibration error left by the constraints alone. The main figure shows two slices: the middle target budget $N = 32$ for the label-count sweep, and the largest label count $m = 10^4$ for the target-budget sweep. The full grid below includes all 30 settings.

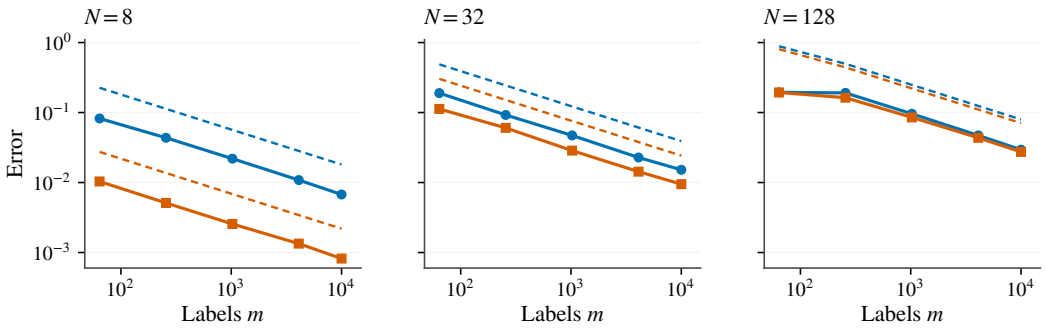


Figure 6: Full passive-label grid. ● $K = 1$; ■ $K = 4$. Solid curves with markers show the maximum MAE over $P_0, \ldots, P_3$; dashed curves show the continuous feasible dual bound for the same color and K. Shading gives approximate 95% Monte Carlo envelopes, simultaneous over the four model MAEs within each setting, based on 1000 repetitions. The small-label regime at $N = 128$ is included, including its departure from the clean $m^{-1/2}$ trend.

Increasing m from 256 to 4096 multiplies the label count by 16 and reduces the four-model maximum MAE by factors between 3.74 and 4.22 across the six (K, N) combinations. At $m = 10^4$, the $K = 1$ to $K = 4$ MAE ratio falls from 8.23 at $N = 8$ to 1.08 at $N = 128$; the corresponding modulus ratios are 8.25 and 1.12. These two effects explain the main figure's emphasis on label benefit and the diminishing contribution of calibration constraints at larger target budgets.

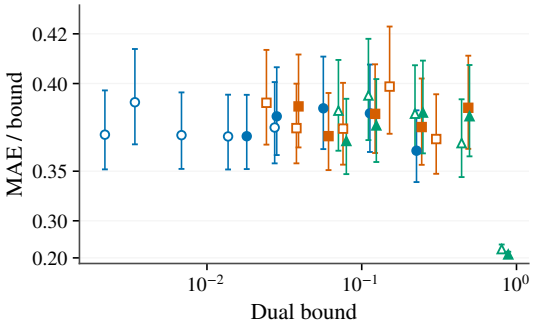


Figure 7: Auxiliary bound check for all 30 settings. Colors/markers: ● $N = 8$, ■ $N = 32$, ▲ $N = 128$; filled/open markers indicate $K = 1/K = 4$. The ratio r is the four-model maximum MAE divided by its feasible dual bound. Vertical positions use r^5 and ticks show the original r, expanding the dense upper cluster. Bars transform the same Monte Carlo envelopes as in Fig. 6. Ratios range from about 0.22 to 0.40, below 1.

The four-model maximum is a representative empirical risk, not the minimax risk over the full calibrated model class. The Monte Carlo envelopes quantify simulation uncertainty within each setting; they are not simultaneous over the 30 settings and are not additional theoretical guarantees. The bound comparison is a consistency check; the minimax characterization is given by Thm. 8.

J Additional Code-Generation Results

The independent-reference curves in Fig. 8 include all eight configurations, using the ordering and encoding of the main figure. Tab. 1 gives endpoint comparisons for both designs. LCB abbreviates LiveCodeBench.

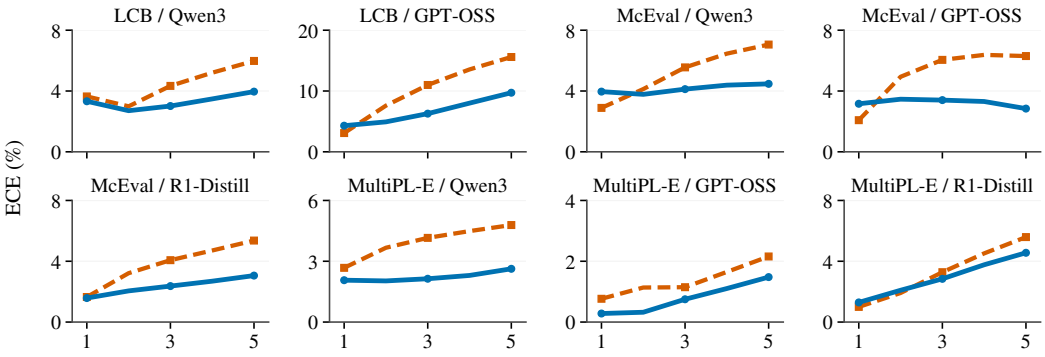


Figure 8: Rank gains with five disjoint rank references and up to five selection candidates. ■ dashed: score-only; ● solid: score+rank. Curves show held-out ECE against selection budget N in the independent $5 + 5$ design. LCB: LiveCodeBench. Both maps are selected using only $N = 1$ Brier score. All configurations and budgets are shown; vertical ranges differ by panel and start at zero.

	Fixed pool ($N = 10$)		Independent $5 + 5$ ($N = 5$)	
Configuration	Score → rank	Gain [95% CI]	Score → rank	Gain [95% CI]
LCB / Qwen3	$9.37 \to 4.63$	$4.74\ [-1.63, 8.97]$	$5.98 \to 3.96$	$2.02\ [-2.40, 4.25]$
LCB / GPT-OSS	$20.34 \to 7.05$	$13.30\ [7.88, 14.77]$	$15.59 \to 9.72$	$5.87\ [0.56, 6.76]$
McEval / Qwen3	$8.48 \to 3.94$	$4.54\ [-1.23, 5.89]$	$7.06 \to 4.47$	$2.58\ [-3.30, 3.89]$
McEval / GPT-OSS	$8.51 \to 3.00$	$5.51\ [-2.49, 8.82]$	$6.30 \to 2.84$	$3.46\ [-1.94, 4.33]$
McEval / R1-Distill	$6.42 \to 4.57$	$1.85\ [-1.91, 4.22]$	$5.36 \to 3.06$	$2.30\ [-2.05, 2.94]$
MultiPL-E / Qwen3	$6.53 \to 2.36$	$4.17\ [-0.34, 5.89]$	$4.79 \to 2.63$	$2.16\ [-1.78, 3.79]$
MultiPL-E / GPT-OSS	$3.44 \to 2.10$	$1.34\ [-1.34, 1.92]$	$2.16 \to 1.48$	$0.68\ [-1.44, 1.85]$
MultiPL-E / R1-Distill	$5.99 \to 2.90$	$3.09\ [-0.85, 5.15]$	$5.59 \to 4.56$	$1.03\ [-0.78, 1.49]$

Table 1: Complete high-budget comparisons. Score → rank gives ECE in percent; gain is score-only ECE minus rank-aware ECE in percentage points. Intervals are 95% paired percentile bootstrap intervals from 4000 resamples of final problem groups, conditional on the fitted reports. MultiPL-E resampling keeps translations of the same underlying problem together.

The endpoint ECE gain is positive in all eight configurations under each design. The intervals are pointwise and conditional on the fitted reports. All reports share the selected candidates and accuracy.

Brier score improves in all eight fixed-pool endpoint comparisons and seven independent-reference comparisons; the exception is MultiPL-E/R1-Distill (0.16401 to 0.16475). At the 5% ECE threshold, the fixed-pool budget prefix expands in seven configurations; MultiPL-E/GPT-OSS already meets the threshold throughout for both reports. These are observed ranges, not certified budget guarantees.

K Empirical Protocol

This section describes the protocol for the eight code-generation configurations in Sec. 5.3.

K.1 Generations and Eight Configurations

We use generations from `lavis-nlp/CALIBRI`. LiveCodeBench contributes Qwen3 and GPT-OSS; McEval and MultiPL-E each contribute Qwen3, GPT-OSS, and R1-Distill. The abbreviated model names refer to Qwen3-Coder-30B, GPT-OSS-20B, and R1-Distill-Qwen-32B. There is no LiveCodeBench/R1-Distill configuration in this release.

Every prompt has ten candidates. Correctness is the binary test outcome (Ni et al., 2024), and the score is the exponential of the mean token log probability (Detommaso et al., 2024; Spiess et al., 2025), including reasoning tokens where present.

K.2 Splits and Fitted Reports

Official training rows are used for fitting. Official validation and test rows are pooled, their sorted problem-group identifiers are permuted, and the first half of the groups form an audit set; the remaining groups form the final set used in the figures. Neither held-out set tunes the reports. MultiPL-E groups all translations of the same HumanEval (Chen et al., 2021) base problem together. LiveCodeBench and McEval use the published prompt identifiers; McEval's numeric suffixes are language-local identifiers rather than a translation alignment.

Benchmark	Train rows / groups	Audit rows / groups	Final rows / groups
LiveCodeBench	527/527	264/264	264/264
McEval	853/853	427/427	427/427
MultiPL-E	1814/80	904/40	943/41

Counts are identical across models within each benchmark. Training folds and held-out splits are disjoint at the group level.

For each configuration and pool design, the score-only and score+rank reports receive the same candidate labels and search over eight model configurations. Six use logistic regression on cubic spline features with rank-knot count $k \in \{4, 6, 8\}$ and inverse regularization $C \in \{0.1, 10\}$. The joint map includes score and rank spline features and their products; the score-only map uses $(k+1)^2 + 2(k+1) - 1$ score knots to match feature capacity. Two use histogram gradient boosting (Friedman, 2001) with ℓ_2 regularization in $\{1, 10\}$, at most seven leaves, 160 iterations, learning rate 0.05, and minimum leaf size 20. Input scores and ranks are clipped before a logit transformation, at $[10^{-6}, 1 - 10^{-6}]$ and $[0.01, 0.99]$, respectively.

Each report selects its model by candidate-averaged out-of-fold Brier score at $N = 1$, using three group-disjoint folds. The selected model is trained on all training candidates and frozen before held-out evaluation. No $N > 1$ metric enters model selection.

K.3 Pool Designs

In the fixed-pool design, candidates are ordered by original score with independent fixed random keys to break ties. A candidate with one-based rank r among the ten scores has feature $\widehat{U} = (r - 1/2)/10$. This feature uses the whole pool and remains fixed as the selection-set size varies over $N = 1, \ldots, 10$.

In the independent $5 + 5$ design, a fixed random permutation splits each pool into five selection candidates and five disjoint rank references. If R_{ref} references fall below a candidate in the same score-and-tie ordering, its feature is $\widehat{U} = (R_{\text{ref}} + 1/2)/6$. Reference labels are not used to fit either report. The selection budget varies over $N = 1, \ldots, 5$, with the partition and reference features held fixed.

In either design, selection always uses the original score, so the raw, score-only, and rank-aware reports share the selected candidates and accuracy. For a selection pool of size L, a candidate of one-based rank r receives winner weight $\binom{r-1}{N-1}/\binom{L}{N}$, interpreted as zero for $r < N$. These weights

average exactly over all size-N subsets. Subsets are not treated as independent prompt samples. The main plot shows all ten budgets of the fixed-pool comparison; Fig. 8 shows all five budgets with disjoint references.

K.4 Metrics and Paired Uncertainty

We average the subset-weighted statistics equally across final prompt rows. Fixed-bin ECE (Roelofs et al., 2022) is computed by taking the absolute value of each pooled bin residual and summing over bins with edges $0, 0.2, \ldots, 1$. It is not the average of per-prompt ECE values. Accuracy, mean confidence, signed residual, and Brier score are also recorded. The figure lines are the held-out point estimates, with no smoothing or omission of intermediate budgets.

The endpoint gain in Tab. 1 is score-only ECE minus rank-aware ECE. Its 95% interval uses 4000 paired percentile bootstrap resamples of final problem groups, with the same resampled groups for both reports. MultiPL-E translations are resampled together; statistics are divided by the sampled row count to retain the equal-row target. These pointwise intervals are conditional on the fitted reports and do not include refitting or supply simultaneous calibration certificates. The observed 5% ECE range is the largest contiguous budget prefix whose point estimates all meet that threshold.

K.5 Generation Costs

The following table distinguishes the information schemes studied in the theory.

Information or experiment	Observations	Labels and generation cost
Calibration constraints	Calibration at $1, \ldots, K$; no observed U required	Acquisition cost not modeled; population equalities are supplied as assumptions.
Noisy moments	Simultaneous bounds ϵ_n	Depends on the procedure producing the intervals.
Joint/passive oracle	m ordinary labels and true U; joint also knows the calibration constraints	m labels and base candidates; obtaining the true CDF is not costed.
Complete pools	n_{pool} labeled candidates per prompt	mn_{pool} labels and candidates; direct estimation for $N \leq n_{\text{pool}}$.
Independent ranks	One label and k independent references per prompt	m labels, $m(k+1)$ candidates; unbiased rank weights for $N \leq k+1$.

Additional references used by a predictor are counted separately from references used to estimate residual moments. These cost models do not change the accounting of the pool experiment: all ten candidates per prompt already exist, and N denotes subset selection size rather than total generation cost. Each fitted report uses 10 training labels per row in the fixed-pool design and 5 in the independent design; the latter also uses five reference scores per row without their labels. Within a design, both reports receive exactly the same fitting labels.

K.6 Rank Permutation Control

Sec. 5.4 uses the LiveCodeBench/Qwen3 independent-reference comparison. Its 527 training prompts supply 2635 selection-candidate labels; the final set has 264 prompts. The score-only and true-rank predictions are reused. Permuted-rank reports use the same three grouped cross-validation folds and eight candidate model configurations, selected by $N = 1$ Brier score. Score-decile boundaries are estimated on the training data. Ranks are permuted within these strata across different prompts, separately within each cross-validation fitting or validation split and within final fitting or evaluation data. The rank multiset is preserved in each stratum, and a given permutation stays fixed across budgets. All reports use the original score-based selection rule.

K.7 Synthetic Experiment Details

The synthetic experiments are specified in Secs. 5.1 and 5.2. The first compares the known calibrated construction with finite-sample ECE estimates over 20 settings, with 100 repetitions of 10^4 prompts

per setting. The second evaluates the dual estimator across four generating models and 30 settings, with 1000 repetitions per setting; all settings and the bound comparison are given in App. I. The displayed synthetic error bars describe Monte Carlo variability, rather than an empirical coverage study of the real-data bootstrap intervals.