THE PRICE OF TEST-TIME SELECTION: AN ACCURACY–EVALUATION FRONTIER

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting a better model response can make its accuracy harder to measure: the favored responses may be rare among previously labeled outputs. We characterize the accuracy–evaluation tradeoff when selection returns one of N candidates and evaluation reuses L labels from the base generator. Candidate availability determines exactly which output distributions are reachable. We measure their evaluation difficulty with a bounded audit profile that matches finite-sample minimax error up to universal constants. Optimizing accuracy subject to this profile gives an exact frontier and a selection rule that discounts correctness in regions with poor label coverage. Continuous-score best-of-N incurs worst-case risk $\Theta(1 \wedge N/L)$, so fixed worst-case precision requires labels proportional to the candidate budget. We turn feasible finite approximations into executable selection rules, including on banks with tied scores. Experiments on four frozen reasoning banks test the evaluation law and the resulting selectors. At matched audit-radius budgets, the policies improve accuracy over uniform–BoN mixtures by 2.92 percentage points on average and closely match fully optimized order-statistic and quadratic baselines. This connects candidate compute, selected accuracy, and the cost of substantiating that accuracy from reusable labels. Code and data: `https://anonymous.4open.science/r/selection-cost`.

1 INTRODUCTION

Selecting among model responses can improve accuracy without changing model parameters. Best-of-N, verifier-based reranking, self-consistency, and minimum-risk decoding use different signals to decide which candidate to return (Cobbe et al., 2021; Wang et al., 2023; Eikema & Aziz, 2022; Snell et al., 2025). Their value is usually measured by accuracy versus inference compute. But establishing that accuracy can require a second resource: labels that adequately cover the selected responses.

For a single fixed system, one can label fresh selected outputs directly. Comparing several prespecified verifiers or candidate budgets for the same generator creates a different opportunity: a common bank of labeled base responses can support each evaluation without collecting new labels for every choice. The value of this reuse depends on which responses selection favors. An answer that is rare under the generator can become common among selected outputs, so the same bank provides less information about the new system. We call evaluation from these reusable base labels an *audit*.

Figure 1 makes this cost visible across four tasks. At the largest candidate budgets, best-of-N improves accuracy by 5.9–26.1 percentage points, but the mean-squared error (MSE) of evaluation from $L = 64$ base labels is 2.6–37.3 times its value at $N = 1$. [1] Selection improves the answers while making their accuracy less precisely estimated from the existing labels.

A change of distribution explains both effects. For one prompt, let P be the base response law and Q the law after selection. The density ratio $w = dQ/dP$ measures how much selection amplifies each region: a response set with 1% of base probability and 10% of selected probability has average amplification ten. If $\eta(z)$ is the probability that response z is correct, selected accuracy is

$$A_w = \int \eta(z) w(z)\, dP(z). \tag{1}$$

[1] Both axes use the same frozen candidate distribution and with-replacement selection protocol; audit MSE uses $L = 64$ independent base labels and the exact discrete selection weights.

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.

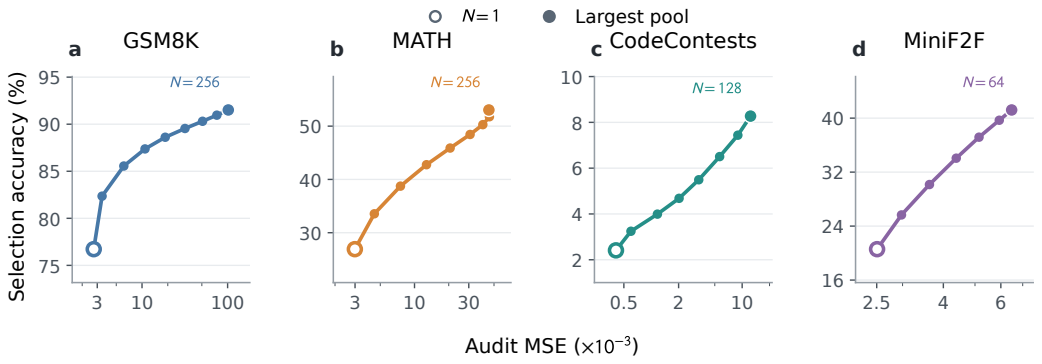


Figure 1: **Better selected answers can be harder to evaluate from old labels.** Best-of-N accuracy versus observed audit MSE using 64 total base labels per task. Paths follow powers-of-two candidate budgets; open markers denote $N = 1$ and large filled markers $N_{\max} = 256, 256, 128, 64$, respectively. Horizontal axes are logarithmic; vertical scales differ.

The same amplification that raises accuracy also gives more influence to responses that may seldom appear in the audit sample.

We ask: *with N candidates and L reusable labels, how much accuracy is attainable at a given evaluation precision?* The *accuracy–audit frontier* answers this question over all laws obtainable by returning one of the candidates. It combines two ingredients. A selected response set cannot be more likely than the event that the pool contains one of its members; these constraints exactly describe the reachable laws (Theorem 3.1). An *audit profile* then measures how difficult each law is to evaluate from L base labels, matching minimax MSE up to universal constants (Theorem 4.2).

This yields a distinction between ranking quality and concentration. Better ranking determines *where* selection places probability; concentration governs its worst-case evaluation cost. All continuous-score best-of-N rules have the same audit profile at fixed N and L, even with very different accuracies. They reach the most concentrated boundary permitted by the candidate budget, with worst-case error $\Theta(1 \wedge N/L)$. A better verifier alone therefore cannot remove the linear label cost of extreme selection.

The frontier makes evaluation difficulty a constraint on selection itself (Theorem 5.2). Its dual favors correct responses while discounting regions with poor label coverage. An interior optimum can require randomized selection, rather than simply a smaller best-of-N budget. A certified finite approximation and an explicit implementation turn feasible distributions into rules returning an actual candidate.

Controlled experiments test the evaluation law, and cross-fitted selectors on GSM8K, MATH, CodeContests, and MiniF2F test its use in policy design. At matched audit budgets, optimizing amplification improves on simple mixture and rank-power rules; richer order-statistic and quadratic policies recover nearly all of the same accuracy. These comparisons distinguish the benefits of flexible selection from those of the finite-sample constraint.

2 Related Work

Theories of test-time selection analyze coverage, verifier error, ROC curves, and reward or win-rate tradeoffs (Huang et al., 2025; Dorner et al., 2026; Gui et al., 2024; Beirami et al., 2025; Mroueh & Nitsure, 2025). These complement empirical studies of sampling and inference scaling (Brown et al., 2024; Schaeffer et al., 2025; Snell et al., 2025). Proximity regularization, including MBR-BoN (Jinnai et al., 2025), aims to limit reward overoptimization (Gao et al., 2023). Our constraint instead measures the labels needed to evaluate selected outputs, even when their gold accuracy improves.

For a fixed selector, evaluation from base labels is an off-policy evaluation problem. Importance weighting and clipping are established tools in this setting and under covariate shift (Ionides, 2008; Li et al., 2015; Wang et al., 2017; Ma et al., 2022; Morrison et al., 2024; Zhang et al., 2025a). Our target law is itself a design variable: we choose how to redistribute probability within what a finite

candidate pool permits, subject to the risk of evaluating that choice. Thus the optimization couples policy value, label coverage, and candidate availability.

Random-set selection (Artstein, 1983) supplies the realizability criterion; majorization (Marshall et al., 2011) and L^1–L^2 interpolation (Holmstedt, 1970) describe concentration and finite-sample risk. Their joint optimization yields the frontier and its selection rule. Appendix C gives the extended comparison.

3 Selection-Induced Distributions

Let $X \sim P_X$ be a prompt and $Z \mid X = x \sim P_x$ a base response. Candidate pairs (Z_i, Y_i) are conditionally i.i.d., with binary gold correctness Y_i. Write

$$\eta_x(z) = \mathbb{P}(Y = 1 \mid X = x, Z = z), \qquad \mu(dx, dz) = P_X(dx)P_x(dz). \tag{2}$$

We assume standard Borel prompt and response spaces. For the continuous rearrangement statements, P_x is nonatomic for almost every prompt. Section 5.1 treats atomic candidate banks directly.

An N-*candidate selector* observes $(X, Z_{1:N})$ and returns an index $J \in [N]$, possibly at random, without observing $Y_{1:N}$. It is permutation equivariant without loss of generality, since symmetrization preserves its output distribution. Its *selected law* $Q_{\pi,x}$ is absolutely continuous with respect to P_x: a base-null set cannot occur in the pool. Define its amplification and accuracy by

$$w_{\pi,x} = \frac{dQ_{\pi,x}}{dP_x}, \qquad A_\pi = \mathbb{P}(Y_J = 1) = \int \eta w_\pi \, d\mu. \tag{3}$$

Thus accuracy depends on the selector through its output law. We can study selection by characterizing these laws, without enumerating selection rules.

To describe the feasible laws, we compare how much probability densities assign to their most amplified regions. Let $w^\downarrow$ denote the decreasing rearrangement of w. We say w is *majorized* by a decreasing density g, written $w \prec g$, when they have equal total mass and $\int_0^t w^\downarrow(s)\, ds \le \int_0^t g(s)\, ds$ for every $t \in [0, 1]$.

Theorem 3.1 (Exact laws reachable with N candidates)**.** *Fix a prompt and write $P = P_x$. A probability law Q is realizable by returning one of N i.i.d. draws from P if and only if, for every measurable B,*

$$Q(B) \le 1 - (1 - P(B))^N. \tag{4}$$

Equivalently, its density satisfies

$$w \prec g_N, \qquad g_N(s) = N(1 - s)^{N-1}. \tag{5}$$

Necessity is immediate: the pool must contain a member of B before one can be returned. Sufficiency shows there are no further distributional restrictions. The criterion also holds jointly across prompts: every Borel kernel obeying it almost surely has a jointly measurable realization (Appendix D). We write

$$\mathcal{K}_N = \{w : w_x \prec g_N \text{ for almost every } x\} \tag{6}$$

for the reachable densities. For every convex Φ with existing integrals,

$$\int \Phi(w_x)\, dP_x \le \int_0^1 \Phi(g_N(s))\, ds. \tag{7}$$

In particular, $\|w_x\|_\infty \le N$ and $\|w_x\|_2^2 \le N^2/(2N - 1)$, sharply. Continuous-score best-of-N has rank density $w_N(u) = Nu^{N-1}$, whose sorted values equal g_N. It attains every convex concentration bound above: it puts as much weight on rare base responses as the candidate budget permits. Thus two scores can induce equally concentrated laws while favoring different responses. Their accuracy depends on the alignment of this amplification with correctness. We next ask how this concentration affects evaluation from base labels.

4 EVALUATION FROM REUSABLE LABELS

Fix w independently of L i.i.d. labeled base draws $D_L = (X_i, Z_i, Y_i)_{i=1}^L$. The base law μ and the selected density w are known.[2] With no structural assumptions on $0 \le \eta \le 1$, the optimal worst-case mean-squared error for estimating $A_w = \int \eta w \, d\mu$ is

$$R_L^*(w) = \inf_{\widehat{A}} \sup_{0 \le \eta \le 1} \mathbb{E}_\eta (\widehat{A} - A_w)^2. \tag{8}$$

The second moment $m_2 = \|w\|_2^2$ gives the familiar variance proxy m_2/L. At small L, a large weight on a tiny region can instead be clipped with little bias. If its base probability is p with $Lp \ll 1$, the audit bank rarely observes it: correctness functions differing only there are hard to distinguish, yet their selected accuracies can differ by its entire selected mass. The following profile captures these unresolved correctness differences.

Definition 4.1 (Audit profile)**.** For a selected-law density w and label budget L, let

$$\mathcal{H}_L = \{h : 0 \le h \le 1, \ \|h\|_2 \le L^{-1/2}\}, \qquad \omega_L(w) = \sup_{h \in \mathcal{H}_L} \int wh \, d\mu. \tag{9}$$

The *audit profile* is $\mathcal{E}_w(L) = \omega_L(w)^2$.

Here h represents a difference between possible correctness functions. Its L^2 constraint makes this difference hard to resolve from L base observations, while $0 \le h \le 1$ enforces bounded correctness. Thus $\omega_L(w)$ measures the largest selected-accuracy signal that can remain hidden in the audit sample. Clipping gives a complementary way to quantify this difficulty:

$$\rho_L(w) = \inf_{\tau \ge 0} \left\{ \frac{1}{L} \int (w \wedge \tau)^2 \, d\mu + \left[\int (w - \tau)_+ \, d\mu \right]^2 \right\}. \tag{10}$$

The two terms control variance after clipping and squared bias from the discarded selected mass.

Theorem 4.2 (Finite-sample evaluation law)**.** *There is a universal $c > 0$ such that every density w and $L \ge 1$ satisfy*

$$\begin{aligned} c\,\omega_L(w)^2 \le R_L^*(w) \le \tfrac{1}{4}\rho_L(w) \le \tfrac{1}{4}\omega_L(w)^2, \\ \rho_L(w) \le \omega_L(w)^2 \le 2\rho_L(w). \end{aligned} \tag{11}$$

For any data-independent $\tau \ge 0$, the centered clipped estimator

$$\widehat{A}_{w,\tau} = \Pi_{[0,1]} \left[\frac{1}{2} + \frac{1}{L} \sum_{i=1}^{L} (w(X_i, Z_i) \wedge \tau)(Y_i - \tfrac{1}{2}) \right] \tag{12}$$

has worst-case risk at most one quarter of the corresponding objective in (10)*.*

The upper bound is an attainable bias–variance tradeoff. The lower bound shows why a different estimator cannot remove poor coverage: two correctness functions can remain indistinguishable in the base sample while giving different selected accuracies (proof in Appendix E). Centering at $1/2$ uses the known range of correctness. Raising τ retains more selected mass at the cost of larger weights; optimizing it balances bias and variance using the selected law alone. When $\|w\|_\infty \le \sqrt{L}\,\|w\|_2$, the witness constraint $h \le 1$ is inactive and the profile reduces to

$$\mathcal{E}_w(L) = \frac{\|w\|_2^2}{L} = \frac{1 + \chi^2(Q\|\mu)}{L}. \tag{13}$$

Outside this regime the profile stays bounded and rate-sharp, even for unbounded density ratios.[3]

Corollary 4.3 (The audit cost of extreme selection)**.** *Every $w \in \mathcal{K}_N$ satisfies*

$$\omega_L(w) \le \omega_L(g_N), \qquad R_L^*(w_N) = \Theta(1 \wedge N/L). \tag{14}$$

Every continuous-score extreme-order selector attains the profile bound. If $\eta_x(Z)$ is atomless for almost every prompt, the unique accuracy oracle returns $\arg\max_{i \le N} \eta_X(Z_i)$ and maximizes the entire audit profile over $\mathcal{K}_N$.

[2] Appendix L develops a black-box replay extension; Figure 7 reports its coefficient diagnostics.

[3] Appendix I gives the reachable-law construction in Figure 3: the quadratic proxy stays constant while minimax risk vanishes. It also quantifies the resulting frontier gap.

For best-of-N, worst-case error stays bounded away from zero when $L \ll N$ and falls as N/L once labels cover the concentration created by selection. Maintaining worst-case precision therefore requires labels proportional to N. The no-ties condition matters at the accuracy endpoint: if correctness is constant, all selectors have the same accuracy but uniform selection is easiest to evaluate.[4]

5 The Accuracy–Audit Frontier

The preceding results separate two constraints: which output laws a candidate budget permits, and how precisely a label bank can evaluate them. Their intersection determines the best accuracy available at a given audit cost.

Definition 5.1 (Accuracy–audit frontier)**.** For candidate budget N, label budget L, and audit radius $r \geq 0$, define

$$\mathcal{F}_{N,L}(r;\eta) = \sup\left\{ \int \eta w \, d\mu : w \in \mathcal{K}_N,\ \omega_L(w) \leq r \right\}. \tag{15}$$

The vertical coordinate is accuracy; the squared horizontal coordinate r^2 bounds worst-case evaluation error through Theorem 4.2. Uniform selection has radius $r_0 = L^{-1/2}$, while continuous-score best-of-N has the largest radius $r_N = \omega_L(g_N)$. To guarantee worst-case root-MSE at most ε with the optimally clipped estimator, it suffices to impose $r \leq 2\varepsilon$. Increasing L improves coverage of a fixed selected law; changing the selector reallocates probability to suit existing coverage, subject to the candidate budget N.

For a bounded score f, let

$$\mathcal{M}_N(f) = \mathbb{E}_{X,Z_{1:N}}\left[\max_{i\leq N} f_X(Z_i)\right], \tag{16}$$

the largest expected score achievable by choosing one candidate.

Theorem 5.2 (Optimal selection at a fixed audit radius)**.** *Let $N \geq 2$, $L > 1$, and $r > r_0$. The primal and dual optima are attained, and*

$$\mathcal{F}_{N,L}(r;\eta) = \inf_{\substack{\lambda\geq 0 \\ h\in\mathcal{H}_L}} \{\lambda r + \mathcal{M}_N(\eta - \lambda h)\}. \tag{17}$$

An optimal primal–dual triple (w^*, λ^*, h^*) has an N-pool realization returning only candidates in

$$\operatorname*{argmax}_{i\leq N}\{\eta_X(Z_i) - \lambda^* h^*_X(Z_i)\}. \tag{18}$$

At $r = r_0$ the unique feasible law is $w \equiv 1$. For $r \geq r_N$, the audit constraint is redundant and $\mathcal{F}_{N,L}(r;\eta) = \mathcal{M}_N(\eta)$.

The proof exchanges the reachable-law and audit-witness optimizations. Poolwise maximization gives the inner value, and strict feasibility of uniform selection gives strong duality for $r > r_0$.

The optimal score discounts correctness in directions that base labels cannot resolve well. The function h^* identifies those directions; its multiplier λ^* is an *audit price*. A loose constraint permits $\lambda^* = 0$ and recovers the accuracy oracle. A tighter constraint can favor a less concentrated output distribution, even when its unconstrained accuracy is lower.

Tie probabilities in (18) must realize an optimal feasible density. Interior solutions can require randomization, so they need not be weaker best-of-N rules. For one prompt with atomless correctness, the adjusted score is flat where the capacity constraint is slack and the density is affine in correctness there.[5]

The frontier is also computable with a controlled approximation error. For one prompt, restricting to M equal correctness-quantile bins while retaining both constraints gives an optimum $\mathcal{F}^{(M)}_{N,L}$ satisfying

$$0 \leq \mathcal{F}_{N,L}(r;\eta) - \mathcal{F}^{(M)}_{N,L}(r;\eta) \leq \min\{1, N/M\}. \tag{19}$$

[4] Appendix F proves the corollary and gives the exact transition at $L = 2N - 1$, plotted in Figure 8.

[5] Appendix H gives the interior structure. Appendix G proves the duality and compares profile-constrained and exact-minimax-risk frontiers; their horizontal scales agree up to universal constants.

Every feasible grid density remains realizable. Independent pilot data can supply an estimated correctness function, with plug-in regret controlled by its estimation error (Proposition J.2). The grid certificate controls numerical approximation in correctness quantiles; the experiments instead fit correctness within a verifier-rank class. The pilot labels used for fitting are separate from the L audit labels.

5.1 Executing the frontier on finite candidate banks

A feasible output law must also be turned into a rule that returns an actual pool member. Finite banks make this step concrete, while requiring explicit treatment of atomic probabilities and score ties.

Proposition 5.3 (Finite-bank implementation)**.** *For base probabilities p_j, an output probability vector q is achievable from N i.i.d. candidates if and only if*

$$q(A) \le 1 - (1 - p(A))^N \qquad \text{for every subset } A, \tag{20}$$

where $p(A)$ and $q(A)$ are probability masses. Every feasible law is a mixture of priority selectors. On M equal-probability rank bins, a feasible bin-constant density admits at most $M - 1$ randomized coordinate exchanges, with $O(M + N)$ execution time and $O(M)$ storage after preprocessing, excluding generation and scoring.

Classical majorization (Marshall et al., 2011) supplies the exchange factorization. Each pool receives fresh independent coins that sample a priority order independently of its candidates; the selector returns the first occurrence in the highest-priority present bin. Averaging over these orders yields exactly the target bin masses. Appendix K gives the construction and proofs.

For score CDF F_S, each candidate occurrence receives the uniform randomized rank $U = F_S(S-) + V\{F_S(S) - F_S(S-)\}$. The auxiliary $V \sim \mathrm{Unif}[0,1]$ is independent of candidates and priority coins, and fresh for every occurrence, including repeated records. For record j with score s_j, let $a_j = F_S(s_j-)$ and $c_j = F_S(s_j)$; the interval length is the base probability mass of its entire score tie. Its exact bank marginal is

$$\bar{w}_j = \frac{1}{c_j - a_j} \int_{a_j}^{c_j} w(u)\, du, \qquad Q(Z = j) = p_j \bar{w}_j. \tag{21}$$

These probabilities independently check poolwise accuracy. Marginalization gives $\omega_L(\bar{w}) \le \omega_L(w)$, so rank radius r certifies finite-bank clipped-estimator MSE at most $r^2/4$. The guarantee is sufficient: averaging within ties can strictly reduce audit difficulty.

The experiments fit correctness on other folds, then execute held-out policies using scores and fresh randomness, without gold labels. Their curve concerns an implementable fitted verifier-rank class, not the unrestricted correctness-oracle frontier in (15).

6 Experiments

We first test whether the audit profile captures finite-sample evaluation error. We then study how an audit budget changes selection on four reasoning tasks, measuring both held-out accuracy and realized audit error.

6.1 Data and evaluation protocol

The public MONKEY BUSINESS bank (Schaeffer et al., 2025) supplies Llama-3-8B-Instruct responses (Dubey et al., 2024) for GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), CodeContests (Li et al., 2022), and MiniF2F (Zheng et al., 2022). The banks contain 183,040 responses across 525 task–problem rows (524 semantic clusters), with 128–512 candidates per problem. Math selection uses Skywork-o1-Open-PRM-Qwen-2.5-1.5B; code selection uses public tests and is judged on hidden tests; proof selection uses a one-second Lean budget and is judged under the published ten-second budget. Scores and operational correctness labels are frozen throughout.[6]

[6]PRM aggregation was chosen on a design fold. Appendix B.3 reports its provenance and comparisons excluding those design questions (Table 7).

In five-fold cross-fitting, each policy uses a tie-integrated isotonic correctness estimate from the other four folds. On held-out problems, it selects from N with-replacement candidates using scores and fresh randomness; the returned label is read after selection. The random-priority implementation uses $M = \max\{512, 32N\}$ rank bins; independent marginal and Monte Carlo checks verify that it executes the intended law.[7]

The policy grid is $N \in \{8, 32, 64\}$, $L \in \{32, 64, 128\}$, and five radii between uniform selection and the bin-integrated BoN endpoint. Nine policies give 8,100 policy–fold settings, with 8,192 pool trials per problem–policy setting. Uncertainty is measured across problems, not pool trials: intervals use paired problem bootstrapping, conditional on the fitted policies. The shared MATH–MiniF2F problem receives shared bootstrap weights.

For a task-level audit, L is the *total* number of labels drawn from the joint prompt–candidate base law, with prompts equally weighted. The design radius r is imposed foldwise. Pooling the held-out folds preserves the clipping guarantee $\rho_L(w_{\text{joint}}) \le \sum_f p_f \rho_L(w_f) \le r^2$, where p_f is the fraction of test problems in fold f. The joint estimator retains worst-case MSE at most $r^2/4$, although its pooled ω_L can exceed r (Lemma K.5). Appendix A specifies the full protocol.[8]

6.2 The cost predicted by the evaluation law

Figure 1 shows an evaluation cost accompanying selection gains on all four tasks. Its size varies: at the largest candidate budgets, audit MSE rises by factors of 37.3, 15.1, 30.2, and 2.6 on GSM8K, MATH, CodeContests, and MiniF2F, respectively (Table 3). These are realized errors under each bank's correctness labels; the theory controls the worst case over correctness.

To isolate the finite-sample law from errors in density estimation or correctness calibration, we use controlled distributions. The selected densities include uniform selection, BoN, their equal mixture, a median order statistic, and a rank-power law. Correctness is constant, monotone, logistic, or a least-favorable alternative. Figure 2 compares the projected centered-clipped estimator with its theoretical upper bound. Across 2,375 estimator settings, none exceeds the bound by more than five Monte Carlo standard errors; the median MSE-to-ω_L^2 ratio is 0.209. A separate two-point Bayes experiment tests the lower-bound construction, with median Bayes-risk-to-ω_L^2 ratio 0.050.

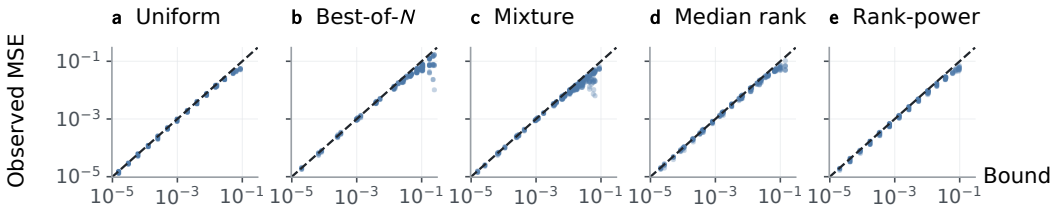


Figure 2: **The finite-sample envelope tracks attainable audit error.** Observed MSE of the projected centered-clipped estimator versus its upper bound in Theorem 4.2, on log–log axes. Panels separate selector families across correctness functions and budgets. Dashed lines denote equality; the mixture weights uniform selection and BoN equally.

6.3 Accuracy under a common audit-radius budget

We next ask whether optimizing the selected distribution improves accuracy within an audit budget. We compare the implemented frontier policy with four principal alternatives. A *uniform–BoN mixture* randomizes between a uniform pool member and the score winner. *Rank-power* searches a small family of power-law rank densities, each mixed with uniform selection to meet the budget. The *full order-statistic hull* optimizes a convex mixture over all N rank orders. The *quadratic frontier* optimizes the same grid class as our policy with the sufficient constraint $m_2/L \le r^2$ in place of

[7] Independent reconstruction matches target marginals within 2.8×10^{-13} in L^1; Table 8 gives optimization and execution checks.

[8] Tables 1 and 2 give bank sizes and replication budgets.

The distinction from a quadratic constraint is clearest in the analytic rare-region construction of Proposition I.1, illustrated in Figure 3. There, the quadratic proxy remains at least $1/4$ while minimax risk is $\Theta(1/L)$. The finite-sample profile recognizes that the troublesome region carries little selected mass and can be clipped cheaply.

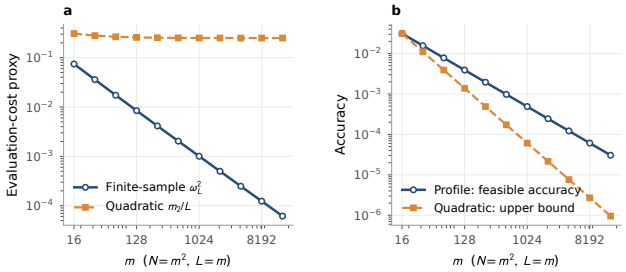


Figure 3: **A large second moment need not imply a large finite-sample cost.** Analytic rare-region construction with $N = m^2$, $L = m$ (Proposition I.1). (a) Exact evaluation-cost proxies. (b) At $r = 2/\sqrt{m}$, profile-feasible accuracy versus an upper bound on every quadratic-constrained policy.

$\omega_L(w) \le r$. All share the independently fitted correctness curve and per-fold (N, L, r) budget. Uniform selection and ordinary BoN are endpoint references; ordinary BoN need not be feasible at an interior audit radius.

Figure 4 shows held-out poolwise accuracy along the five audit budgets at $N = 32$ and $L = 64$. The greatest gains over simple mixtures appear in the mathematical tasks and MiniF2F. CodeContests shows less separation, consistent with its smaller gains in the full-grid comparison.

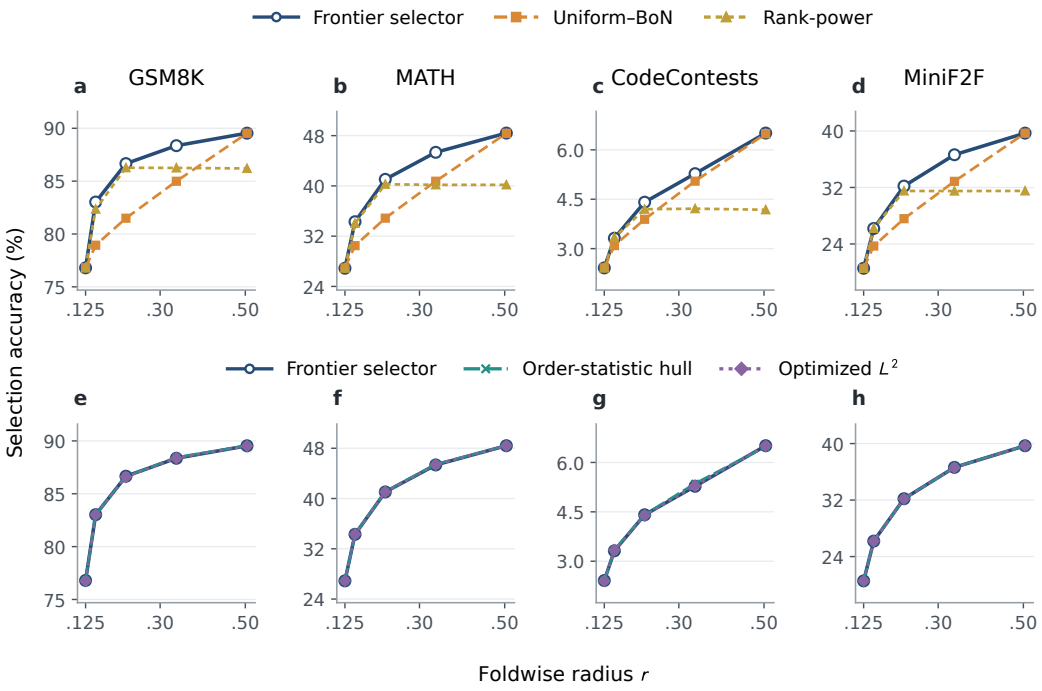


Figure 4: **Feasible distributions yield useful poolwise selection rules.** Held-out poolwise accuracy versus foldwise audit-radius budget r at $N = 32$, $L = 64$. Top: simple selector families. Bottom: fully optimized alternatives. Each path joins five budgets, averaging the cross-fitted policies by held-out problem count.

For the aggregate comparison, we first average the 27 interior settings within each problem, then average problems within each task and give the four tasks equal weight. Appendix Figure 6 reports the paired differences. The frontier gains 2.924 percentage points over uniform–BoN mixtures (95% interval $[2.480, 3.390]$) and 1.254 points over rank-power ($[1.051, 1.474]$). Against the full order-statistic hull, the difference is -0.024 points ($[-0.042, -0.006]$); against the quadratic frontier

it is -0.001 points ($[-0.012, 0.011]$). The all-order hull has a small but resolved aggregate advantage. Tables 4 and 5 report the full comparisons.

The gains over fixed mixtures and rank-power show the value of optimizing where selection places probability. The richer order-statistic class captures nearly all of this benefit. The quadratic constraint also recovers nearly the same accuracy: 500 of the 540 fitted interior frontier policies lie in the equality regime of (13). The two cost expressions therefore coincide for most fitted policies; Appendix B.4 connects this identity to agreement of the fitted frontiers. Together with the rare-region construction, this identifies when quadratic control is sufficient and when it can be overly restrictive.

6.4 From audit guarantees to realized error

Figure 5 compares the same frontier and uniform–BoN policies using joint audit MSE. At the middle radius with $N = 32, L = 64$, the MATH frontier has exact finite-bank accuracy 41.03% and observed MSE 0.01009, compared with 34.86% and 0.00646 for the uniform–BoN mixture. Both meet the same worst-case budget, but the higher-accuracy frontier policy has higher realized error. The guarantee and the task-specific tradeoff answer different questions.[9]

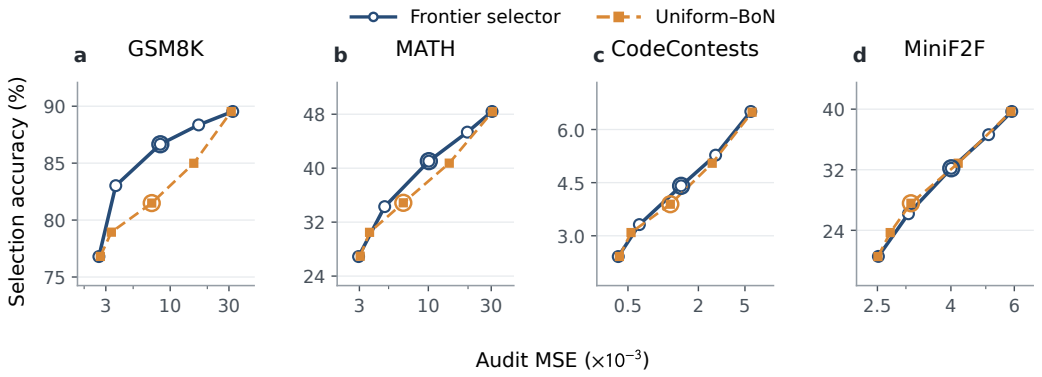


Figure 5: **A common guarantee leaves room for different realized errors.** The frontier and uniform–BoN mixture from Figure 4, plotted against observed MSE of the joint task-level audit using $L = 64$ total base labels (horizontal log scale, displayed in units of 10^{-3}). Both coordinates use the same finite-bank output law; highlighted points mark the middle audit-radius budget.

Score ties further separate design and realized costs. At the middle budget, integrating out tie randomness leaves the two math-bank radii nearly unchanged but reduces them substantially on CodeContests and MiniF2F (Table 6). The contraction in Section 5.1 explains why a valid continuous-rank certificate can correspond to an easier finite-bank audit.

7 Discussion

The frontier makes label coverage a constraint on selection design, separating the value of better rankings from the cost of more concentrated outputs.

With L fresh, independent labels from a fixed selected system, a sample mean has variance at most $1/(4L)$ regardless of N. Our frontier instead designs selection around reusable-label coverage. Its guarantees apply to each independently specified policy; choosing policies adaptively using audit labels requires further statistical control.

The experiments use frozen reasoning banks; new model generations remain to be tested. Sequential search and dependent proposals require new reachable-law descriptions. For fixed-pool selection, candidate compute determines what can be selected, while label coverage determines how precisely the gain can be substantiated.

[9]Table 9 also compares prediction of realized MSE across held-out problems and tasks, including label-budget-aware KL and TV controls.

REPRODUCIBILITY STATEMENT

Appendices D–K give the proofs and finite-bank construction. Appendix A specifies data, cross-fitting, selectors, and audit protocols; Appendix B reports complete comparisons and numerical checks. Real-bank experiments reuse frozen scores and outcomes. Code, processed data, and experiment results are available at `https://anonymous.4open.science/r/selection-cost`.

AI USE STATEMENT

We used generative AI tools to check grammar, spelling, and word choice, to check the consistency of mathematical notation across derivations and proofs, and to help organize the open-source codebase. The authors are responsible for the final content.

REFERENCES

Shashaank Aiyer, Yishay Mansour, Shay Moran, and Han Shao. A theoretical framework for statistical evaluability of generative models. *arXiv preprint arXiv:2604.05324*, 2026.

Gholamali Aminian, Idan Shenfeld, Amir R. Asadi, Ahmad Beirami, and Youssef Mroueh. Best-of-N through the smoothing lens: KL divergence and regret analysis. In *International Conference on Learning Representations*, 2026.

Zvi Artstein. Distributions of random sets and random selections. *Israel Journal of Mathematics*, 46 (4):313–324, 1983. doi: 10.1007/BF02762891.

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, pp. 3580–3602. PMLR, 2025.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Fan Chen, Audrey Huang, Noah Golowich, Sadhika Malladi, Adam Block, Jordan T. Ash, Akshay Krishnamurthy, and Dylan J. Foster. The coverage principle: How pre-training enables post-training. In *International Conference on Learning Representations*, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Florian E. Dorner, Yatong Chen, André F. Cruz, and Fanny Yang. ROC-n-reroll: How verifier imperfection affects test-time scaling. In *International Conference on Learning Representations*, 2026.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10978–10993. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.754.

Lorenzo Fattorini. Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93(2):269–278, 2006. doi: 10.1093/biomet/93.2.269.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023.

Lin Gui, Cristina Gârbacea, and Victor Veitch. BoNBoN alignment for large language models and the sweetness of best-of-n sampling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2851–2885, 2024. doi: 10.52202/079017-0094.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Tord Holmstedt. Interpolation of quasi-normed spaces. *Mathematica Scandinavica*, 26:177–199, 1970. doi: 10.7146/math.scand.a-10976.

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, pp. 25075–25126. PMLR, 2025.

Edward L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008. doi: 10.1198/106186008X320456.

Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-N sampling with minimum Bayes risk objective for language model alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9321–9347. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.472.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-time reward hacking in large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Andreas Kleiner, Benny Moldovanu, and Philipp Strack. Extreme points and majorization: Economic applications. *Econometrica*, 89(4):1557–1593, 2021. doi: 10.3982/ECTA18312.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 297–306. Association for Computing Machinery, 2011. doi: 10.1145/1935826.1935878.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38, pp. 608–616. PMLR, 2015.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022. doi: 10.1126/science.abq1158.

Qiang Liu and Jason D. Lee. Black-box importance sampling. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 952–961. PMLR, 2017.

Cong Ma, Banghua Zhu, Jiantao Jiao, and Martin J. Wainwright. Minimax off-policy evaluation for multi-armed bandits. *IEEE Transactions on Information Theory*, 68(8):5314–5339, 2022. doi: 10.1109/TIT.2022.3162335.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2 edition, 2011. doi: 10.1007/978-0-387-68276-1.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Samantha Morrison, Constantine Gatsonis, Issa J. Dahabreh, Bing Li, and Jon A. Steingrimsson. Robust estimation of loss-based measures of model performance under covariate shift. *Canadian Journal of Statistics*, 52(4):e11815, 2024. doi: 10.1002/cjs.11815.

Youssef Mroueh and Apoorva Nitsure. Information theoretic guarantees for policy alignment in large language models. *Transactions on Machine Learning Research*, 2025.

Arpan Mukherjee, Marcello Bullo, Debabrota Basu, and Deniz Gündüz. Test-time verification via optimal transport: Coverage, ROC, & sub-optimality. In *International Conference on Learning Representations*, 2026.

Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting Monte Carlo estimators. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4267–4276. PMLR, 2018.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 53132–53176. PMLR, 2025.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. doi: 10.2140/pjm.1958.8.171.

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 *International Conference on Learning Representations*, 2025.

Ved Sriraman and Adam Block. Revisiting the (sub)optimality of best-of-N for inference-time alignment. In *Proceedings of the 39th Conference on Learning Theory*, volume 336, pp. 5980–6028. PMLR, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3589–3597. PMLR, 2017.

Zhuohao Yu, Zhiwei Steven Wu, and Adam Block. From curiosity to caution: Mitigating reward hacking for best-of-N with pessimism. In *International Conference on Learning Representations*, 2026.

Zhen Zhang, Xin Liu, Shaoli Wang, and Jiaye Teng. Minimax optimal two-stage algorithm for moment estimation under covariate shift. In *International Conference on Learning Representations*, 2025a.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025b.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. miniF2F: A cross-system benchmark for formal olympiad-level mathematics. In *International Conference on Learning Representations*, 2022.

A Experimental Details

A.1 Candidate banks, selection signals, and outcomes

The empirical base laws are frozen subsets of the public MONKEY BUSINESS candidate banks (Schaeffer et al., 2025). All four primary configurations use Llama-3-8B-Instruct responses (Dubey et al., 2024). Table 1 gives their sizes. A base-law identifier records the benchmark, generator,

decoding configuration, and raw-bank revision. The original subbanks were selected by sorting candidate identifiers with a deterministic hash using seed 20260801; candidate membership is therefore unchanged by the order of files or parallel workers. For each question, P_x is the uniform distribution on its retained candidates. Pools are sampled with replacement from this law, so a candidate may appear more than once in a pool.

Table 1: **Primary candidate banks.** K is the number of retained candidates per question. $N_{\max}$ is used for the selection trajectories; the implemented-strategy comparison uses $N \in \{8, 32, 64\}$ in every task.

Task	Questions	K	Candidate records	$N_{\max}$
GSM8K	127	512	65,024	256
MATH	128	512	65,536	256
CodeContests	140	256	35,840	128
MiniF2F	130	128	16,640	64
Total	525	—	183,040	—

Math selection uses the frozen step rewards of Skywork-o1-Open-PRM-Qwen-2.5-1.5B, aggregated by the geometric mean on GSM8K and the minimum on MATH. The retained model revision is `98d69606595eedbdbbbf0a7d28efdcd462ba6a67`. GSM8K correctness is the normalized final-answer outcome; MATH correctness is the published symbolic-equivalence outcome. CodeContests selection uses the fraction of valid public tests passed, while correctness requires all hidden tests to pass. MiniF2F selection uses the one-second Lean checking score; its operational correctness label is the published ten-second checking outcome. Existing score and checking statuses are retained, rather than replacing operational outcomes by idealized labels. None of the implemented selectors reads a test question's correctness labels when choosing a winner.

Before policy fitting, PRM aggregation was selected using the labels of design fold 3: 24 GSM8K and 26 MATH questions. The candidates were minimum, geometric mean, and final-step reward. We retained aggregations within one standard error of the best question-macro average precision, selected the lowest candidate-level Brier score among them, and broke remaining ties in the fixed order geometric mean, minimum, final step. The selected aggregation was then frozen across policy fits. The full-bank comparisons include the design questions; Appendix B.3 separately evaluates the remaining questions, whose labels entered neither this aggregation choice nor their own correctness-curve fit.

The 525 task–question rows correspond to 524 semantic groups in the frozen crosswalk. MATH problem 142 and its MiniF2F formalization share a group and fold 4. Other questions are assigned to five folds using the same frozen semantic-group identifiers. This grouping accounts for the annotated overlap; it does not assume that every possible semantic dependence between benchmarks has been identified. Repeated pools, audit trials, and parameter settings are replications on these banks, not additional independent questions.

A.2 Rank representation and fitting

For the verifier score $s_x(z)$, let F_x denote its CDF under P_x. Each drawn occurrence receives an independent auxiliary random variable $V \sim \mathrm{Unif}[0, 1]$ and rank

$$U = F_x(s_x(z)-) + V\{F_x(s_x(z)) - F_x(s_x(z)-)\}. \tag{22}$$

This rank is uniform even for discrete scores. In particular, it gives uniform tie-breaking for score best-of-N, without ordering tied responses by their labels. Writing the tie interval for candidate j as $[a_{xj}, b_{xj}]$, a rank-density kernel w induces the exact candidate weight

$$\bar{w}_{xj} = \frac{1}{b_{xj} - a_{xj}} \int_{a_{xj}}^{b_{xj}} w(u)\, du. \tag{23}$$

Finite-bank accuracy is consequently $K^{-1} \sum_j \bar{w}_{xj} Y_{xj}$ for one question. All weights used for the new audits are computed by this integration, rather than evaluating the kernel at a tied score's average rank.

Each task is fitted separately. In each fold rotation, the other four folds provide training labels for a shared, question-independent rank kernel. For every training question, we integrate its piecewise-constant correctness profile over M equal rank bins. We then average those profiles with equal question weight and fit a nondecreasing isotonic regression to the resulting bin values. The fitted values are the plug-in correctness objective $\widehat{\eta}$. This procedure uses complete tie intervals; it differs from fitting at tied-score midpoints and interpolating the fitted curve. With score aggregation fixed, labels in each evaluation fold do not enter correctness-curve fitting, baseline selection, or the optimization constraints. The grid size is $M = \max\{512, 32N\}$, giving 512, 1,024, and 2,048 bins for the three candidate budgets. These grids align all primary-bank tie boundaries.

A.3 Implemented selectors and matched budgets

The strategy experiment crosses four tasks, five folds, $N \in \{8, 32, 64\}$, $L \in \{32, 64, 128\}$, and five audit radii. If $r_0 = L^{-1/2}$ and r_N is the BoN radius of the finite rank-grid density, the radii are

$$r_j = r_0 + (r_N - r_0)(j/20)^2, \qquad j \in \{0, 5, 10, 15, 20\}. \tag{24}$$

There are 900 task–fold–budget settings. All nine selectors below use the same training objective and the same frozen candidate laws.

The *frontier selector* solves the finite capacity- and audit-profile-constrained program. Its feasible density is factored into random priority permutations by the constructive majorization procedure. An independent permutation is drawn for each pool, and the highest-priority bin present in that pool supplies the returned candidate. Thus execution always returns an observed member of the N-candidate pool. Drawing directly from a desired winner distribution over the entire bank is not used as a substitute for this step.

Uniform selection returns a uniformly chosen pool occurrence, and *ordinary BoN* returns the largest randomized rank. These are reference endpoints. Ordinary BoN generally exceeds an interior radius budget, while uniform selection generally leaves that budget unused. The *Uniform–BoN mixture* uses density $1 + \alpha(w_N - 1)$ with the largest $\alpha \in [0, 1]$ satisfying the audit constraint. It is executed by choosing between the corresponding two poolwise rules.

For *rank-power selection*, the candidate directions are $(a+1)u^a$ for $a \in \{0.25, 0.5, 1, 2, 4\}$, integrated exactly over each rank bin. Each direction is mixed with uniform to satisfy the audit budget, and the best fitted objective determines the exponent. The resulting density is executed with the same priority construction as the frontier.

The *order-statistic hull* optimizes a probability vector $\beta \in \mathbb{R}_+^N$ over *all* N order statistics. If b_k is the bin-integrated density of the kth ascending rank in an N-pool, its program maximizes $\int \widehat{\eta} \sum_k \beta_k b_k$ subject to $\sum_k \beta_k = 1$ and $\omega_L(\sum_k \beta_k b_k) \le r_j$. Execution first samples k from β, then returns that rank in the observed pool. The *L^2-constrained frontier* instead optimizes over the complete finite capacity region with constraint $\int w^2 \le L r_j^2$. It is not restricted to a mixture direction. Since $\omega_L(w) \le \|w\|_2 / \sqrt{L}$, this is a sufficient, possibly stronger, version of the same audit-radius constraint.

Two more restricted paths are retained for comparison with the initial numerical study. The *order-statistic envelope* picks the best of six individual order statistics, each mixed with uniform to meet the budget. Their ranks are $k = \min\{N, \max\{1, \lceil q(N+1) \rceil\}\}$ for $q \in \{0.5, 0.75, 0.9, 0.95, 0.99, 1\}$. The *$L^2$-matched Uniform–BoN path* uses the Uniform–BoN direction with its mixture weight constrained by $\|w\|_2 / \sqrt{L} \le r_j$. These paths are distinct from the full hull and full L^2 program above; gains against them do not measure gains against those stronger optimizers.

The comparison matches the rank-space audit constraint of each fitted fold, not observed MSE. Integrating out the auxiliary rank can decrease that fold's candidate-space radius. Combining different fitted policies across folds is a separate operation: their common radius bound does not itself imply that the pooled cross-fitted law has $\omega_L \le r_j$. Lemma K.5 instead bounds its ρ_L by the question-weighted average of the foldwise profiles squared, preserving the clipped-estimator guarantee $r_j^2/4$ for L total joint labels. We also compute both ω_L and ρ_L directly for every joint law; the observed pooling differences are reported in Appendix B. The optimization objective is the fitted rank-class objective. Held-out accuracy is measured separately after each policy has been fixed.

A.4 POOL EXECUTION, AUDIT TRIALS, AND TRAJECTORIES

Table 2 summarizes replication. Each of the 900 settings contains nine policies, giving 8,100 policy–fold settings and 212,625 question–policy settings. Each question–policy setting is executed on 8,192 independently sampled candidate pools. Its measured accuracy is checked against the exact candidate-weight calculation in (23).

Table 2: **Replication in the implemented-strategy study.** An audit repetition draws a new label sample from the indicated frozen base law. The three audit targets are reported separately.

Computation	Settings	Repetitions per setting
Actual pools, per question and policy	212,625	8,192
Single-question audit	212,625	2,000
Joint held-out-fold audit	8,100	10,000
Joint all-question, cross-fitted audit	1,620	10,000
Matched BoN trajectory audit	33	100,000

Audits use the centered clipped estimator in (12), projected onto $[0, 1]$. The clipping threshold minimizes the known-law objective ρ_L and depends only on the candidate weights and L, not on the audit labels. A single-question audit draws L labels from that question's base law. A joint fold audit first draws a held-out question uniformly and then a candidate uniformly within that question, repeating this procedure for L *total* labels. The all-question audit similarly targets the entire task, assigning each question the policy fitted without its fold. This last construction yields 1,620 task–budget–policy summaries and weights questions equally rather than weighting fold means equally.

We estimate MSE by averaging squared error against the exact finite-bank selected accuracy. Its Monte Carlo standard error is the sample standard deviation of squared errors divided by the square root of the number of audit repetitions. We also retain the unprojected estimator as a diagnostic: its exact MSE is the variance of $(\bar{w} \wedge \tau)(Y - 1/2)$ divided by L, plus its squared bias. This identity and the independently computed candidate probabilities provide checks separate from the optimization objective.

The 33 BoN points in Figure 1 use powers of two from $N = 1$ to the task-specific $N_{\max}$ in Table 1. Both coordinates use the same task, candidate distribution, score, tie-breaking rule, and correctness labels. Accuracy averages 8,192 actual pool executions per question. Audit MSE uses 100,000 independent repetitions of $L = 64$ total labels from the task's joint base law. Its exact BoN candidate weight is $(b_{xj}^N - a_{xj}^N)/(b_{xj} - a_{xj})$. These points were recomputed directly from the frozen banks because the earlier audit export did not retain the pool-sampling protocol in its aggregation keys.

A.5 QUESTION-LEVEL UNCERTAINTY AND PREDICTIVE DIAGNOSTICS

Accuracy intervals are percentile intervals from 5,000 question-level bootstrap resamples. For a comparison between two policies, their differences are paired within question. The aggregate interior comparison first averages the 27 settings $N \in \{8, 32, 64\}$, $L \in \{32, 64, 128\}$, and $j \in \{5, 10, 15\}$ within each question, then averages over questions. The four-task macro average gives each task equal weight. Its bootstrap draws one multiplicity for each semantic group, shares that multiplicity across the known MATH–MiniF2F overlap, and normalizes each task's mean separately. These intervals are conditional on the fitted cross-fold policies; the policies are not refitted inside each bootstrap draw. Pointwise intervals over individual settings are not simultaneous confidence bands for the full grid.

The MSE-prediction diagnostic uses all 212,625 question–policy records. The target is base-ten log MSE, with a numerical floor of 10^{-14}; all observed values exceed this floor. For ω_L^2, ρ_L, and m_2/L, the feature is the base-ten logarithm with the same numerical floor. KL and TV use $\log(1 + \max\{d, 0\})$ to retain zero-distance cases, clipping numerical roundoff below zero. The budget controls are $\log_{10} L$ and $\log_{10} N$. Each predictor is a linear regression with an intercept; training weights give equal total weight to each included task. Thus the comparison is between specified calibrated predictors, not an ordering of the underlying distances under arbitrary regression models.

One validation scheme holds out a complete question fold, including every selector and setting for each question in that fold. The other holds out a complete task and also removes any matching semantic group from its training tasks. Predictions are collected only on the held-out rows. The reported R^2 is pooled over these out-of-fold predictions; its interval resamples semantic groups while holding predictions fixed. This is a diagnostic calibration study on the already fitted strategy matrix, not a nested procedure that relearns all upstream policies for each outer split. The complete budget-aware comparison is in Table 9.

A.6 Controlled validation and numerical verification

The controlled audit experiment uses five known selected-law families: uniform, BoN, half-uniform/half-BoN, median order statistic, and a capped rank-power family. Correctness functions include constant, linear, logistic, hump-shaped, oscillatory, and least-favorable two-point cases. The retained grids use $N \in \{2, 8, 32, 128, 512, 2048\}$ and label budgets between 2 and 16,384, including the transition values $2N - 1$. The rank integration grid has 65,536 points. There are 2,375 centered-clipped projected-estimator cells, with 2,000 or 10,000 Monte Carlo trials each, and 335 two-point Bayes-risk cells. The latter compare two known correctness models under a prior; they do not identify the observed MSE of one estimator with a minimax lower bound.

The implemented-strategy experiment uses CVXPY with Clarabel and checks capacity, radius feasibility, and objective certificates on every returned policy. The nominal solver tolerance is 10^{-8}, with tighter retries when needed. Numerical feasibility repairs mix with uniform and are checked against an L^1 budget of 2×10^{-6}; the original and final densities are retained. The final objective-certificate tolerance is 2×10^{-6} and the realization tolerance is 2×10^{-9}. A deterministic hash sample of approximately 5% of interior settings is also solved with ECOS. The achieved residuals and all cross-solver outcomes are summarized in Table 8 and its discussion.

The implementation, matched trajectories, and clustered analysis constitute a new analysis of previously examined frozen data. The execution protocol, code, inputs, and random seeds were fixed for this analysis; this is not a claim of preregistration before the original results were seen. The new random streams use master seed 20260912 and purpose-specific hashed seeds. No new model generation, verifier inference, or checker run was required. Computations ran on a 64-core CPU server with approximately 251 GiB of memory, using 48 single-threaded workers for the main matrix and 32 for analysis; no GPU or proprietary model API was used.

B Additional Experimental Results and Diagnostics

B.1 Matched selection trajectories

Table 3 gives the endpoints of the protocol-matched trajectories in Figure 1. The MSE multiplier compares each task's endpoint with $N = 1$ under the same $L = 64$ joint-base-label experiment. Accuracy improves in all four tasks, while endpoint MSE increases by factors from 2.58 to 37.35. The trajectory need not be monotone at every intermediate budget: in MATH, for example, MSE decreases between $N = 128$ and $N = 256$.

Table 3: **Endpoints under a matched selection and audit protocol.** Accuracy is in percent. Brackets give question-bootstrap 95% intervals for the selected accuracy; the MSE multiplier is a point estimate.

Task	$N_{\max}$	Base	Selected [95% CI]	MSE multiplier
GSM8K	256	76.72	91.51 [86.99, 95.41]	37.35×
MATH	256	26.91	53.04 [45.36, 60.78]	15.08×
CodeContests	128	2.41	8.27 [4.45, 12.55]	30.20×
MiniF2F	64	20.58	41.21 [32.47, 49.86]	2.58×

Base MSE is 0.002728 on GSM8K, 0.002988 on MATH, 0.000409 on CodeContests, and 0.002507 on MiniF2F. At the largest candidate budgets, these values become 0.101871, 0.045076, 0.012364, and 0.006475, respectively. These are errors of the specified clipped base-label estimator on the frozen task laws, rather than estimates of each law's exact minimax risk.

B.2 HELD-OUT ACCURACY OF THE IMPLEMENTED STRATEGIES

Table 4 reports the four-task macro comparisons against every baseline. The interior aggregation averages 27 budget settings within each question before comparing question means, as described in Appendix A. This prevents a question with many repeated settings from contributing as if it supplied many independent observations. Table 5 separates the four tasks for the principal constrained baselines.

Figure 6 visualizes these paired comparisons, including their problem-level uncertainty and the smaller differences among fully optimized policies.

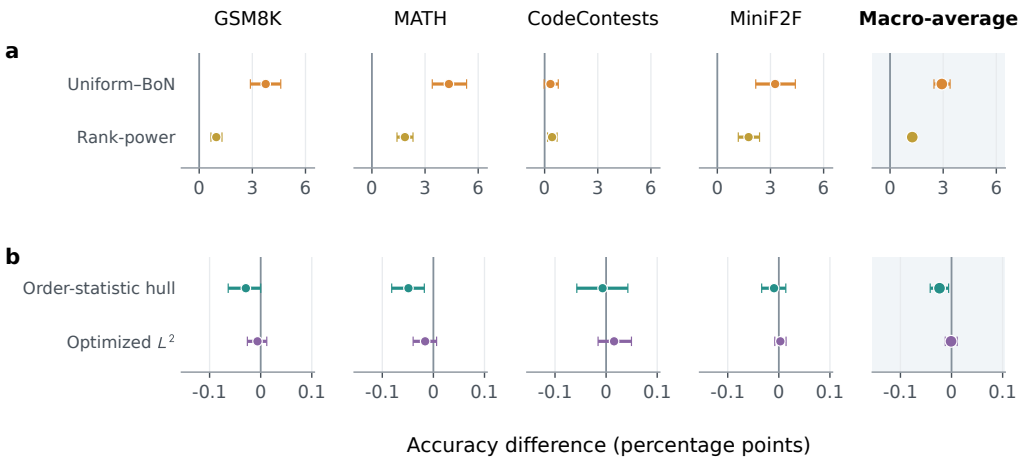


Figure 6: **Accuracy gains over simple rules, with close agreement among optimized policies.** Frontier-minus-baseline accuracy in percentage points, with 95% problem-bootstrap intervals conditional on the fitted policies. Each problem averages 27 interior settings; the macro-average weights tasks equally. Positive values favor the frontier. The two rows use different horizontal scales to resolve the smaller optimized-policy differences.

Table 4: **Actual-pool accuracy differences against all baselines.** Entries are frontier minus baseline in percentage points, macro-averaged over tasks and interior settings. Intervals are paired semantic-group bootstrap intervals conditional on the fitted policies. The last two rows are reference endpoints, not matched interior-budget competitors.

Baseline	Difference	95% lower	95% upper
Uniform–BoN mixture	$+2.924$	$+2.480$	$+3.390$
Rank-power	$+1.254$	$+1.051$	$+1.474$
All-order-statistic hull	-0.024	-0.042	-0.006
L^2-constrained frontier	-0.001	-0.012	$+0.011$
Order-statistic envelope	$+1.871$	$+1.626$	$+2.117$
L^2-matched Uniform–BoN path	$+2.991$	$+2.541$	$+3.461$
Uniform selection	$+8.285$	$+7.251$	$+9.320$
Ordinary BoN	-5.237	-6.133	-4.361

The gains over Uniform–BoN are supported in GSM8K, MATH, and MiniF2F; CodeContests has an interval of $[-0.025, 0.770]$ percentage points. The all-order hull is slightly better on the macro average, with a difference of approximately 0.024 points. The L^2 frontier comparison does not resolve a difference at this precision. Thus the broad optimized families recover nearly all of the held-out accuracy obtained by the general construction on these banks. The comparison with the two restricted envelope/path baselines measures a different limitation and is labeled separately.

Table 5: **Task-specific accuracy differences.** Entries use the same within-question interior aggregation as Table 4, in percentage points.

Task	Uniform–BoN	Rank-power	All-order hull	L^2 frontier
GSM8K	$+3.755$	$+0.965$	-0.029	-0.006
MATH	$+4.347$	$+1.865$	-0.049	-0.016
CodeContests	$+0.330$	$+0.421$	-0.007	$+0.016$
MiniF2F	$+3.263$	$+1.766$	-0.009	$+0.003$

Equal rank-radius budgets need not produce equal finite-bank radii or equal MSE. Table 6 gives one common setting, $N = 32$, $L = 64$, and radius index $j = 10$, with rank budget approximately 0.219733. For example, the MATH Uniform–BoN policy at this budget has accuracy 0.348642 and MSE 0.006459, compared with 0.410294 and 0.010089 for the frontier policy. The accuracy gain here comes with a larger observed MSE, despite the common theoretical constraint.

Pooling the five fitted laws produces four frontier settings with a joint ω_L slightly above their common per-fold radius: GSM8K at $(N, L) = (32, 32), (64, 32), (64, 64)$ and MATH at $(64, 32)$, all with $j = 15$. The largest excess is 0.0016703. These are pooling differences, not floating-point feasibility residuals. Direct evaluation gives $\rho_L \le r_j^2$ for every non-BoN pooled policy in the completed matrix, so their known-law clipped-estimator upper bounds remain at most $r_j^2/4$.

Table 6: **Joint finite-bank audit of one implemented frontier setting.** All tasks use $N = 32$, $L = 64$, and $j = 10$. Accuracy is computed from the exact candidate marginals; MSE uses 10,000 repetitions with L total task-level base labels.

Task	Exact accuracy	Candidate-space radius	Observed MSE
GSM8K	0.866160	0.219709	0.008315
MATH	0.410294	0.219635	0.010089
CodeContests	0.043945	0.151573	0.001423
MiniF2F	0.321541	0.143074	0.004009

B.3 Evaluation beyond the score-design questions

PRM aggregation was selected on design fold 3 before the five-fold policy fits, as specified in Appendix A. The full-bank comparisons above hold those scores fixed and include the design questions. To separate score design from evaluation, we repeat the aggregate comparison after excluding the 24 GSM8K and 26 MATH design questions. This leaves 103, 102, 140, and 130 questions in the four tasks, respectively: 475 task–question rows and 474 semantic groups. For each remaining math question, both score aggregation and its correctness-curve fit use only other folds.

Table 7 uses the existing pool executions, with no policy refitting. We again average the 27 interior settings within each question and give each task equal weight. The 95% intervals use 5,000 paired semantic-group bootstrap resamples of the retained questions, sharing weights across the MATH–MiniF2F overlap and holding fitted policies fixed. The gain over Uniform–BoN is 2.965 percentage points; the close comparisons with the two optimized baselines also persist.

Table 7: **Accuracy comparisons excluding PRM score-design questions.** Frontier-minus-baseline differences in percentage points, using the 475 retained task–question rows. The full-bank comparison is in Table 4.

Baseline	Difference	95% lower	95% upper
Uniform–BoN mixture	$+2.9654$	$+2.5015$	$+3.4397$
Rank-power	$+1.2638$	$+1.0366$	$+1.4933$
All-order-statistic hull	-0.0219	-0.0406	-0.0036
L^2-constrained frontier	-0.0005	-0.0130	$+0.0121$

B.4 WHERE THE OPTIMIZED POLICIES AGREE

The quadratic regime in (13) gives a direct explanation for the close accuracy of the two frontier policies. For each frozen interior frontier density, we compute $\|w\|_\infty/(\sqrt{L}\|w\|_2)$ on its equal-mass rank grid. Of the 540 task–fold–budget fits, 500 (92.6%) have ratio at most one: 125 of 135 on GSM8K, 129 on MATH, 111 on CodeContests, and all 135 on MiniF2F. The classification is unchanged by a numerical tolerance of 10^{-7}. For these fitted densities, the bounded profile equals the quadratic cost. More generally, whenever a profile-optimal density lies in this regime, it is also feasible for the quadratic program. Since the quadratic feasible set is contained in the profile feasible set, their optimal fitted objectives then coincide. The two solvers can still choose different optimizers with different held-out accuracies. The rare-region construction in Proposition I.1 shows why this agreement need not extend to all selected laws.

The all-order hull gives a complementary comparison of fitted and held-out objectives. The general frontier has a higher fitted correctness objective in all 540 interior settings, while the exact held-out task averages favor the hull by 0.009–0.027 percentage points. These exact candidate-marginal calculations exclude execution Monte Carlo noise. The reversal is consistent with a small generalization benefit from restricting the selector class when optimizing an estimated correctness curve.

B.5 FINITE-SAMPLE BOUNDS AND IMPLEMENTATION CHECKS

In the controlled experiment, none of the 2,375 projected-estimator cells exceeds the finite-sample upper envelope by five Monte Carlo standard errors; the largest standardized excess is 3.00. The median MSE/ω_L^2 is 0.2091. The 335 least-favorable two-point Bayes comparisons have median Bayes-risk-to-ω_L^2 ratio 0.0495. These controlled checks complement the exact inequalities: they test the implemented estimator and illustrate finite-sample regimes without treating a single correctness function as the minimax worst case.

All 900 implemented-strategy settings completed. Independent verification reconstructed every serialized policy's output distribution, including the direct order-statistic and mixture rules as well as the priority programs. It also recomputed finite-bank accuracy by a separate discrete summation. Table 8 summarizes the resulting residuals.

Table 8: **Numerical and implementation checks.** Each entry is the maximum over its applicable completed settings. Objective gaps concern the final serialized policies, after any numerical feasibility repair. MC z denotes a Monte Carlo residual divided by its stabilized standard error.

Diagnostic	Maximum
Independently reconstructed distribution, L^1 error	2.80×10^{-13}
Independent finite-bank accuracy calculation, absolute error	3.85×10^{-14}
Final primary-solver primal–dual objective gap	5.95×10^{-8}
Numerical feasibility repair, L^1 change	3.48×10^{-9}
Clarabel–ECOS objective disagreement	2.01×10^{-7}
Actual-pool versus exact accuracy, absolute MC z	5.16
Candidate selection frequency, absolute MC z	5.57
Unprojected MSE identity, absolute MC z	4.99

For the MC diagnostics, accuracy residuals use denominator $\max\{\mathrm{SE}, 1/R\}$ and selection-frequency residuals use $\mathrm{SE} + 1/R$, where $R = 8192$ pool trials and the standard errors use exact finite-bank probabilities. MSE residuals use the empirical standard error with denominator $\max\{\widehat{\mathrm{SE}}, 10^{-15}\}$. These are implementation checks, not simultaneous hypothesis tests.

The ECOS cross-check covers 27 interior settings; 13 return `optimal_inaccurate`, and those statuses are retained alongside the objective comparisons. Cross-solver agreement is an additional numerical check, not an exact-arithmetic certificate. The largest standardized excesses over the finite-sample upper bound are 3.83 for single-question audits and 2.69 for fold-level joint audits. All specified Monte Carlo checks fall within the execution threshold of eight standard errors. The full implementation test suite contains 397 passing tests; one existing solver stress test reports a precision warning.

The completed 900-setting optimization and execution matrix consumes 11,325 CPU seconds, or approximately 3.15 core-hours, summed over its worker processes. Its parallel worker phase takes about 255 seconds after input preparation. In a separate single-process throughput test, the frontier selector's median amortized cost is 6.36, 10.54, and 19.92 microseconds per pool for $N = 8$, 32, and 64, respectively. This timing includes the priority randomness but excludes optimization, rank computation, compilation warm-up, candidate generation, and verifier inference; it is batch throughput, not online single-request latency.

B.6 Predicting observed MSE

The minimax profile is a uniform complexity measure, while predicting the MSE realized by a particular correctness function is a separate empirical question. Table 9 evaluates the calibrated predictors specified in Appendix A. Including L in the KL and TV controls puts their access to label-budget information on the same footing as the audit-profile quantities.

Table 9: **Complete budget-aware MSE prediction comparison.** Entries are held-out R^2 for predicting log MSE, with 95% semantic-group bootstrap intervals. The test matrix has 212,625 records from 524 semantic groups. Regressions and upstream policies remain fixed within each bootstrap resample.

Predictor	Question-fold holdout	Task holdout
ω_L^2	0.660 [0.646, 0.673]	0.557 [0.542, 0.571]
ρ_L	0.656 [0.643, 0.669]	0.560 [0.545, 0.573]
m_2/L	0.655 [0.641, 0.668]	0.550 [0.535, 0.564]
KL only	0.608 [0.593, 0.621]	0.523 [0.508, 0.537]
TV only	0.667 [0.652, 0.681]	0.610 [0.594, 0.623]
KL and $\log L$	0.692 [0.677, 0.706]	0.604 [0.589, 0.618]
TV and $\log L$	0.751 [0.735, 0.766]	0.690 [0.674, 0.703]
$\log L$ only	0.080 [0.074, 0.082]	-0.198 [-0.213, -0.186]
$\log N$ and $\log L$	0.092 [0.086, 0.093]	-0.189 [-0.204, -0.177]

The three quantities ω_L^2, ρ_L, and m_2/L have similar predictive performance, and their ordering is not identical under the two holdout schemes. TV together with the label budget is the strongest of these specified regressions. Accordingly, the empirical study does not identify the minimax profile as the best predictor of dataset-specific MSE. Its role in the frontier is the finite-sample worst-case guarantee, which does not require such a predictive ranking.

B.7 Secondary generator and verifier slices

The original selection-only experiment also contains a Llama-3-70B-Instruct MATH slice with 128 candidates per question and a paired 128-candidate MATH-8B slice scored by Qwen2.5-Math-PRM-7B (Zhang et al., 2025b). The Skywork-70B slice uses final-step reward, and the Qwen-PRM slice uses minimum step reward. The Qwen model revision is `0610740060112df12585d00a1c5f4624d2f59051`. Under with-replacement score BoN, the 70B slice rises from 0.4794 at $N = 1$ to 0.6885 at $N = 256$. The Qwen-PRM slice rises from 0.2718 to 0.5552 over the same range. These results concern selection scaling on additional frozen configurations; the implemented-frontier and joint-audit comparisons use the four primary 8B configurations in Table 1.

B.8 Optional black-box replay diagnostics

The shadow extension in Appendix L has a separate experiment on MATH and CodeContests. It uses $N \in \{8, 32, 128\}$, $L \in \{64, 256\}$, $B \in \{1, 8, 64\}$ companion pools, and $S \in \{1, 8, 64, \infty\}$ internal replays. The selectors are score BoN, half-uniform/half-BoN, and soft-rank selection with inverse temperature one. Regressing the empirical second moment on $1/B$ and $1/(BS)$ estimates the three coefficients in (59). Independent reference probes supply their comparison targets.

Twelve of the 18 task–selector–candidate-budget fits pass the frozen reference-precision and moment-order checks. Figure 7 shows those comparisons. Their R^2 values are at least 0.987; the largest

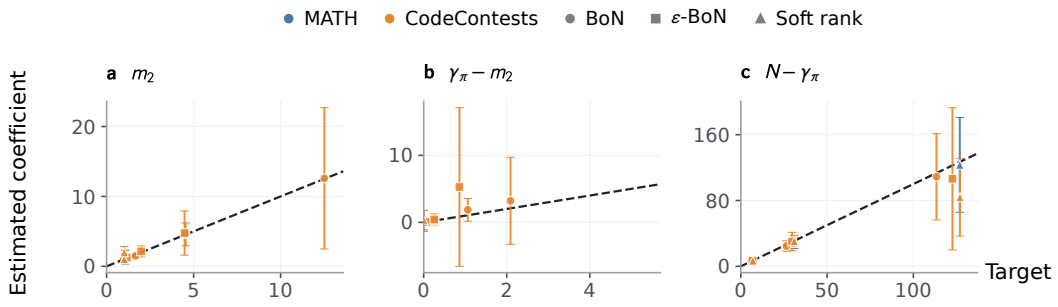


Figure 7: **Shadow-replay coefficient checks.** Estimated coefficients versus independently measured targets for the 12 reference-qualified fits. Error bars are 95% Monte Carlo intervals. The legend's ε-BoN is the half-uniform/half-BoN mixture. The six fits with inadequate reference moments are discussed in the text.

standardized coefficient differences for m_2, $\gamma_\pi - m_2$, and $N - \gamma_\pi$ are 2.23, 1.18, and 1.79, respectively, using the combined standard errors of each estimate and its independent target. The six remaining fits are the MATH BoN and half-uniform/half-BoN cases. Their independently estimated targets are insufficiently precise or fail the empirical moment order, so their high regression R^2 values alone do not test the coefficient identity.

Interval diagnostics provide a different assessment. All 180 exported $L = 64$ summaries have mean interval width one at the 90%, 95%, and 99% levels after intersection with $[0, 1]$. At $L = 256$, the unweighted mean 95% width over the 180 summaries is 0.8703, its minimum is 0.3889, and 120 summaries still have width one. Coverage in these summaries is one, but the width-one intervals contain the entire parameter space. The range term in the empirical-Bernstein bound explains their limited precision at small label budgets.

These exported summaries pool probe-bank-resampled and fresh-companion trials with the same displayed settings. They therefore describe the retained diagnostic output, rather than a fresh-only coverage or unbiasedness experiment. Their standardized mean residuals can be large—the maximum absolute ratio of the exported mean error to its reported standard error is approximately 30.0—and are not interchangeable with the independently probed coefficient checks. The replay evidence is used here to assess the variance decomposition; it does not establish practically narrow observable confidence intervals for the evaluated budgets.

The computational distinction is also substantial. Companion pools cost $LB(N-1)$ candidate draws: for example, $L = B = 64$ and $N = 128$ require 520,192 draws. Here these are lookups in frozen banks; a live generator would have to produce the companions. When fresh selected-output labels are affordable, their sample mean has variance at most $1/(4L)$ without replay. Shadow replay instead trades additional unlabeled computation for access to a selector whose density is unavailable; the present experiments measure that tradeoff's variance components.

C Extended Related Work

Test-time selection and regularization. Verifiers rank sampled mathematical solutions (Cobbe et al., 2021), while self-consistency and minimum-risk decoding aggregate information from several responses (Wang et al., 2023; Eikema & Aziz, 2022). Studies of repeated sampling and adaptive inference establish how these choices affect compute–accuracy scaling (Brown et al., 2024; Schaeffer et al., 2025; Snell et al., 2025). Theoretical analyses explain selected accuracy through coverage and reward model error (Huang et al., 2025), verifier ROC curves (Dorner et al., 2026), and reward or win-rate tradeoffs (Gui et al., 2024; Beirami et al., 2025; Mroueh & Nitsure, 2025). These perspectives characterize selection quality; the evaluation axis of our frontier asks how reliably that quality can be measured from base labels.

A close connection is the use of distributional proximity to control selection. Beirami et al. (2025) analyze the KL divergence and win rate of the best-of-N alignment policy, including the distinction

between a KL upper bound and an exact expression. MBR-BoN (Jinnai et al., 2025) combines a reward score with a minimum-Bayes-risk objective serving as proximity regularization, with the aim of reducing reward hacking. Our adjusted score also discourages some concentration, but its penalty is determined by finite-label evaluation: the audit budget L and the base law specify which changes in correctness the labels can resolve. The optimization ranges over all laws reachable with N candidates.

Pessimistic selection addresses a related but distinct source of uncertainty. Yu et al. (2026) subtract a learned uncertainty penalty from the reward score to reduce reward hacking. Our dual penalty instead prices the error of evaluating the selected law from L base labels. This distinction also affects implementation: if a fixed corrected score is continuously distributed conditional on the prompt, maximizing it over N candidates remains an extreme-order selector, with the same audit profile by Corollary 4.3. Changing the ranking can improve accuracy without reducing that worst-case evaluation cost. Interior frontier points can instead require randomization among maximizers of the adjusted score.

Further analyses study reward overoptimization, smoothing, coverage, and selection guarantees (Gao et al., 2023; Khalaf et al., 2025; Aminian et al., 2026; Sriraman & Block, 2026; Chen et al., 2026). Batched and sequential verification samplers have also been compared through coverage and optimal transport (Mukherjee et al., 2026). Our fixed-pool model does not characterize arbitrary adaptive stopping or sequential search. Within that model, the evaluation cost persists even when the score orders correctness perfectly: the issue is label coverage, not necessarily verifier error.

Off-policy evaluation and finite-sample risk. For a fixed selector, estimating $\int \eta w \, d\mu$ from base-distributed labels is an off-policy evaluation problem. Importance weighting and truncation are established tools (Ionides, 2008). Li et al. (2015) study finite-time minimax value estimation under a behavior policy different from the target policy; Wang et al. (2017) develop optimal and adaptive procedures for contextual bandits. Particularly close is Ma et al. (2022): for bounded rewards in a multi-armed bandit, their Switch estimator is minimax rate-optimal for all sample sizes when the behavior policy is known. Thus finite-sample minimaxity and the use of clipping or switching are not, by themselves, the distinction of the present work.

Our audit result expresses the difficulty of a known selected law through a bounded L^1–L^2 profile, uniformly over correctness functions without smoothness or a parametric reward model. This profile can then be optimized over the entire pool-reachable class. Its second-moment regime recovers the familiar $\int w^2 \, d\mu/L$ dependence; outside that regime, the bounded profile accounts for amplified regions whose selected mass is small enough to discard. The frontier uses this finite-sample quantity as a constraint on selection itself. It is exact for the profile, while its relation to the exact minimax-risk frontier is up to universal constants (Appendix G).

The minimax infimum ranges over all estimators using the specified information, including control-variate and doubly robust estimators built from the same observations. Its lower bound therefore concerns information missing from the base sample, not a particular importance-weighting formula. Accurate reward predictions can still reduce instance-specific error, as in Wang et al. (2017). Additional labeled data or assumptions linking those predictions to correctness provide information beyond this unrestricted L-label model. The pilot labels used to fit our selectors are likewise a separate resource from the audit budget.

Evaluation under covariate shift similarly asks how data collected under one law can support conclusions under another (Morrison et al., 2024; Zhang et al., 2025a). Our selection constraint specifies which target laws can arise from a fixed candidate budget. Generative-model evaluability (Aiyer et al., 2026) addresses another aspect of measurement: whether distributional metrics or bounded tests can be evaluated from i.i.d. target samples. Here the functional is bounded accuracy, and the difficulty arises because its labels are available under the base law. Direct labeling of selected outputs removes this particular distribution mismatch, as discussed in the main text.

Random-set geometry and interpolation. The pool is a random set, and returning one of its members is a measurable selection. The capacity criterion in Theorem 3.1 specializes random-set selection theory (Artstein, 1983) to an i.i.d. pool. Its decreasing-rearrangement form and convex concentration bounds follow from majorization (Marshall et al., 2011). In particular, continuous-score

best-of-N is an extreme law regardless of whether its score is useful: concentration and correctness alignment are different properties.

The evaluation profile uses classical L^1–L^2 interpolation (Holmstedt, 1970). The L^1 component measures the bias from discarding selected mass, and the L^2 component measures the variance of the retained weights. The bounded dual functions in $\mathcal{H}_L$ describe correctness differences that base labels cannot resolve well. Selection geometry provides the feasible densities; interpolation supplies their evaluation cost. Coupling these two structures with the linear accuracy objective produces the dual adjusted-score characterization and the implementable finite-grid approximation. This also explains why a divergence constraint alone cannot replace the pool-realizability condition.

Replay and additional evaluation resources. Replay evaluation (Li et al., 2011), Monte Carlo inclusion probabilities (Fattorini, 2006), and black-box importance sampling (Liu & Lee, 2017) provide related ways to reuse observations or estimate weights without a closed-form density ratio. Shadow replay exploits a specific identity of candidate selection: an anchor's conditional selection probability, multiplied by N, equals its selected-law amplification. This estimates the weight through observable choices rather than through calibrated verifier scores.

The repeated companion pools and internal selector calls form a nested Monte Carlo construction (Rainforth et al., 2018). Here its second moment admits the exact decomposition in (60). Consequently, more companion pools and more internal replays reduce different terms; neither replaces the labels needed for the intrinsic known-law evaluation problem. Replay is therefore an extension of the same evaluation framework when an analytic density is unavailable, with additional Monte Carlo cost. Its unbiasedness concerns the unprojected estimator for a stateless selector fixed independently of the audit labels, under the independence conditions in Appendix L.

D Proof of the Selected-Law Capacity Theorem

Conditional on $X = x$, the selected index is measurable with respect to the candidate pool and independent protocol randomness. Conditional independence of candidate pairs gives

$$\mathbb{E}[Y_J \mid X = x, Z_{1:N}, J] = \eta_x(Z_J).$$

Integrating first over the selected law and then over X proves $A_\pi = \int \eta w_\pi d\mu$. Applying iterated expectation to an independent base draw gives the importance-weight representation in (3).

Fix a prompt, write $P = P_x$, and let Z_J be a selected coordinate. For any measurable B,

$$\mathbb{P}(Z_J \in B) \le \mathbb{P}(\{Z_1, \dots, Z_N\} \cap B \neq \varnothing) = 1 - (1 - P(B))^N,$$

which proves necessity of (4). Conversely, regard $\{Z_1, \dots, Z_N\}$ as a random finite set. The random-selection theorem of Artstein (1983) states that Q is the marginal law of a random element of this set if and only if these hitting-probability inequalities hold. Disintegrating the resulting joint law conditional on $Z_{1:N}$ gives a randomized protocol, and symmetrization preserves Q. Taking $P(B) = 0$ in (4) also gives $Q \ll P$.

For the conditional assertion, put $\nu(dx, dz) = P_X(dx)Q_x(dz)$ and consider the joint random finite set

$$\mathsf{S} = \{(X, Z_1), \dots, (X, Z_N)\} \subseteq \mathcal{X} \times \mathcal{Z}.$$

For every x in a common P_X-full set, majorization implies the Hall inequality simultaneously for all measurable subsets of $\mathcal{Z}$. Therefore, for every measurable $C \subseteq \mathcal{X} \times \mathcal{Z}$, conditional capacity gives

$$\begin{aligned}\nu(C) &= \mathbb{E}[Q_X(C_X)] \\ &\le \int [1 - (1 - P_x(C_x))^N] P_X(dx) = \mathbb{P}(\mathsf{S} \cap C \neq \varnothing).\end{aligned}$$

Equip the product standard Borel space with a compatible Polish topology. Then S is a random finite closed set, so Artstein's theorem couples a point of law ν to S. Regular conditional distributions lift the coupling to the original ordered tuple and disintegrate the selected point given that tuple. This produces a jointly measurable randomized index. A uniform tie break among duplicate coordinates and symmetrization complete the construction.

For a density w, nonatomicity yields

$$\sup_{P(B)=t} \int_B w \, dP = \int_0^t w^{\downarrow}(s) ds.$$

Thus (4) is equivalent to (5), because

$$\int_0^t N(1-s)^{N-1} ds = 1 - (1-t)^N.$$

The convex-function characterization of majorization proves (7) (Marshall et al., 2011). Taking $\Phi(t) = t^p$ gives

$$\int_0^1 g_N(s)^p ds = \frac{N^p}{p(N-1)+1}. \tag{25}$$

The L^2 claim follows at $p = 2$, the essential-supremum claim follows as $p \to \infty$, and the equimeasurable best-of-N rank kernel proves sharpness. □

E Proof of the Minimax Audit Law

Fix $\tau \geq 0$ and abbreviate $g = w \wedge \tau$. Before projection, the estimator in (12) has bias

$$\int (g - w)(\eta - \tfrac{1}{2}) d\mu,$$

whose absolute value is at most $\frac{1}{2}\int (w-\tau)_+ d\mu$. Since $(Y - 1/2)^2 = 1/4$, its variance is at most $\frac{1}{4L}\int g^2 d\mu$. Projection onto $[0,1]$ cannot increase distance to $A_w \in [0,1]$. Optimizing over τ proves $R_L^*(w) \leq \rho_L(w)/4$.

For every $h \in \mathcal{H}_L$ and every τ,

$$\begin{aligned} \int wh &\leq \int (w \wedge \tau) h + \int (w - \tau)_+ \\ &\leq L^{-1/2} \left\| w \wedge \tau \right\|_2 + \int (w - \tau)_+. \end{aligned}$$

Squaring and minimizing over τ gives $\omega_L(w)^2 \leq 2\rho_L(w)$.

For the reverse inequality, first suppose a $\tau > 0$ satisfies

$$\int \min\{1, w/\tau\}^2 d\mu = 1/L. \tag{26}$$

Then $h = \min\{1, w/\tau\}$ lies in $\mathcal{H}_L$. With $B = \int (w-\tau)_+ d\mu$,

$$\int wh = B + \tau/L, \qquad L^{-1} \int (w \wedge \tau)^2 d\mu = (\tau/L)^2.$$

Therefore $\rho_L(w) \leq B^2 + (\tau/L)^2 \leq \omega_L(w)^2$. If no positive threshold solves (26), the support of w has measure at most $1/L$. Then $h = \mathbf{1}\{w > 0\}$ is feasible, $\omega_L(w) = 1$, and the threshold $\tau = 0$ proves the same conclusion.

For the minimax lower bound, fix $h \in \mathcal{H}_L$ and define

$$\eta_\pm = \tfrac{1}{2} \pm \tfrac{1}{4} h.$$

The two target values differ by $\frac{1}{2}\int wh$. Pointwise, their Bernoulli divergence is

$$\frac{h}{2} \log \frac{2+h}{2-h} \leq \frac{2}{3} h^2, \qquad 0 \leq h \leq 1.$$

Hence the L-sample Kullback–Leibler divergence is at most $2/3$, and Pinsker's inequality bounds total variation by $1/\sqrt{3}$. Le Cam's two-point inequality yields

$$R_L^*(w) \geq \frac{1 - 1/\sqrt{3}}{32} \left(\int wh \right)^2.$$

Taking the supremum over h proves Theorem 4.2, with $c = (1 - 1/\sqrt{3})/32$. □

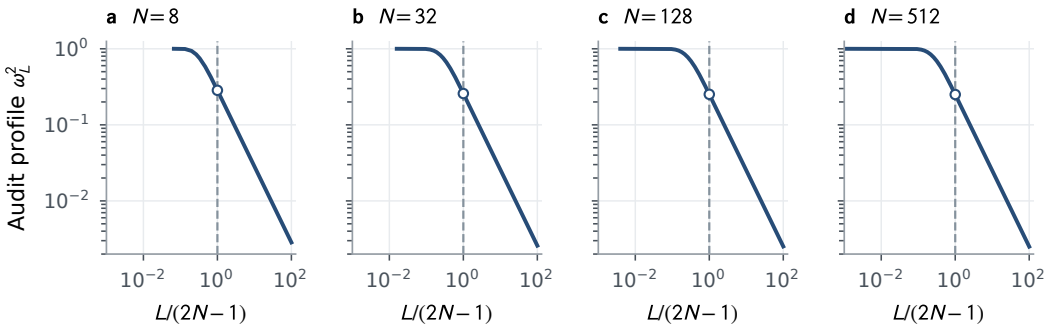


Figure 8: **Best-of-N audit transition.** Each panel shows the exact audit profile $\omega_L(w_N)^2$ versus normalized label budget $L/(2N-1)$ at one candidate budget, on log–log axes. The dashed line marks $L = 2N-1$.

F The Exact Best-of-N Profile and Extremality

For $N \geq 2$, define

$$a_{N,L} = \frac{(2N-1)(L-1)}{(2N-2)L}.$$

Then

$$\omega_L(w_N) = \begin{cases} 1 - \dfrac{N-1}{2N-1} a_{N,L}^N, & 1 \leq L < 2N-1, \\ \dfrac{N}{\sqrt{L(2N-1)}}, & L \geq 2N-1. \end{cases} \tag{27}$$

For $N = 1$, $\omega_L(w_1) = L^{-1/2}$.

The KKT conditions for (9) give an optimizer $h(u) = \min\{1, cw_N(u)\}$. At $L = 1$, $h \equiv 1$ is feasible and the modulus equals one. If $L \geq 2N-1$, the box constraint is inactive, so

$$h = \frac{w_N}{\sqrt{L}\,\|w_N\|_2}, \qquad \omega_L(w_N) = \frac{\|w_N\|_2}{\sqrt{L}} = \frac{N}{\sqrt{L(2N-1)}}.$$

If $1 < L < 2N-1$, let a denote the saturation point. Then $h(u) = (u/a)^{N-1}$ below a and $h(u) = 1$ above it. The norm constraint gives

$$\frac{a}{2N-1} + 1 - a = \frac{1}{L}, \qquad a = a_{N,L},$$

and direct integration gives the first branch of (27).

Figure 8 visualizes (27). Below $L = 2N-1$, the box constraint on the least-observable correctness difference is active. Beyond this point, the profile follows the quadratic regime $\omega_L^2 = N^2/[L(2N-1)]$. The plotted values are computed directly from the closed form, with integer label budgets; they are not averages over different selector families.

For any N-candidate density, Theorem 3.1 gives $w_x \prec g_N$ almost surely. Hence the joint random variable $w_X(Z)$ is majorized by $g_N(U)$ for $U \sim \mathrm{Unif}[0,1]$. The functional ω_L is convex and rearrangement invariant, hence monotone under majorization. Therefore $\omega_L(w) \leq \omega_L(g_N) = \omega_L(w_N)$. Combining (27) with Theorem 4.2 proves (14).

It remains to identify the accuracy endpoint. Conditional on $X = x$ and the candidate pool, the expected correctness of returning candidate i is $\eta_x(Z_i)$, so choosing its maximum is pointwise optimal. If $V_x = \eta_x(Z)$ is atomless, the winner is unique:

$$J^* = \operatorname*{argmax}_{i \leq N} \eta_X(Z_i). \tag{28}$$

A candidate observed at z wins precisely when the other $N-1$ utilities are below $\eta_x(z)$; symmetry gives

$$w_{\mathrm{orc},x}(z) = NG_x(\eta_x(z))^{N-1}. \tag{29}$$

The probability integral transform makes $G_x(\eta_x(Z))$ uniform. Thus the oracle density is conditionally equimeasurable with Nu^{N-1} and g_N, which proves Theorem 4.3. □

G Proof of the Accuracy–Audit Frontier

We record the equivalent penalty formulation and optimality conditions that complement Theorem 5.2. Under its assumptions,

$$\mathcal{F}_{N,L}(r;\eta) = \inf_{\xi \in L^\infty_+(\mu)} \left\{ r \max\{\|\xi\|_\infty, \sqrt{L}\,\|\xi\|_2\} + \mathcal{M}_N(\eta - \xi) \right\}. \tag{30}$$

This dual optimum is also attained. There is a primal–dual optimum (w^*, λ^*, h^*) obeying

$$\begin{aligned} h^* &\in \operatorname*{argmax}_{h \in \mathcal{H}_L} \int w^* h \, d\mu, \\ \lambda^*(\omega_L(w^*) - r) &= 0, \\ w^* &\in \operatorname*{argmax}_{w \in \mathcal{K}_N} \int (\eta - \lambda^* h^*) w \, d\mu. \end{aligned} \tag{31}$$

To state the exact-risk comparison, define

$$\mathcal{F}^{\mathrm{risk}}_{N,L}(\epsilon;\eta) = \sup\{\langle \eta, w\rangle : w \in \mathcal{K}_N,\ R^*_L(w) \le \epsilon^2\}.$$

Then

$$\mathcal{F}_{N,L}(2\epsilon;\eta) \le \mathcal{F}^{\mathrm{risk}}_{N,L}(\epsilon;\eta) \le \mathcal{F}_{N,L}(\epsilon/\sqrt{c};\eta), \tag{32}$$

where c is the universal constant in Theorem 4.2.

We begin with a useful exact representation of the audit radius. For $t, \kappa \ge 0$, define the closed perspective

$$\phi_\kappa(t) = \max_{0 \le u \le 1} \left\{ tu - \frac{\kappa}{2} u^2 \right\} = \begin{cases} t^2/(2\kappa), & 0 \le t \le \kappa, \\ t - \kappa/2, & t > \kappa, \end{cases} \tag{33}$$

with $\phi_0(t) = t$. Lagrange duality for the quadratic constraint in $\mathcal{H}_L$ gives

$$\begin{aligned} \omega_L(w) &= \min_{\alpha \ge 0} \left\{ \frac{\alpha}{L} + \int \sup_{0 \le u \le 1} (wu - \alpha u^2) d\mu \right\} \\ &= \min_{\kappa \ge 0} \left\{ \frac{\kappa}{2L} + \int \phi_\kappa(w) d\mu \right\}. \end{aligned} \tag{34}$$

The integrand is a supremum of affine functions of (w, κ) and is jointly convex. For $\kappa > 0$, its pointwise maximizer is $h = \min\{1, w/\kappa\}$.

We next establish compactness of the N-reachable region in $L^2(\mu)$. Conditional majorization is equivalently described by $w \ge 0$, $\int w_x dP_x = 1$ almost surely, and, for every rational t, the conditional stop-loss inequalities

$$T_t(w)(x) := \int (w_x - t)_+ dP_x \le \ell_N(t) := \int_0^1 (g_N(u) - t)_+ du. \tag{35}$$

The map $u \mapsto (u - t)_+$ is one-Lipschitz, so

$$\|T_t(w) - T_t(v)\|_{L^1(P_X)} \le \|w - v\|_{L^1(\mu)}.$$

Consequently, the normalization, positivity, and stop-loss inequalities pass to an L^2 limit after extracting a common almost-sure subsequence. Thus $\mathcal{K}_N$ is norm closed. It is convex and $0 \le w \le N$, hence weakly compact in the reflexive space $L^2(\mu)$. The set $\mathcal{H}_L$ is weakly compact, and its support function ω_L is weakly lower semicontinuous. The primal feasible set in (15) is therefore weakly compact and its optimum is attained.

Since $\omega_L(1) = r_0$ and $w = 1$ is strictly feasible when $r > r_0$, scalar convex duality gives

$$\mathcal{F}_{N,L}(r;\eta) = \inf_{\lambda \ge 0} \left\{ \lambda r + \sup_{w \in \mathcal{K}_N} [\langle \eta, w\rangle - \lambda \omega_L(w)] \right\}. \tag{36}$$

For fixed λ, Sion's minimax theorem (Sion, 1958) yields

$$\sup_{w\in\mathcal{K}_N}\inf_{h\in\mathcal{H}_L}\langle \eta-\lambda h, w\rangle = \inf_{h\in\mathcal{H}_L}\sup_{w\in\mathcal{K}_N}\langle \eta-\lambda h, w\rangle .$$

For every bounded f,

$$\sup_{w\in\mathcal{K}_N}\langle f, w\rangle = \mathcal{M}_N(f). \tag{37}$$

Indeed, no selector can return a value above the maximum in its observed pool, and poolwise argmax attains equality. Substitution in (36) proves (17).

Set $\xi=\lambda h$. The least multiplier representing a fixed nonnegative ξ is

$$\inf\{\lambda : \xi\in\lambda\mathcal{H}_L\} = \max\{\|\xi\|_\infty, \sqrt{L}\,\|\xi\|_2\}.$$

Minimizing over λ proves (30). For attainment, note that $\mathcal{M}_N$ is the support function of $\mathcal{K}_N$ and is weakly lower semicontinuous. Weak compactness gives attainment on bounded multiplier sets. Evaluating the inner supremum at $w=1$ lower-bounds the dual objective by

$$\int \eta\, d\mu + \lambda(r-r_0),$$

which is coercive because $r>r_0$.

The constant witness $h=r_0$ proves $\omega_L(w)\ge r_0$ for every density. If $w\neq 1$, put $s^2=\|w-1\|_2^2>0$ and, for sufficiently small $\varepsilon>0$, define

$$h_\varepsilon = \frac{r_0+\varepsilon(w-1)}{\sqrt{1+L\varepsilon^2 s^2}}.$$

Because $0\le w\le N$ and $0<r_0<1$, this function lies in $\mathcal{H}_L$, while

$$\int w h_\varepsilon d\mu = \frac{r_0+\varepsilon s^2}{\sqrt{1+L\varepsilon^2 s^2}} > r_0$$

for small positive ε. Thus $w=1$ is unique at r_0. Conversely, Theorem 4.3 proves $\omega_L(w)\le r_N$ throughout $\mathcal{K}_N$, and an extreme-order selector attains r_N. This proves both endpoint claims.

The saddle-point conditions give (31) and complementary slackness. By Theorem 3.1, w^* has an N-pool realization. Put $f=\eta-\lambda^* h^*$. Equality in (37) says that the nonnegative gap

$$\max_{i\le N} f_X(Z_i) - f_X(Z_J)$$

has expectation zero under any realization of w^* and therefore vanishes almost surely. This proves (18).

Finally, Theorem 4.2 gives the inclusions

$$\{\omega_L\le 2\epsilon\}\subseteq\{R_L^*\le\epsilon^2\}\subseteq\{\omega_L\le\epsilon/\sqrt{c}\},$$

which prove (32). □

H Structure of Interior Frontier Points

The frontier's optimality conditions imply more than a penalized score. The next result gives a one-prompt structural characterization. Its interval averaging is related to majorization ironing (Kleiner et al., 2021); the additional complementarity comes from the clipped finite-label audit constraint.

Corollary H.1 (Audit ironing)**.** *Consider one prompt, suppose $\eta(Z)$ is atomless, and let $r_0<r<r_N$. There exists a comonotone primal–dual optimum (w^*,λ^*,h^*) for which the audit constraint is active and the adjusted score $b^*=\eta-\lambda^* h^*$ has an atom. In common decreasing-quantile coordinates, write*

$$a=\eta^{\downarrow},\qquad q=(w^*)^{\downarrow},\qquad b=a-\lambda^*\min\{1,q/\kappa^*\},$$

where $\kappa^*>0$ is chosen so that $\|\min\{1,q/\kappa^*\}\|_2^2=1/L$. Define

$$D(t)=1-(1-t)^N-\int_0^t q(s)ds\ge 0. \tag{38}$$

Then the decreasing càdlàg representative of b obeys

$$\int_{(0,1)} D(t)\, d[-b(t)] = 0. \tag{39}$$

On every connected open interval $I \subseteq \{D > 0\}$, b is constant, say $b = c_I$, and

$$q(t) = \frac{\kappa^*}{\lambda^*}[a(t) - c_I] \quad \textit{for almost every } t \in I. \tag{40}$$

On every open contact interval where $D = 0$, instead $q = g_N$ almost everywhere. Thus on each open component of the contact or slack regions the optimizer has, respectively, an extreme-capacity or audit-ironed affine water-filling form.

Proof. Start from a primal optimum and rearrange its density comonotonically with η. This preserves its distribution, hence both the majorization and audit constraints, and weakly increases accuracy by Hardy–Littlewood. We may therefore choose an optimal density q decreasing in the same coordinate as $a = \eta^{\downarrow}$. Hard-pool capacity forces $q > 0$ almost everywhere: otherwise its strict-subset support would have selected mass one, contradicting (4). Since $L > 1$, the quadratic constraint in (9) is active, its multiplier satisfies $\kappa^* > 0$, and the audit witness is the unique canonical solution $h = \min\{1, q/\kappa^*\}$. Thus it is comonotone with q.

We record the remaining simultaneous-comonotonicity argument. The KKT condition says that q maximizes $\int bv$ over $v \prec g_N$, where $b = a - \lambda^* h$. Every equimeasurable rearrangement of q remains feasible. If two equal-measure sets had strictly larger q but strictly smaller b on the first set, swapping their q values would preserve feasibility and strictly increase $\int bq$, a contradiction. Thus q and b are comonotone. On each level set of q, the witness h is constant, so b inherits the decreasing ordering of a. Choosing this representative makes a, q, h, b simultaneously decreasing.

Put $G_N(t) = 1 - (1-t)^N$, $Q(t) = \int_0^t q$, and $D = G_N - Q$. Majorization gives $D \geq 0$. Because b is decreasing, selecting the largest b in each pool shows that g_N is a support maximizer; q is another support maximizer by the KKT condition. Their objective values agree, and integration by parts gives

$$0 = \int_0^1 b(t)[g_N(t) - q(t)]dt = \int_{(0,1)} D(t) d[-b(t)].$$

Both factors in the Stieltjes integral are nonnegative. Therefore b is constant on each connected component I of $\{D > 0\}$. There

$$a - \lambda^* \min\{1, q/\kappa^*\} = c_I.$$

The saturated branch $q \geq \kappa^*$ on a positive-measure subset would force $a = c_I + \lambda^*$ there, contradicting atomlessness. This proves (40). On an open contact interval, absolute continuity of $D = 0$ gives $q = g_N$ almost everywhere.

If the audit constraint were slack, or if b were atomless, the poolwise maximizer would be the unique extreme rearrangement of g_N and would have radius r_N. Hence $r_0 < r < r_N$ forces $\lambda^* > 0$, an active constraint, and a positive-mass plateau of b. □

I Why the Quadratic Surrogate Can Fail

In the high-coverage regime, the audit profile is quadratic by (13). The following result shows that replacing it globally by that quadratic expression can change the frontier by a polynomial factor, even after imposing the exact finite-pool capacity region.

Proposition I.1 (Polynomial failure of the L^2 surrogate)**.** *Let $N = m^2$, $L = m$, and $m \geq 16$. There are an atomless one-prompt base law, a binary correctness function η, and a valid kernel $w \in \mathcal{K}_N$ such that*

$$\omega_L(w) \leq \frac{3}{2\sqrt{m}}, \qquad \frac{\|w\|_2^2}{L} \geq \frac{1}{4}. \tag{41}$$

Thus $R_L^*(w) = \Theta(m^{-1})$, although the untruncated quadratic proxy is constant. More strongly, at $r = 2/\sqrt{m}$, define

$$\widetilde{\mathcal{F}}_{N,L}(r;\eta) = \sup\{\langle \eta, v\rangle : v \in \mathcal{K}_N,\ \|v\|_2^2 \leq Lr^2\}.$$

Then

$$\mathcal{F}_{N,L}(r;\eta) \geq \frac{1}{2m}, \qquad \widetilde{\mathcal{F}}_{N,L}(r;\eta) \leq \frac{2}{m^{3/2}}. \tag{42}$$

The exact frontier is therefore larger by a factor at least $\sqrt{m}/4 = N^{1/4}/4$.

Proof. Take the base space $[0,1]$, choose a measurable set A of mass $p = m^{-3}$, and put $\eta = \mathbf{1}_A$. Define

$$a = \frac{m^2}{2}, \qquad b = \frac{1-ap}{1-p}, \qquad w = a\mathbf{1}_A + b\mathbf{1}_{A^c}.$$

Arrange A first in decreasing coordinates and let $Q(t) = \int_0^t w^{\downarrow}$. With $G(t) = 1-(1-t)^{m^2}$, the function $G - Q$ is concave separately on $[0,p]$ and $[p,1]$. Its endpoint values are nonnegative: at zero and one they vanish, while

$$G(p) \geq m^2 p - \binom{m^2}{2} p^2 \geq \frac{1}{m} - \frac{1}{2m^2} \geq ap.$$

Concavity therefore gives $Q(t) \leq G(t)$ everywhere, so $w \in \mathcal{K}_N$ by Theorem 3.1.

The exact audit witness equals one on A and equals

$$c_m = \sqrt{\frac{1/m - p}{1-p}} = \sqrt{\frac{m^2-1}{m^3-1}} \leq \frac{1}{\sqrt{m}}$$

on A^c. It has squared norm $1/m$ and is of the KKT form $\min\{1, w/\kappa\}$ with $\kappa = b/c_m < a$, so it is optimal. Consequently,

$$\omega_L(w) = ap + (1-ap)c_m \leq \frac{3}{2\sqrt{m}}, \qquad \frac{\|w\|_2^2}{L} \geq \frac{pa^2}{m} = \frac{1}{4}.$$

The universal lower bound $\omega_L(w) \geq m^{-1/2}$ and Theorem 4.2 give the stated risk order. At $r = 2/\sqrt{m}$, the same kernel is feasible and has accuracy $ap = 1/(2m)$. Conversely, every v in the quadratic-surrogate class has $\|v\|_2 \leq 2$, and therefore

$$\int_A v \leq \sqrt{P(A)}\,\|v\|_2 \leq 2m^{-3/2}.$$

This proves (42). □

J Certified Computation and Plug-In Stability

The exact frontier admits a finite convex approximation with a deterministic certificate. This separates statistical exactness from the possible complexity of realizing an arbitrary measurable fixed-pool rule.

Proposition J.1 (Certified quantile discretization)**.** *Consider one prompt. Let* $a = \eta^{\downarrow}$*, partition* $[0,1]$ *into* M *equal bins* I_j*, and put* $a_j = M\int_{I_j} a$*. Let* $\mathcal{F}^{(M)}_{N,L}(r;\eta)$ *be the optimum of*

$$\begin{aligned} \max_{q,\kappa} \quad & \frac{1}{M}\sum_{j=1}^{M} a_j q_j \\ \textit{subject to} \quad & q_1 \geq \cdots \geq q_M \geq 0, \qquad \frac{1}{M}\sum_{j=1}^{M} q_j = 1, \\ & \frac{1}{M}\sum_{j=1}^{k} q_j \leq 1-(1-k/M)^N, \quad k < M, \\ & \frac{\kappa}{2L} + \frac{1}{M}\sum_{j=1}^{M} \phi_\kappa(q_j) \leq r, \qquad \kappa \geq 0. \end{aligned} \tag{43}$$

This is a finite convex program, every feasible grid density is realizable by an N*-pool protocol, and*

$$0 \leq \mathcal{F}_{N,L}(r;\eta) - \mathcal{F}^{(M)}_{N,L}(r;\eta) \leq \min\{1, N/M\}. \tag{44}$$

Proof. Let P_M denote conditional expectation onto the equal-bin partition. By Hardy–Littlewood, take an optimal frontier density q decreasing with a, and set $q_M = P_M q$. It remains decreasing and preserves total mass. At bin endpoints its cumulative integral equals that of q; between endpoints it is the corresponding chord. Since $G_N(t) = 1 - (1-t)^N$ is concave, every majorization inequality remains valid. Moreover,

$$\omega_L(P_M q) = \sup_{h \in \mathcal{H}_L} \langle q, P_M h \rangle \le \omega_L(q),$$

because $P_M h \in \mathcal{H}_L$ by Jensen. Thus q_M is feasible for (43), using (34). Finally,

$$\int aq - \int aq_M = \int (a - P_M a) q \le N \left\| a - P_M a \right\|_1 \le \frac{N}{M}.$$

The last inequality uses $0 \le a \le 1$ and monotonicity, whose total variation across the bins is at most one. The trivial accuracy bound one gives (44). Joint convexity follows from (33), and Theorem 3.1 gives implementability. □

Proposition J.2 (Plug-in stability)**.** *Let $1 \le p \le \infty$ and let q be its conjugate exponent. If $\left\| \widehat{\eta} - \eta \right\|_p \le \varepsilon$, then every $w \in \mathcal{K}_N$ satisfies*

$$| \langle \widehat{\eta} - \eta, w \rangle | \le \varepsilon C_{N,q}, \qquad C_{N,q} = \frac{N}{[q(N-1)+1]^{1/q}}, \tag{45}$$

with $C_{N,\infty} = N$. If $\widehat{w}$ maximizes the frontier under $\widehat{\eta}$ and w^* maximizes it under η, at the same (N, L, r), then

$$\langle \eta, w^* \rangle - \langle \eta, \widehat{w} \rangle \le 2\varepsilon C_{N,q}. \tag{46}$$

Proof. Hölder's inequality and (25) give

$$| \langle \widehat{\eta} - \eta, w \rangle | \le \left\| \widehat{\eta} - \eta \right\|_p \left\| w \right\|_q \le \varepsilon C_{N,q}.$$

Applying this inequality to w^* and $\widehat{w}$, and using optimality of $\widehat{w}$ under $\widehat{\eta}$, proves (46). □

K From Finite Candidate Banks to Executable Selectors

The experiments draw candidate pools from frozen empirical distributions. This section connects those atomic distributions to the selected-law formulation and gives the construction used to execute the fitted grid policies. The subset-capacity argument is finite-dimensional; the subsequent decomposition uses classical majorization (Marshall et al., 2011). Neither step requires treating an empirical score distribution as continuous.

K.1 Exact capacity on an atomic candidate space

Fix a prompt and let its candidate types be $[K] = \{1, \dots, K\}$, with base probabilities $p = (p_1, \dots, p_K)$. Types can be individual candidate records, even when some records share the same verifier score. Draw $Z_1, \dots, Z_N$ independently from p, with replacement. A selector returns one drawn occurrence and induces an output probability vector q. For $A \subseteq [K]$, write $p(A) = \sum_{j \in A} p_j$ and define $q(A)$ similarly. Let $G_N(t) = 1 - (1-t)^N$.

Proposition K.1 (Finite-bank selected laws)**.** *The achievable output probability vectors are exactly*

$$\mathcal{C}_N(p) = \left\{ q \in \mathbb{R}_+^K : \sum_{j=1}^K q_j = 1, \quad q(A) \le G_N\big(p(A)\big) \quad \textit{for all } A \subseteq [K] \right\}. \tag{47}$$

Moreover, every $q \in \mathcal{C}_N(p)$ is a mixture of deterministic priority-selector laws. For a permutation σ of $[K]$, the priority selector returns the first present type in the order $\sigma(1), \dots, \sigma(K)$; its law is

$$q^\sigma_{\sigma(k)} = (1 - P^\sigma_{k-1})^N - (1 - P^\sigma_k)^N, \qquad P^\sigma_k = \sum_{i=1}^k p_{\sigma(i)}, \quad P^\sigma_0 = 0. \tag{48}$$

Proof. A selected type can belong to A only if at least one candidate belongs to A. This gives (47) by independence. In particular, a type with $p_j = 0$ necessarily has $q_j = 0$.

For sufficiency, let $a \in \mathbb{R}^K$ be any linear reward vector and choose σ such that $a_{\sigma(1)} \geq \cdots \geq a_{\sigma(K)}$. For $A_k = \{\sigma(1), \ldots, \sigma(k)\}$, summation by parts gives

$$\begin{aligned} a^\top q &= a_{\sigma(K)} + \sum_{k=1}^{K-1} \big(a_{\sigma(k)} - a_{\sigma(k+1)}\big) q(A_k) \\ &\leq a_{\sigma(K)} + \sum_{k=1}^{K-1} \big(a_{\sigma(k)} - a_{\sigma(k+1)}\big) G_N(P_k^\sigma). \end{aligned}$$

The priority selector with order σ attains equality at every prefix simultaneously: it returns a type in A_k precisely when the pool hits A_k. Subtracting successive prefix probabilities gives (48). Thus $\mathcal{C}_N(p)$ and the convex hull of the finitely many vectors q^σ have identical support functions. Both are nonempty compact convex sets, so they are equal. To realize a convex combination, sample its priority order independently of the pool and then execute that order. If a winning type appears more than once, any choice among its occurrences has the same type law. □

This proves the subset criterion directly for atomic banks. For a nonuniform p, it is this weighted subset criterion that should be used; ordinary unweighted vector majorization is recovered by equal base probabilities, as follows.

K.2 An equal-bin realization with a compact randomized program

Let $U_1, \ldots, U_N$ be independent uniform ranks on $[0, 1]$. Partition the interval into M equal bins $I_j = [(j-1)/M, j/M)$, assigning the endpoint 1 to the last bin. A grid density has value w_j on I_j, so its desired selected bin mass is $b_j = w_j/M$. Define the decreasing vector

$$v_j = \left(1 - \frac{j-1}{M}\right)^N - \left(1 - \frac{j}{M}\right)^N, \qquad j = 1, \ldots, M. \tag{49}$$

Here $j = 1$ denotes the highest *priority position*, not the largest verifier rank. A permutation assigns these priority positions to score bins; in particular, high-score bins can receive high priority.

For probability vectors, $b \prec v$ means $\sum_{j=1}^k b_j^\downarrow \leq \sum_{j=1}^k v_j$ for $k < M$ and equal total mass. Since the largest mass of a subset of k bins is the sum of the k largest b_j, Theorem K.1 gives

$$b \text{ is realizable} \iff b \prec v, \qquad \sum_{j=1}^k v_j = G_N(k/M). \tag{50}$$

These endpoint inequalities also certify the continuous majorization constraint for the bin-constant density: between successive endpoints its decreasing rearrangement has a linear cumulative integral, which lies below the corresponding concave function G_N.

Proposition K.2 (Compact priority realization)**.** *If $b \prec v$, there are a permutation matrix P, coordinate transpositions $\mathsf{S}_1, \ldots, \mathsf{S}_s$, and numbers $t_1, \ldots, t_s \in [0, 1]$, with $s \leq M - 1$, such that*

$$b = \mathsf{P}\mathsf{T}_s \cdots \mathsf{T}_1 v, \qquad \mathsf{T}_i = (1 - t_i)\mathsf{I} + t_i \mathsf{S}_i. \tag{51}$$

Independently for each pool, sample independent Bernoulli(t_i) coins, replace each T_i by S_i or I accordingly, and use the resulting permutation of priority positions to select a present bin. Selecting its first occurrence gives rank density w. After factorization and bin lookup preprocessing, executing one pool requires $O(M + N)$ operations and $O(M)$ auxiliary storage.

Proof. We first recall the constructive two-coordinate argument behind (51). Suppose descending vectors $x \prec y$ of length m have the same total mass. If $m = 1$, they coincide. Otherwise $x_1 \leq y_1$ by majorization and $x_1 \geq y_m$ because x_1 is at least the common coordinate average. If $x_1 = y_k$ for some k, freeze that coordinate. Otherwise choose adjacent entries with $y_k > x_1 > y_{k+1}$ and apply a two-coordinate mix with

$$t = \frac{y_k - x_1}{y_k - y_{k+1}}.$$

It replaces (y_k, y_{k+1}) by (x_1, d), where $d = y_k + y_{k+1} - x_1 \in [y_{k+1}, y_k]$. Freeze the coordinate equal to x_1.

After removing this coordinate, the remaining decreasing vector is $(y_1, \dots, y_{k-1}, d, y_{k+2}, \dots, y_m)$. For a prefix of length $\ell < k$, all its entries are at least x_1, and hence its sum is at least $x_2 + \cdots + x_{\ell+1}$. For $\ell \geq k$, its sum is $y_1 + \cdots + y_{\ell+1} - x_1$, which is at least $x_2 + \cdots + x_{\ell+1}$ by $x \prec y$. Total remaining masses agree. The same argument applies when an exact matching coordinate is simply removed. Induction therefore needs at most $m - 1$ mixes. Keeping track of the original coordinate names and the frozen coordinates gives (51), including the final output permutation. This is the classical T-transform construction, here applied to the priority masses in (49).

For the randomized implementation, let $\mathsf{R}_i \in \{\mathsf{I}, \mathsf{S}_i\}$ be the sampled transposition. Each realization of $\mathsf{PR}_s \cdots \mathsf{R}_1$ is a permutation, hence specifies a valid priority selector with bin law $\mathsf{PR}_s \cdots \mathsf{R}_1 v$. Independence gives

$$\mathbb{E}[\mathsf{PR}_s \cdots \mathsf{R}_1 v] = \mathsf{PT}_s \cdots \mathsf{T}_1 v = b.$$

Conditional on all candidate bin indices, each rank remains uniform within its bin. The priority rule and the first-occurrence tie rule use only those indices and independent coins. The selected rank is therefore uniform within its winning bin, giving density $Mb_j = w_j$.

For execution, initialize an array of M priority positions, perform at most $M - 1$ sampled exchanges, and scan the N candidate bin indices for the highest priority. The stored exchanges, their probabilities, the output map, and the working array each use $O(M)$ space. The cost includes $O(M)$ random coins but excludes the one-time factorization and the cost of producing or scoring the candidates. □

Fresh independent coins are drawn for every pool. Reusing one sampled priority permutation across a batch preserves a marginal averaged over that permutation, but couples the trials and does not produce the independent repetitions used in our accuracy and error calculations.

K.3 Score ties and exact marginalization back to candidate records

Return to an atomic bank with verifier score s_j for type j and let $S = s_Z$. Write F_S for its base score CDF. For each drawn occurrence, generate an independent $V \sim \mathrm{Unif}[0, 1]$ and define

$$U = F_S(S-) + V\big(F_S(S) - F_S(S-)\big). \tag{52}$$

For a type with $p_j > 0$, put $a_j = F_S(s_j-)$ and $c_j = F_S(s_j)$, so $c_j - a_j > 0$ is the total base mass of its score tie. Each tie maps uniformly onto an interval whose length equals that tie's probability. These intervals partition $[0, 1]$ up to endpoints, proving that U is uniform. The auxiliary draw is repeated even when the same candidate record appears twice in a pool. Tie midpoints do not have this uniform distribution.

Proposition K.3 (Finite-bank marginals of a rank policy)**.** *Apply the bin-based priority construction above to the randomized ranks in* (52)*. Its selected density relative to the original bank is*

$$\bar{w}_j = \frac{1}{c_j - a_j} \int_{a_j}^{c_j} w(u)\, du, \qquad Q(Z = j) = p_j \bar{w}_j. \tag{53}$$

In particular, all records in a score tie receive the same density ratio. For correctness means η_j, its accuracy is exactly

$$A = \sum_{j=1}^{K} p_j \bar{w}_j \eta_j = \int_0^1 w(u) \eta_{\mathrm{bank}}(u)\, du, \qquad \eta_{\mathrm{bank}}(u) = \mathbb{E}[\eta_Z \mid U = u]. \tag{54}$$

Proof. Conditional on the candidate bin indices, the chosen index depends only on those indices and independent priority coins. Within each bin its selected augmented record (Z, V) thus has the base conditional law. Because bin j receives total mass b_j, the density on this augmented space is $w(U)$. Integrating out V, which is uniform on the tie interval conditional on a candidate record, gives (53). Multiplication by η_Z followed by conditional expectation given U gives (54). □

For the uniform empirical banks, $p_j = 1/K$. A bin-constant w is integrated exactly over every tie interval, including partial overlaps with bin boundaries. No normalization estimated from selected

winners is needed: (53) already gives $\sum_j p_j \bar{w}_j = 1$. In the supplementary primary grid, $K \in \{128, 256, 512\}$ and $M = \max\{512, 32N\}$ with $N \in \{8, 32, 64\}$, so M is a multiple of K and all tie endpoints also align with grid boundaries.

K.4 Audit guarantees after removing the auxiliary rank

Let $\widetilde{\mu}$ be the joint base law of (X, Z, V) and put $\widetilde{w}(X, Z, V) = w_X(U)$. The marginal density on (X, Z) is $\bar{w} = \mathbb{E}[\widetilde{w} \mid X, Z]$. To distinguish the underlying measures in this subsection, write $\omega_{L,\nu}$ for the modulus in (9) computed under a base law ν.

Lemma K.4 (Marginalization does not increase audit radius)**.** *For $L \geq 1$,*

$$\omega_{L,\mu}(\bar{w}) \leq \omega_{L,\widetilde{\mu}}(\widetilde{w}) = \omega_{L,P_X \otimes du}(w). \tag{55}$$

Consequently, a rank policy with radius at most r admits the centered clipped estimator on the finite bank with worst-case MSE at most $r^2/4$.

Proof. Every witness $h(X, Z)$ satisfying $0 \leq h \leq 1$ and $\mathbb{E}_\mu h^2 \leq 1/L$ is also admissible on the augmented space, with the same norm. Moreover,

$$\mathbb{E}_\mu[\bar{w}h] = \mathbb{E}_{\widetilde{\mu}}[\widetilde{w}h].$$

Taking suprema proves the inequality. For the equality, average any augmented witness over (X, U). Its range remains in $[0, 1]$, its squared norm cannot increase by conditional Jensen, and its pairing with $\widetilde{w} = w_X(U)$ is unchanged. Conversely, every witness depending only on (X, U) lifts to the augmented space. Since U is conditionally uniform for every prompt, the two suprema are equal.

The proof of Theorem 4.2 uses integration and the binary-label identity $(Y - 1/2)^2 = 1/4$, not nonatomicity. It therefore applies to the finite bank with density $\bar{w}$, giving

$$R^*_L(\bar{w}) \leq \tfrac{1}{4}\rho_{L,\mu}(\bar{w}) \leq \tfrac{1}{4}\omega_{L,\mu}(\bar{w})^2 \leq \tfrac{1}{4}r^2.$$

Here $\rho_{L,\mu}$ is (10) under μ, and the clipping threshold is fixed from the known law and label budget, independently of the audit draws. □

The finite-bank radius can be strictly smaller than the augmented rank radius. For example, if all candidate scores tie, every rank-only policy has $\bar{w} \equiv 1$, even if w is highly concentrated within the single tie interval. The inequality above therefore supplies a sufficient evaluation budget rather than identifying the two radii. The same finite-sample upper bound holds for the deterministic operational labels in the frozen banks. The minimax lower bound in Theorem 4.2, by contrast, ranges over arbitrary Bernoulli correctness kernels; replaying one deterministic label bank is not a lower-bound experiment for that larger statistical class.

K.5 Pooling cross-fitted policies under a common error budget

A task-level audit mixes the held-out folds, each with its own fitted policy. Their separate radius constraints need not imply the same constraint on the pooled modulus. The clipping representation nevertheless preserves the common error guarantee for a pooled audit with L total labels.

Lemma K.5 (Foldwise certification of a pooled audit)**.** *Let the base law first draw a group f with probability α_f, then draw from μ_f, where $\sum_f \alpha_f = 1$. Let w_f be a density under μ_f, fixed independently of the audit draws, and let $w(f, z) = w_f(z)$ be the resulting joint density under μ. For the same total label budget L,*

$$\rho_{L,\mu}(w) \leq \sum_f \alpha_f \rho_{L,\mu_f}(w_f) \leq \sum_f \alpha_f \omega_{L,\mu_f}(w_f)^2. \tag{56}$$

In particular, if every group has radius at most r, the pooled centered clipped estimator with its globally optimal scalar threshold has worst-case MSE at most $r^2/4$ using L total base labels.

Proof. For any density w, define

$$J_\mu(g) = \frac{1}{L}\mathbb{E}_\mu g^2 + \big[\mathbb{E}_\mu(w - g)\big]^2, \qquad 0 \leq g \leq w.$$

We first show that allowing arbitrary retained weights g does not improve on scalar clipping. There is a unique $\tau_* > 0$ satisfying $\tau_*/L = \mathbb{E}_\mu(w - \tau_*)_+$: the difference of the two sides is continuous and strictly increasing, is -1 at zero, and is nonnegative at $\tau_* = L$. Put $g_* = w \wedge \tau_*$. For any admissible g with finite second moment, expanding the two objectives gives

$$J_\mu(g) - J_\mu(g_*) = \frac{1}{L}\mathbb{E}_\mu(g - g_*)^2 + \left[\mathbb{E}_\mu(g - g_*)\right]^2 + \frac{2}{L}\mathbb{E}_\mu[(g_* - \tau_*)(g - g_*)] \;\geq\; 0.$$

The last integrand is nonnegative: on $\{w < \tau_*\}$ both factors are nonpositive, and elsewhere its first factor is zero. Infinite second moment makes the inequality immediate. Consequently $\rho_{L,\mu}(w) = \inf_{0 \leq g \leq w} J_\mu(g)$.

In group f, take its optimal clipped weight g_f and let $B_f = \mathbb{E}_{\mu_f}(w_f - g_f)$. The piecewise weight $g(f, z) = g_f(z)$ is admissible for the pooled problem. Jensen's inequality gives

$$\rho_{L,\mu}(w) \leq \frac{1}{L}\sum_f \alpha_f \mathbb{E}_{\mu_f} g_f^2 + \left(\sum_f \alpha_f B_f\right)^2 \leq \sum_f \alpha_f \left\{\frac{1}{L}\mathbb{E}_{\mu_f} g_f^2 + B_f^2\right\} = \sum_f \alpha_f \rho_{L,\mu_f}(w_f).$$

The second inequality in (56) and the MSE bound follow from Theorem 4.2. □

For the cross-fitted task, α_f is its fraction of problems in fold f. Apply Theorem K.4 within each fold and then Theorem K.5. The common rank-space design budget therefore certifies the pooled clipped estimator, including when its exact pooled modulus is slightly larger than that budget. These are different certificates: a bound on ρ does not assert the same bound on ω^2. All statements condition on the fitted fold-specific rules; they do not include uncertainty from refitting those rules.

K.6 Fitting, numerical certificates, and the target of evaluation

For each held-out fold, the empirical construction first forms each training prompt's tie-integrated correctness profile, averages these profiles with equal prompt weights, and fits an increasing isotonic profile $\widehat{\eta}$. It then optimizes the rank density using only that fitted profile and the prescribed (N, L, r). Test-prompt labels enter (54) and the audit calculation, but not the profile fit, priority factorization, pool selection, or tie-breaking for that prompt. The exact finite-bank expectation and independently randomized execution provide two calculations of the same selected accuracy. Aggregating cross-fitted policies gives each prompt its own out-of-fold rule; it does not turn folds or repeated pools into independent research samples.

All identities above concern feasible probabilities in exact arithmetic. The numerical implementation retains both the original solver density and the density actually executed. Any uniform-mixture feasibility repair, its change in the fitted objective, the reconstruction residual in (51), and the final primal–dual gap are recorded separately. Thus solver tolerance, implementation accuracy, and statistical uncertainty are distinct quantities.

The optimized empirical curve is the frontier of its fitted verifier-rank class. It is not the correctness oracle over every prompt-dependent candidate-record policy in (47). In particular, the N/M correctness-quantile approximation guarantee in Theorem J.1 is not an error certificate for an arbitrary empirical verifier-rank fit. Likewise, matching a rank-space radius matches a worst-case guarantee, not the realized MSE of two policies. The replay calculations condition on the frozen banks and the fitted rules: new independent pool draws are draws from those empirical laws, rather than new generations from the underlying language model.

L Evaluation by Black-Box Replay

L.1 Construction and resource costs

Shadow replay estimates a labeled response's density ratio by repeatedly inserting it into fresh pools and recording how often it is selected.

For one pool, let $p_i = \mathbb{P}(J = i \mid X, Z_{1:N})$ be the selector's conditional choice probability. For each labeled base response (X_ℓ, Z_ℓ, Y_ℓ), called the *anchor*, and each $b \le B$, draw $N-1$ companions from P_{X_ℓ} to form a pool with the anchor. Let $p_{\ell b}$ be the anchor's conditional selection probability. If this probability is exposed, set $S = \infty$ and $\widetilde{p}_{\ell b} = p_{\ell b}$. If only sampled selections are exposed, let $\widetilde{p}_{\ell b}$ average S independent replays of the same pool. Use $1/S = 0$ at $S = \infty$ and define

$$\widetilde{W}_{\ell,B,S} = \frac{N}{B}\sum_{b=1}^{B}\widetilde{p}_{\ell b},$$

$$\widetilde{A}_{\mathrm{sh}} = \frac{1}{2} + \frac{1}{L}\sum_{\ell=1}^{L}\widetilde{W}_{\ell,B,S}(Y_\ell - \tfrac{1}{2}), \qquad \widehat{A}_{\mathrm{sh}} = \Pi_{[0,1]}(\widetilde{A}_{\mathrm{sh}}). \tag{57}$$

For a stateless permutation-equivariant protocol fixed independently of the audit data, with fresh conditionally independent companions and replay randomness, exchangeability gives

$$\mathbb{E}[\widetilde{W}_{\ell,B,S} \mid X_\ell, Z_\ell] = w_{\pi,X_\ell}(Z_\ell), \qquad \mathbb{E}\widetilde{A}_{\mathrm{sh}} = A_\pi. \tag{58}$$

To describe its variance, define

$$m_2 = \mathbb{E}[w_{\pi,X}(Z)^2], \qquad \gamma_\pi = N\mathbb{E}\left[\sum_{i=1}^{N} p_i^2\right],$$

$$\Lambda_{B,S} = m_2 + \frac{\gamma_\pi - m_2}{B} + \frac{N - \gamma_\pi}{BS}. \tag{59}$$

m_2 measures how unevenly the selected law amplifies base responses, while γ_π measures how concentrated the selector's choice is within a fixed pool. The quantity $\Lambda_{B,S}$ combines these two sources with the finite numbers of companion pools and internal replays. Theorem L.1 below states the full assumptions and proves $1 \le m_2 \le \gamma_\pi \le N$ together with

$$\mathbb{E}[\widetilde{W}^2_{\ell,B,S}] = \Lambda_{B,S},$$

$$\mathbb{E}(\widetilde{A}_{\mathrm{sh}} - A_\pi)^2 = \frac{1}{L}\left\{\frac{\Lambda_{B,S}}{4} - (A_\pi - \tfrac{1}{2})^2\right\}, \tag{60}$$

$$\mathbb{E}(\widehat{A}_{\mathrm{sh}} - A_\pi)^2 \le \min\left\{1, \frac{\Lambda_{B,S}}{4L}\right\}. \tag{61}$$

Companion generation costs $LB(N-1)$ candidates. A sampled interface uses LBS selector calls; a deterministic or probability-exposing interface uses LB calls.

The three terms in (59) isolate distinct resources. More labels reduce the outer $1/L$ factor. More companion pools remove conditional Monte Carlo noise and recover m_2 as $B \to \infty$. More internal replays reveal whether a randomized selector is genuinely diffuse; at $S = 1$, the one-pool second moment is N regardless of γ_π. For a deterministic extreme-order rule, $\gamma_\pi = N$ and $m_2 = N^2/(2N-1)$, so the label rate remains $\Theta(1 \wedge N/L)$.

The same observed anchor scores give an empirical-Bernstein confidence interval without analytic density ratios. Its radius includes a sampling variance term and a range term proportional to $N/(L-1)$. The derivation below establishes coverage for $L \ge 2$.

L.2 Variance and confidence guarantees

Theorem L.1 (Shadow-pool audit and exact variance decomposition)**.** *Let $L, B \in \mathbb{N}$ and $S \in \mathbb{N} \cup \{\infty\}$. The anchors are i.i.d. base draws. The permutation-equivariant N-candidate protocol*

is stateless, fixed independently of the audit data, and does not observe anchor labels. Conditional on each anchor prompt, companions are independent draws from the same P_x; shadow pools are conditionally independent across (ℓ, b); their transcripts are independent of Y_ℓ given (X_ℓ, Z_ℓ); and sampled replays use fresh independent protocol randomness. If selection probabilities are unavailable, $\widetilde{p}_{\ell b}$ is the average of S independent sampled selections. Then (58)–(61) *hold with $\Lambda_{B,S}$ from* (59)*. In particular, the unprojected estimator is unbiased and has the exact variance in* (60)*; projection is used only for the risk bound. Companion generation costs $LB(N-1)$ candidates, and the selector-call costs are LBS for sampled selections and LB for deterministic or probability-exposing interfaces.*

For an observed pool, write $p_i = \mathbb{P}(J = i \mid X, Z_{1:N})$. Permutation equivariance and exchangeability give, for every measurable $C \subseteq \mathcal{Z}$ and almost every x,

$$\begin{aligned} Q_{\pi,x}(C) &= \mathbb{E}\left[\sum_{i=1}^{N} p_i \mathbf{1}\{Z_i \in C\} \mid X = x\right] \\ &= N \int_C \mathbb{E}[p_1 \mid X = x, Z_1 = z] dP_x(z). \end{aligned}$$

Comparison of Radon–Nikodym derivatives yields the forward representation

$$w_{\pi,x}(z) = N\mathbb{E}[p_1 \mid X = x, Z_1 = z]. \tag{62}$$

The anchor and its companions have exactly the distribution on the right-hand side. Averaging independent companion pools and, when needed, unbiased internal-randomness replays proves the first identity in (58).

Consider one shadow pool and put $\widetilde{R} = N\widetilde{p}_1$. For a sampled interface with $S < \infty$, conditional on that pool,

$$\mathbb{E}[\widetilde{R}^2 \mid X, Z_{1:N}] = N^2 \left\{ p_1^2 + \frac{p_1(1-p_1)}{S} \right\}.$$

Exchangeability gives

$$N^2\mathbb{E}[p_1^2] = N\mathbb{E}\left[\sum_{i=1}^{N} p_i^2\right] = \gamma_\pi, \qquad N^2\mathbb{E}[p_1] = N.$$

The same calculation holds at $S = \infty$ after setting $1/S = 0$. Conditional variance decomposition for the average of B pools gives

$$\begin{aligned} \mathbb{E}[\widetilde{W}^2_{\ell,B,S}] &= \mathbb{E}[w_{\pi,X}(Z)^2] + \frac{1}{B}\left\{ \gamma_\pi + \frac{N - \gamma_\pi}{S} - m_2 \right\} \\ &= \Lambda_{B,S}. \end{aligned}$$

Furthermore, $m_2 \geq (\mathbb{E}w)^2 = 1$. Conditional Jensen applied to (62) gives $m_2 \leq \gamma_\pi$, and $1/N \leq \sum_i p_i^2 \leq 1$ gives $1 \leq \gamma_\pi \leq N$.

Protocol randomness and companions are independent of the anchor label given (X_ℓ, Z_ℓ). Therefore

$$\begin{aligned} \mathbb{E}\widetilde{A}_{\mathrm{sh}} &= \frac{1}{2} + \mathbb{E}[w_{\pi,X}(Z)(Y - \tfrac{1}{2})] \\ &= \frac{1}{2} + A_\pi - \frac{1}{2}\mathbb{E}[w_{\pi,X}(Z)] = A_\pi, \end{aligned}$$

where promptwise normalization gives $\mathbb{E}w_{\pi,X}(Z) = 1$. Since $(Y - 1/2)^2 = 1/4$, one anchor score has variance

$$\frac{1}{4}\Lambda_{B,S} - (A_\pi - \tfrac{1}{2})^2.$$

The anchor scores are i.i.d., proving (60). Projection onto $[0,1]$ cannot increase distance to $A_\pi \in [0,1]$, and the squared distance between two points in this interval is at most one. This proves (61). □

Confidence interval. For $L \geq 2$, define an observable confidence interval from the anchor scores. Let $\mathsf{T}_\ell = \frac{1}{2} + \widetilde{W}_{\ell,B,S}(Y_\ell - \frac{1}{2})$, let $\overline{\mathsf{T}} = L^{-1}\sum_\ell \mathsf{T}_\ell$ and $\widehat{v} = (L-1)^{-1}\sum_\ell(\mathsf{T}_\ell - \overline{\mathsf{T}})^2$, and put $a = \log(4/\delta)$. The empirical-Bernstein radius

$$r_\delta = \sqrt{\frac{2\widehat{v}a}{L}} + \frac{7Na}{3(L-1)} \tag{63}$$

satisfies $A_\pi \in [\overline{\mathsf{T}} - r_\delta, \overline{\mathsf{T}} + r_\delta] \cap [0,1]$ with probability at least $1-\delta$.

The scores $\mathsf{T}_1, \ldots, \mathsf{T}_L$ are i.i.d. and each lies in $[\frac{1}{2} - \frac{N}{2}, \frac{1}{2} + \frac{N}{2}]$, an interval of width N. Apply the empirical Bernstein inequality of Maurer & Pontil (2009) after affine rescaling to $[0,1]$, and apply it again to the reflected scores. Giving each tail failure probability $\delta/2$ yields $|\overline{\mathsf{T}} - A_\pi| \leq r_\delta$ with probability at least $1-\delta$. Intersecting the interval with $[0,1]$ preserves coverage. □