AUDITING BEST-OF-k SELECTION: MINIMAX LABEL COMPLEXITY OF QUALITY CURVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Best-of-k selection returns the response with the highest judge score among k candidates. As k grows, selection can favor responses whose quality the judge overestimates, causing actual quality to decline. We study how many trusted outcome labels, such as verified correctness or human ratings, are needed to estimate selected-response quality across candidate budgets. Given pools of L responses ordered by judge score, the quality curve averages the outcome of the highest-ranked response in a uniformly random k-subset, for every $k \leq L$. Under independent queries at chosen ranks with outcomes in $[0, 1]$, we establish matching upper and lower bounds of $\Theta(\varepsilon^{-2} \log L[\log \log L + \log(1/\delta)])$ on the minimax expected label cost for simultaneous error ε at confidence $1 - \delta$. The lower bound allows adaptive query allocation and data-dependent stopping. We prove that an estimator reusing labels across candidate budgets attains this rate under several closed-form rank allocations. These include Envelope, our default allocation, which minimizes the largest importance ratio and can be constructed in linear time. For fixed corpora, we derive simultaneous error bounds for sampling without replacement and recover the same minimax rate when sufficiently many pools are available. Experiments across diverse tasks show that Envelope achieves accuracy comparable to numerically optimized designs. Audited curves also guide the choice of candidate budget, and hence of inference-time computation, for new inputs.

1 INTRODUCTION

Best-of-k selection uses additional inference-time computation to improve a language model's response: it generates k candidates and returns the one with the highest judge score. Judge scores are imperfect measures of quality, so a low-quality response can receive a high score. Increasing k can amplify such overestimation, allowing the selected score to rise while actual quality falls. This decline in quality under stronger selection is a form of reward overoptimization (Gao et al., 2023; Huang et al., 2025).

Choosing k requires knowing how the quality of the selected response changes with the number of candidates. For a fixed task, generator, and judge, a *quality curve* records expected selected-response quality at each candidate budget. It reveals how much extra computation helps and whether increasing the budget eventually harms quality. For a specified generation cost, the curve guides the choice of candidate budget for new inputs.

Measuring this curve requires trusted outcome labels, such as verified correctness or human ratings, alongside the judge scores used for selection. These labels can be expensive, so an *audit* estimates the curve from a limited number of them. The challenge is to choose which responses to label so that estimates are reliable at every candidate budget.

The same labeled responses can estimate quality at several candidate budgets. Consider 32 responses to one input, ordered by judge score. For $k = 1$, every response is equally likely to be selected; for $k = 32$, only the highest-ranked response is returned. For intermediate k, each rank has a known selection probability. If each label is weighted by its importance ratio, the ratio of its rank's selection probability to its query probability, one set of labels suffices to estimate the entire curve. This leads to our central question: *How many trusted labels are needed to estimate the entire quality curve, and how should those labels be allocated?*

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.

1.1 CONTRIBUTIONS

Minimax label complexity. Under independent queries at chosen ranks, we establish matching upper and lower bounds

$$\Theta\big(\varepsilon^{-2} \log L \,[\log\log L + \log(1/\delta)]\big) \tag{1}$$

on the minimax label cost for simultaneous error ε at confidence $1-\delta$, where L is the candidate pool size. The lower bound allows adaptive query allocation and data-dependent stopping.

Explicit rate-optimal allocations. We construct several rank allocations that require no numerical optimization and prove that an estimator reusing labels across candidate budgets attains this rate under each. Our default, Envelope, allocates queries in proportion to each rank's maximum selection probability across candidate budgets. It minimizes the largest importance ratio and can be constructed in linear time. We compare finite-budget accuracy across these allocations and numerically optimized designs.

Finite-corpus guarantees. When an audit pays once per distinct label, we derive computable simultaneous error bounds for sampling without replacement. We establish the same minimax rate when sufficiently many candidate pools are available.

Empirical accuracy and budget selection. Experiments across diverse tasks show that Envelope achieves full-curve accuracy comparable to numerically optimized designs at matched label budgets. In budget-selection experiments, a candidate budget chosen from an offline audit achieves utility on new inputs close to that obtained using complete audit labels.

1.2 RELATED WORK

Quality curves and label acquisition. WebGPT reuses candidate pools to estimate selected-response quality from validation rewards (Nakano et al., 2021, Appendix I), and PPE evaluates best-of-k curves using correctness labels (Frick et al., 2025). These computations use evaluation outcomes for every candidate. We study how many trusted labels must be acquired to recover the entire curve. Lourie et al. (2024) provide simultaneous bands for tuning curves, which track the best observed validation score. Our curve measures the trusted quality of a response selected by a separate judge score, so quality can decline as k grows.

Shared sampling and minimax design. Each candidate budget defines a known distribution over selected ranks, so our audit is an instance of shared-sample estimation of multiple expectations (Roy & Evangelou, 2024; Chen et al., 2025). The minimax-variance objective of Chen et al. (2025) and Gaussian-width designs (Katz-Samuels et al., 2020) motivate our numerical comparators. Envelope specializes the normalized maximum likelihood construction (Shtar'kov, 1987) to the best-of-k family, minimizing the largest importance ratio. We derive its closed form and use the family's geometry to establish the minimax label cost for the entire curve under rank queries, including adaptive sampling and stopping.

Prediction-assisted evaluation. Active Testing and Active Evaluation Acquisition reduce evaluation cost by selecting informative inputs (Kossen et al., 2021; Li et al., 2025); selective human audits also identify the best alternative using noisy judges (Ao et al., 2026). Prediction-powered inference corrects aggregate estimates using labeled prediction errors (Angelopoulos et al., 2023a; Boyeau et al., 2025; Chen et al., 2026), with stratified and active extensions allocating labels to reduce variation in prediction errors (Fisch et al., 2024; Zrnic & Candès, 2024). These general estimation results leave open the minimax label cost of estimating the best-of-k curve simultaneously. Our bounds characterize this cost, and our estimator incorporates prediction-based corrections.

2 QUALITY CURVES AND AUDIT OBJECTIVES

2.1 POOLS, RANKS, AND SELECTED-RESPONSE QUALITY

A pool contains L candidate responses to one input. Rank r is a response's position when the full pool is ordered by judge score, from $r = 1$ (lowest) to $r = L$ (highest). Scores and the tie-breaking

order are set before auditing and remain unchanged as labels are revealed. A trusted outcome lies in $[0, 1]$; for correctness, it is one for a correct response and zero otherwise. Write μ_r for the mean quality at rank r over the pools of interest.

Choose a uniformly random subset of k responses and return its highest-ranked member. There are $\binom{r-1}{k-1}$ subsets whose maximum rank is r, giving

$$a_{kr} = \frac{\binom{r-1}{k-1}}{\binom{L}{k}} \mathbf{1}\{r \geq k\}, \qquad \theta_k = \sum_{r=1}^{L} a_{kr}\mu_r. \tag{2}$$

The *quality curve* is $(\theta_1, \ldots, \theta_L)$. The weights depend only on L, so the unknown rank means determine the task- and judge-specific curve. With binary outcomes and a judge that ranks all correct responses above incorrect ones, this subset average equals the pass@k estimator of Chen et al. (2021).

A wrong answer at the top rank is selected with probability $a_{kL} = k/L$, so increasing k can amplify judge errors. Figure 1 illustrates rising, peaked, and declining curves. We allow arbitrary rank means, including higher scores associated with worse quality.

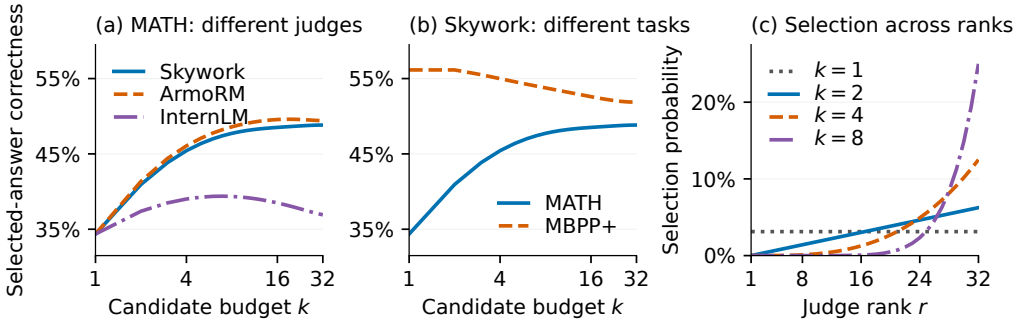

Figure 1: Quality and rank selection at $L = 32$. (a) Three judges rank the same MATH responses; InternLM's curve peaks at $k = 7$. (b) Skywork's curve rises on MATH and declines on MBPP+. Complete PPE labels determine these curves. (c) Selection probabilities a_{kr} for $k \in \{1, 2, 4, 8\}$ concentrate near the top rank as k grows. At $k = L$, all mass is at $r = L$.

2.2 Using a curve to choose the candidate budget

Let c_k be a specified cost of using k candidates, expressed on the quality scale. It may represent a preference for fewer candidates, monetary cost, or latency after an appropriate conversion. Given an estimated curve, choose $\widehat{k} \in \arg\max_k(\widehat{\theta}_k - c_k)$. On the event $\|\widehat{\theta} - \theta\|_\infty \leq \varepsilon$,

$$\max_k(\theta_k - c_k) - (\theta_{\widehat{k}} - c_{\widehat{k}}) \leq 2\varepsilon. \tag{3}$$

This follows by bounding the errors at the best true budget and at $\widehat{k}$. The bound holds for every cost vector and remains valid when the budget is chosen after inspecting the audited curve.

An *audit set* contains independent pools from a fixed input distribution, generator, decoding configuration, judge, and outcome definition. If responses are conditionally iid given the input and the judge scores each separately, a random k-subset has the same selection law as k fresh responses. Let $Q(k)$ be expected quality on a new input. If the audit error is at most e_a and the audit-set curve differs uniformly from Q by at most e_g, population utility regret is at most $2(e_a + e_g)$. Appendix I.3 bounds both errors, allowing dependence between ranks within a pool. Section 5.2 evaluates the resulting budget choices on separate test inputs.

2.3 Query access and minimax label cost

A query at rank r returns a trusted label $Y \sim P_r$ on $[0, 1]$, with mean μ_r, independently of past observations conditional on the chosen rank. The audit may use observed labels to choose subsequent

ranks and when to stop. We call this access model a *rank oracle*. The *label budget* T counts queries to this oracle; the *candidate budget* k counts responses available for selection.

A fixed corpus implements the oracle by drawing a pool uniformly with replacement after each rank choice. Conditional on the corpus, these draws are independent even if response qualities within a pool are related. A fresh independent pool for each query gives the population version. In Section 4, an audit pays once per distinct label and may sample without replacement.

Definition 1 (Minimax label cost)**.** A procedure is (ε, δ)-accurate if it stops almost surely at a finite time τ and returns $\widehat{\theta}$ with $\mathbb{P}_P(\left\|\widehat{\theta} - \theta\right\|_\infty \le \varepsilon) \ge 1 - \delta$ for every tuple $P = (P_1, \dots, P_L)$ of laws on $[0, 1]$. The minimax label cost $m^*(L, \varepsilon, \delta)$ is the infimum of $\sup_P \mathbb{E}_P \tau$ over these procedures.

3 Rank allocation and minimax guarantees

3.1 Envelope allocation and shared estimation

Envelope allocates more queries to ranks that have a high chance of being selected at some candidate budget. For each rank, take its largest selection probability over k and normalize:

$$b_r = \max_k a_{kr}, \qquad B_L = \sum_r b_r, \qquad q_r^{\mathrm{env}} = b_r / B_L. \tag{4}$$

Draw ranks independently from q^{env}. Each label contributes to every candidate budget, weighted by the ratio of its selection probability to its query probability. For a fixed prediction $g_r \in [0, 1]$, the estimator is

$$\widehat{\theta}_k = G_k + \frac{1}{T} \sum_{t=1}^{T} \frac{a_{kR_t}}{q_{R_t}} (Y_t - g_{R_t}), \qquad G_k = \sum_r a_{kr} g_r. \tag{5}$$

It is unbiased for every fixed full-support q. The default $g_r = 1/2$ needs no training labels; Envelope depends only on L. Predictions can reduce residual variation through this standard prediction-plus-correction form (Angelopoulos et al., 2023a;b); Appendix H.1 compares raw scores and fitted predictions.

Algorithm 1 computes the allocation with L coefficient evaluations and linear memory, using log-binomial values without constructing the $L \times L$ matrix (Appendix A.1). The same allocation serves every label budget. Appendix D.2 reports construction times and comparisons with numerical designs.

3.2 Minimax label complexity

Theorem 2 (Minimax label complexity)**.** *There are universal constants $c, C > 0$ such that, for every integer $L \ge 2$ and $0 < \varepsilon, \delta \le 1/8$,*

$$c \frac{\log L}{\varepsilon^2} \left[\log\log L + \log(1/\delta)\right] \le m^*(L, \varepsilon, \delta) \le C \frac{\log L}{\varepsilon^2} \left[\log\log L + \log(1/\delta)\right]. \tag{6}$$

The upper bound uses estimator (5), iid ranks from q^{env}, and a deterministic number of queries T. The lower bound permits arbitrary adaptive queries, stopping times, and estimators.

Algorithm 1 Envelope with independent queries

Require: Pool size L, label budget T, a rank oracle, fixed predictions g (default $1/2$).
1: For each rank r, set $k_r^* = \min\{L, \lfloor (L+1)/(L-r+1) \rfloor\}$ and $b_r = a_{k_r^*, r}$.
2: Set $q_r^{\mathrm{env}} = b_r / \sum_s b_s$ and $G_k = \sum_r a_{kr} g_r$.
3: Draw $R_1, \dots, R_T$ independently from q^{env}; query the label Y_t at each R_t.
4: **return** the shared estimates $\widehat{\theta}_k$ in (5) for every k.

A single candidate budget chosen before the audit requires order $\varepsilon^{-2} \log(1/\delta)$ labels: sample ranks from its selection law. Theorem 2 characterizes, up to universal constants, the extra label cost of measuring every candidate budget.

3.3 UNIFORM UPPER BOUND

A random k-subset misses the top s ranks with probability at most $e^{-ks/L}$. Candidate budgets near $k = L/s$ therefore put substantial probability in that neighborhood. Geometric candidate budgets probe $O(\log L)$ rank scales, giving an envelope mass B_L of order $\log L$. Controlling these scales simultaneously contributes the $\log \log L$ term.

Cover the relevant ranks. The largest importance ratio $D(q) = \max_{k,r} a_{kr}/q_r$ bounds both the magnitude and the second moment of the weighted residuals, and hence the influence of a queried label. Envelope uniquely minimizes it:

$$\inf_q \max_{k,r} \frac{a_{kr}}{q_r} = B_L, \qquad c_B \log(eL) \le B_L \le H_L \le 1 + \log L,$$

where the infimum is over full-support distributions, $H_L = \sum_{s=1}^{L} 1/s$, and $c_B > 0$ is universal. If $b_r/q_r \le C$ at every rank, summing $b_r \le Cq_r$ gives $B_L \le C$, with equality only at Envelope. A rank with s responses at or above it has selection probability at most $1/s$, giving the harmonic upper bound. Disjoint rank intervals, each carrying substantial probability at some budget, give the logarithmic lower bound. At $L = 32$, Envelope reduces this ratio from Uniform's 32 to 2.631.

Use the relation between candidate budgets. A Bernstein bound and union bound over L separate targets give cost $O(B_L \log(L/\delta)/\varepsilon^2)$. Nearby budgets have similar selection probabilities, so a tighter bound is possible. Each rank's selection probability first rises and then falls with k: small subsets rarely include it, while large subsets usually include a higher rank.

The normalized weights a_{kr}/b_r therefore have total variation at most two in k. Grouping budgets by accumulated variation gives a small set of representative functions. The covering number $N(\eta, \mathcal{F}, L_2(Q))$ counts the representatives needed for root-mean-square error η under Q.

Lemma 3 (Covering the budget family)**.** *For $f_k(r, y) = a_{kr}(y - g_r)/b_r$, let $\mathcal{F} = \{f_k : k \le L\}$. For every probability measure Q on observations and $0 < \eta \le 1$,*

$$N(\eta, \mathcal{F}, L_2(Q)) \le 1 + 2/\eta^2, \qquad \sup_k \mathbb{E} f_k^2 \le 1/B_L, \tag{7}$$

where the second expectation uses the envelope oracle law.

Each normalized target has second moment at most $1/B_L$. At this scale, the covering bound needs only $O(B_L)$ representatives, whose logarithm is order $\log B_L$. The local maximal inequality of van der Vaart & Wellner (2011) controls the approximation at all finer scales; concentration (Bousquet, 2002) then gives the following bound.

Theorem 4 (Envelope upper bound)**.** *For $L \ge 2$, $T \ge 1$, $0 < \delta < 1$, and every fixed predictor $g \in [0, 1]^L$, Envelope with estimator (5) satisfies*

$$\mathbb{P}\left\{ \left\| \widehat{\theta} - \theta \right\|_\infty > C \left[\sqrt{\frac{B_L \log(eB_L/\delta)}{T}} + \frac{B_L \log(eB_L/\delta)}{T} \right] \right\} \le \delta, \tag{8}$$

where C is universal.

More generally, the bound holds with $D(q)$ replacing B_L (Proposition C.1). Harmonic weights $q_r \propto 1/(L - r + 1)$ and a uniform mixture of selection laws over $\mathcal{K} = \{1, 2, 4, \ldots\} \cap [1, L] \cup \{L\}$ also attain the rate: their ratios are H_L and $|\mathcal{K}|$. We use Envelope as the default because it has a closed form and minimizes this ratio. Actual error also depends on residual variances and covariance across budgets; Appendix C.3 compares the resulting finite-budget accuracy.

A union bound gives this rate on $\mathcal{K}$. The covering argument extends it to every integer budget without increasing the label-complexity order (Appendix C.4). Full proofs appear in Appendix A.

3.4 ADAPTIVE LOWER BOUND

Theorem 5 (Adaptive lower bound)**.** *For $L \ge 2$ and $0 < \varepsilon, \delta \le 1/8$, every (ε, δ)-accurate procedure satisfies, even for Bernoulli outcome laws,*

$$\sup_P \mathbb{E}_P \tau \ge \frac{J \log(J/\delta)}{160\varepsilon^2}, \qquad J = 1 + \lfloor \log_8 L \rfloor. \tag{9}$$

The proof uses hard instances with independent quality changes of either sign at the top rank and in intervals at distances $[1, 8)$, $[8, 64)$, and so on below it. Mixtures of candidate budgets reveal these signs from the top rank downward. Once preceding signs are recovered, their known contributions can be subtracted; each remaining interval then has enough mass to identify its sign despite all farther intervals and an estimation error of ε. Thus uniform curve accuracy recovers all J signs. The top rank alone supplies one sign when $L < 8$.

Under an independent uniform prior, the signs' posterior factorizes even after adaptive stopping. Joint recovery with failure probability δ requires order $\varepsilon^{-2} J \log(J/\delta)$ queries (Appendix A.3).

4 Auditing a finite corpus

A fixed corpus has M ranked pools and ML outcomes $y_{ir} \in [0, 1]$. An audit pays once per revealed record and targets the corpus curve $\theta_k = M^{-1} \sum_{i,r} a_{kr} y_{ir}$, so all randomness comes from sampling records. We first analyze fixed quotas, then establish the finite-corpus minimax cost.

4.1 Fixed quotas for distinct labels

For a label budget $L \le n \le ML$, choose integer quotas $1 \le n_r \le M$ summing to n; a *census* labels all ML records. Independently sample a uniform subset within each rank and estimate

$$\widehat{\theta}_k = \sum_r a_{kr} \overline{Y}_r. \tag{10}$$

This estimator is unbiased. Sampling without replacement reduces its covariance through a finite-population correction (Proposition B.1).

Algorithm 2 converts Envelope into proportional quotas, with at least one label per rank and at most M. Appendix B.3 also analyzes additive quotas, which reserve one label per rank before allocating the remainder.

Algorithm 2 Envelope with a distinct-label budget

Require: M records per rank, $L \le n \le ML$.
1: Compute q^{env}. If $n = L$, set $x_r = 1$; if $n = ML$, set $x_r = M$.
2: Otherwise choose c with $\sum_r x_r = n$, where $x_r = \min\{M, \max\{1, cq_r^{\mathrm{env}}\}\}$.
3: Set $n_r = \lfloor x_r \rfloor$ and assign the remaining labels by decreasing fractional part.
4: **for** each rank r independently **do**
5: Sample n_r distinct pools uniformly; reveal their rank-r labels and compute $\overline{Y}_r$.
6: **end for**
7: **return** $\widehat{\theta}_k = \sum_r a_{kr} \overline{Y}_r$ for every k.

Theorem 6 (Fixed-quota curve error)**.** *For deterministic quotas $1 \le n_r \le M$ and independent uniform subsets within ranks, define*

$$h_r = \min\left\{ \frac{M - n_r + 1}{M n_r}, \frac{(M - n_r)(n_r + 1)}{M n_r^2} \right\}, \quad V_k = \sum_r a_{kr}^2 h_r, \quad \kappa_n = \max_r b_r h_r.$$

For $0 < \delta < 1$, if $V_k > 0$ for some k, let $u_\delta > 0$ be the unique solution of

$$\sum_{k: V_k > 0} 2 \exp(-2u_\delta^2 / V_k) = \delta, \qquad \text{set } u_\delta = 0 \text{ if all } V_k = 0.$$

For every fixed outcome array,

$$\mathbb{P}\left\{ \left\| \widehat{\theta} - \theta \right\|_\infty > R_n(\delta) \right\} \le \delta, \quad R_n(\delta) = \min\left\{ 1, u_\delta, \sqrt{2\kappa_n \log(384 B_L/\delta)} \right\}. \tag{11}$$

For $n > L$, Algorithm 2 satisfies $\kappa_n \le 2B_L/(n - L)$. At a census the error is zero.

The radius is computable before querying labels, so a budget n with $R_n(\delta) \le \varepsilon$ can be chosen in advance. The scalar equation gives the tighter certificate in our experiments; the covering term establishes the dependence on L. The factor h_r applies finite-population corrections to the sample and its complement (Bardenet & Maillard, 2015). Appendix B.3 proves the bound and describes this choice.

4.2 FINITE-CORPUS MINIMAX COMPLEXITY

Let m^*_{fin} be the infimum, over accurate procedures that may choose records and stop adaptively, of the worst-case expected number of distinct labels.

Corollary 7 (Finite-corpus minimax complexity)**.** *For* $L \geq 2$, $0 < \varepsilon, \delta \leq 1/16$, *and* $M \geq \log(2/\delta)/(2\varepsilon^2)$,

$$m^*_{\text{fin}}(M, L, \varepsilon, \delta) = \Theta\left(\frac{\log L}{\varepsilon^2}[\log\log L + \log(1/\delta)]\right).$$

Cached Envelope queries attain the upper bound; the lower bound allows all adaptive finite-corpus audits.

For the lower bound, we generate a corpus with independent records and transfer a finite-corpus audit to the oracle problem. Concentration of the corpus curve introduces no additional L dependence. Caching Envelope queries gives the upper bound (Appendix B.2). This scheme pays only for new labels and can use label budgets below L without querying every rank. Fixed quotas also attain this order when L is at most a constant multiple of the statistical cost.

5 EXPERIMENTS

We evaluate the finite-budget accuracy of rate-optimal allocations, measured by mean maximum error $\mathbb{E}\max_k |\widehat{\theta}_k - \theta_k|$, and the resulting candidate-budget choices for new inputs. Complete labels supply reference curves; each audit observes only the labels it pays for.

Data and comparisons. The 34 fixed-corpus settings comprise 15 PPE task–judge pairs ($L = 32$), four SummEval rating dimensions ($L = 16$), four WMT22 translation rankings ($L = 14$), and 11 mathematical settings from SAL and RLHFlow ($L \leq 2048$) (Frick et al., 2025; Fabbri et al., 2021; Freitag et al., 2022; Beeching et al., 2024; Xiong et al., 2024). SummEval and WMT pools are fixed menus of system outputs, ranked by G-Eval and GEMBA GPT-4 (Liu et al., 2023; Kocmi & Federmann, 2023); the budget-selection study uses the SAL and RLHFlow generation pools. Appendices D and I.1 specify the data and rankings.

Rank allocations in Table 1 share an estimator within each protocol. Algorithm 2 converts rank allocations to distinct-label quotas; MMV optimizes quotas using finite-corpus variances (Appendix D.2, Equation (D.1)). Means and Monte Carlo SEs use 1,000 paired audits per corpus. Appendix E gives additional comparisons.

Table 1: Audit designs. MMV (minimax-variance) and GW (Gaussian-width) use numerical optimization with fixed unit proxies; the other rank allocations have explicit formulas.

Method	Allocation rule or design objective
Envelope	$q_r \propto \max_k a_{kr}$; minimizes $D(q)$ (Section 3.3) (Shtar'kov, 1987).
Harmonic	$q_r \propto 1/(L - r + 1)$; construction and analysis in Appendix C.
Grid-mixture	$q_r = \lvert\mathcal{K}\rvert^{-1} \sum_{k\in\mathcal{K}} a_{kr}$; geometric candidate budgets (Appendix C).
Uniform	$q_r = 1/L$; equal probability for every rank.
Target-mixture	$q_r = L^{-1} \sum_{k=1}^{L} a_{kr}$; equal mixture over all k (Veach & Guibas, 1995).
MMV	Minimize $\max_k \sum_r a_{kr}^2/q_r$ for independent queries (Chen et al., 2025).
GW	Simulated maximum-error proxy, inspired by Katz-Samuels et al. (2020).
Whole-pool	Sample inputs uniformly and label all L responses per input.

5.1 FULL-CURVE ACCURACY ACROSS TASKS AND CANDIDATE POOLS

The explicit rate-optimal allocations improve on Uniform in all six summaries (Table 2). Envelope has error comparable to MMV and GW; Harmonic and Grid-mixture exchange order across collections. Figure 2 illustrates how one audit reconstructs rising, peaked, and declining curves.

Table 2: Distinct-label audits across tasks: mean maximum error ($\times 10^3$), with Monte Carlo SEs. The first four columns use $n = 512$; the last two use $n = 2L$. SAL uses Llama 1B. Collection summaries average the settings within each column.

Method	PPE	SummEval	WMT	SAL 256	SAL 1024	SAL 2048
Whole-pool	114.8 (2.0)	14.4 (0.2)	20.3 (0.4)	346.8 (5.5)	348.0 (5.4)	350.6 (5.2)
Uniform	105.7 (2.1)	12.6 (0.2)	17.9 (0.4)	303.0 (6.2)	304.8 (6.3)	299.4 (6.1)
Target-mixture	61.2 (1.0)	14.2 (0.3)	15.4 (0.3)	60.4 (0.8)	38.0 (0.6)	27.7 (0.4)
Grid-mixture	42.8 (0.6)	10.2 (0.2)	11.8 (0.2)	56.3 (0.7)	29.1 (0.3)	19.6 (0.2)
Harmonic	44.1 (0.6)	9.4 (0.2)	11.3 (0.2)	60.0 (0.8)	31.3 (0.4)	20.9 (0.2)
MMV	41.7 (0.5)	7.8 (0.1)	10.9 (0.1)	55.3 (0.7)	28.9 (0.3)	21.1 (0.2)
GW	41.4 (0.5)	8.1 (0.2)	10.8 (0.1)	55.3 (0.7)	28.9 (0.4)	19.7 (0.2)
Envelope	41.3 (0.5)	8.0 (0.2)	10.8 (0.1)	55.1 (0.7)	28.4 (0.4)	19.5 (0.2)

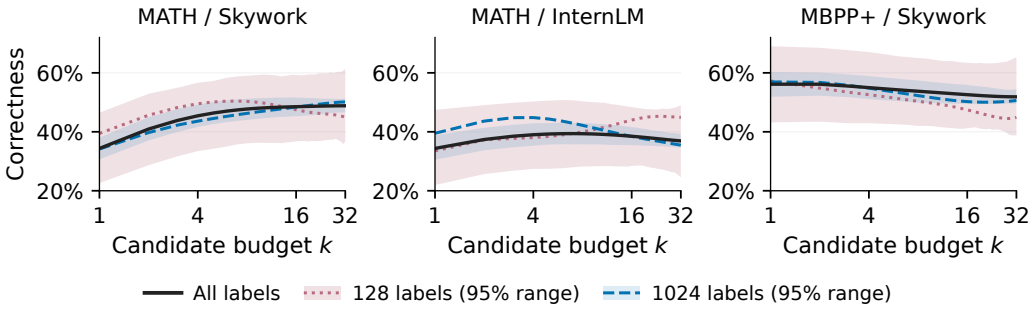

Figure 2: Envelope curve reconstruction with 128 or 1,024 distinct labels, sampled without replacement within ranks. Black curves use all labels; colored curves show the first audit replication. Shading gives the corresponding pointwise 2.5th–97.5th percentiles over 1,000 audits. Appendix F shows all 15 PPE reference curves.

We test whether this comparison persists across generators and judges on the nine configurations in Table 3, each a combination of task, generator, and judge. The four judge checkpoints share a Llama3.1-8B base: outcome reward models (ORMs) score answers, and process reward models (PRMs) score reasoning steps. Figure 8 in Appendix I.1 shows the quality curves of these configurations.

Table 3: Large-pool audits at $n = 2L$: mean maximum error ($\times 10^3$), with Monte Carlo SEs. Llama uses $L = 2048$; DeepSeek and Mistral use $L = 1024$. Judge subscripts D and M identify DeepSeek and Mistral training data.

ID	Task	Generator	Judge	Target-mixture	Harmonic	MMV	GW	Envelope
1	MATH500	Llama 1B	PRM_D	27.8 (0.4)	20.5 (0.2)	20.6 (0.2)	19.8 (0.2)	19.7 (0.2)
2	MATH500	Llama 3B	PRM_D	28.3 (0.4)	21.1 (0.2)	21.2 (0.2)	20.6 (0.2)	20.5 (0.2)
3	MATH500	DeepSeek 7B	ORM_D	37.9 (0.6)	32.0 (0.4)	29.0 (0.3)	29.1 (0.4)	28.8 (0.3)
4	MATH500	DeepSeek 7B	ORM_M	38.0 (0.6)	32.1 (0.4)	29.5 (0.3)	28.6 (0.3)	28.7 (0.3)
5	MATH500	Mistral 7B	ORM_M	37.2 (0.6)	31.5 (0.4)	28.6 (0.3)	28.4 (0.4)	28.3 (0.3)
6	MATH500	Mistral 7B	PRM_M	36.0 (0.5)	31.9 (0.4)	29.1 (0.3)	28.3 (0.4)	28.2 (0.3)
7	GSM8K	DeepSeek 7B	ORM_D	21.8 (0.3)	18.7 (0.2)	18.1 (0.2)	18.1 (0.2)	18.1 (0.2)
8	GSM8K	DeepSeek 7B	ORM_M	23.7 (0.3)	21.2 (0.3)	19.6 (0.2)	19.7 (0.2)	19.6 (0.2)
9	GSM8K	Mistral 7B	ORM_M	25.5 (0.3)	22.0 (0.3)	20.7 (0.2)	20.6 (0.2)	20.6 (0.2)

Across all nine configurations, Envelope has similar error to MMV and GW; Harmonic is comparable on Llama and modestly worse elsewhere. Target-mixture has higher error than Envelope and Harmonic throughout with the same estimator. Independent queries show a similar pattern (Table 19).

Appendices C, G, and H examine allocation geometry, scaling, and the value and label cost of score prediction.

5.2 Budget selection for new inputs

For each configuration in Table 3, we split inputs into audit and test sets. The estimated audit-set curve $\widehat{\theta}_{\mathrm{A}}$ selects one candidate budget for subsequent inputs. Its test utility regret is

$$\widehat{k} = \min \arg\max_{1 \le k \le L} \{\widehat{\theta}_{\mathrm{A}}(k) - c_k\}, \qquad c_k = \rho \frac{k-1}{L-1}, \\ R_{\mathrm{test}}(\widehat{k}) = \max_{1 \le k \le L} \{\theta_{\mathrm{test}}(k) - c_k\} - \{\theta_{\mathrm{test}}(\widehat{k}) - c_{\widehat{k}}\}. \tag{12}$$

We compute θ_{test} from complete test labels. The primary price $\rho = .1$ penalizes using L candidates by ten percentage points of accuracy. Appendix I.4 compares five prices for this cost.

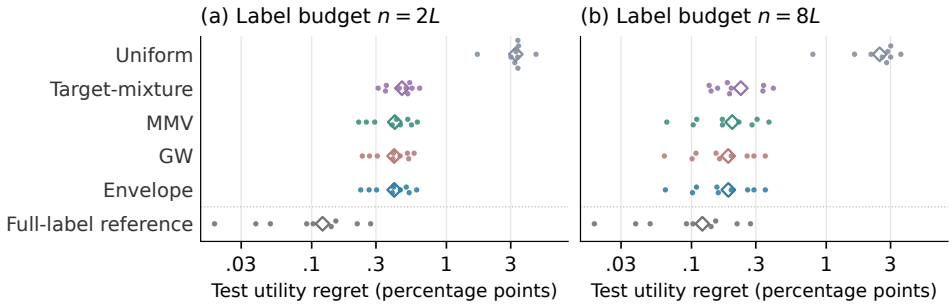

Figure 3: Test utility regret at $\rho = .1$ for the nine configurations in Table 3 (dots), with means (diamonds). Vertical offsets only prevent overlap; axes are logarithmic. The final row chooses k using all audit labels. Table 21 gives individual results and Monte Carlo SEs.

Figure 3 shows similar test utility for Envelope, MMV, and GW. Larger audits approach the full-label reference, whose regret reflects audit–test variation. Harmonic and Grid-mixture have slightly lower mean regret than Envelope at $8L$ in each of the nine configurations (Appendix I.4). Regret evaluates only the chosen budget, so its ranking can differ from full-curve error.

At $n = 8L$, audits use 3.2% of MATH500 and 1.21% of GSM8K audit labels. Envelope's mean extra test utility loss over the full-label choice is at most 0.122 percentage points across the nine configurations. With the same PRM, Llama 1B and 3B select mean budgets of 138 and 29 (Table 22), illustrating generator-specific budget choices.

6 Discussion and conclusion

Under rank queries, the label cost of auditing the entire best-of-k quality curve exceeds that of auditing one candidate budget fixed in advance by only a polylogarithmic factor in L. Closed-form rank allocations attain this minimax rate, which we also recover for finite corpora with sufficiently many pools. Empirically, Envelope achieves full-curve accuracy comparable to numerically optimized designs, and budgets selected through audits achieve utility on held-out inputs close to that obtained using complete audit labels. Simultaneous accuracy bounds the utility loss of the selected candidate budget, so audits provide a statistical basis for decisions about inference-time computation.

Limitations and future work. The audit measures quality as defined by the trusted outcome, and the label-cost analysis excludes the cost of generating and scoring candidates. Guarantees for new inputs assume that audit pools are drawn independently from the target distribution and that the consistency conditions in Appendix I.3 hold. When reading a new input incurs a fixed cost, grouping labels by input can improve accuracy at a fixed total cost (Appendix E.4), motivating joint allocation across inputs and ranks. Bounds that adapt to outcome variance could further sharpen the guarantees.

REPRODUCIBILITY STATEMENT

Appendices A–C and Appendix I.3 give complete proofs and assumptions. Appendices D and I.1 describe data sources, outcome definitions, rankings, and numerical designs; Appendix I.4 specifies the input splits and budget-selection protocol. Code, fixed seeds, data-acquisition instructions, figure construction, and numerical results are available at https://anonymous.4open.science/r/best-of-k-auditing-E814/.

AI USE STATEMENT

We set the direction of this research and decided each claim in the paper and how it is presented. Generative AI tools assisted us in several parts of the work. To aid and polish writing, we used them to revise the wording of our drafts for clarity, grammar, and consistent terminology across sections. For retrieval and discovery, we used them to search for and summarize related work. For research ideation and execution, they helped us refine the problem formulation and research questions and gave feedback on the methodology and experiments. They also helped us implement and debug the experiment code, clean and reformat the public datasets, and interpret the results. For proving mathematical claims, they suggested proof strategies for parts of the analysis, checked our derivations for errors and gaps, and proposed corrections, which we verified before incorporating them.

The LLM judges and reward models in our experiments are objects of study. Their scores come from public releases, and the correctness labels come from those releases or from rule-based answer checkers. The only synthetic outcomes, in Appendix G, are Bernoulli draws from the probability model stated there, generated by our simulation code.

We have reviewed all AI-assisted work. We also tested the AI-assisted code, and all reported results can be reproduced with fixed seeds from our anonymous code repository. We take responsibility for the final content of this work.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000. URL https://doi.org/10.1126/science.adi6000.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. arXiv:2311.01453, 2023b. URL https://arxiv.org/abs/2311.01453.

Ruicheng Ao, Hongyu Chen, Siyang Gao, Hanwei Li, and David Simchi-Levi. Best arm identification with LLM judges and limited human audits. arXiv:2601.21471, 2026. URL https://arxiv.org/abs/2601.21471.

Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015. doi: 10.3150/14-BEJ605. URL https://doi.org/10.3150/14-BEJ605.

Edward Beeching, Lewis Tunstall, and Sasha Rush. Scaling test-time compute with open models. Hugging Face blog, December 16, 2024. URL https://huggingface.co/spaces/HuggingFaceH4/blogpost-scaling-test-time-compute. Companion Search and Learn repository: https://github.com/huggingface/search-and-learn.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus. Mathématique*, 334(6):495–500, 2002. doi: 10.1016/S1631-073X(02)02292-6. URL https://comptes-rendus.academie-sciences.fr/mathematique/item/CRMATH_2002__334_6_495_0/.

Pierre Boyeau, Anastasios Nikolas Angelopoulos, Tianle Li, Nir Yosef, Jitendra Malik, and Michael I. Jordan. AutoEval done right: Using synthetic data for model evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5276–5290. PMLR, 2025. URL https://proceedings.mlr.press/v267/boyeau25a.html.

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

Yilei Chen, Aldo Pacchiano, and Ioannis Ch. Paschalidis. Multiple-policy evaluation via density estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8794–8820. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25aw.html.

Yiqun T. Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. Efficient inference for noisy LLM-as-a-judge evaluation. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61131. arXiv:2601.05420.

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:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Víctor Elvira, Luca Martino, David Luengo, and Mónica F. Bugallo. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019. doi: 10.1214/18-STS668. URL https://doi.org/10.1214/18-STS668.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL https://aclanthology.org/2021.tacl-1.24/.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 111489–111514. Curran Associates, Inc., 2024. doi: 10.52202/079017-3541. URL https://papers.nips.cc/paper_files/paper/2024/hash/c9fcd02e6445c7dfbad6986abee53d0d-Abstract-Conference.html.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 46–68. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.wmt-1.2. URL https://aclanthology.org/2022.wmt-1.2/.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. González, and Ion Stoica. How to evaluate reward models for RLHF. In *International Conference on Learning Representations*, volume 2025, pp. 18128–18163, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2e01083b381b4865919b4915ef32e3d2-Abstract-Conference.html.

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

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

Julian Katz-Samuels, Lalit Jain, Zohar Karnin, and Kevin Jamieson. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10371–10382. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/75800f73fa80f935216b8cfbedf77bfa-Abstract.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://www.jmlr.org/papers/v17/kaufman16a.html.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203. European Association for Machine Translation, 2023. URL https://aclanthology.org/2023.eamt-1.19/.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Yang Li, Jie Ma, Miguel Ballesteros, Yassine Benajiba, and Graham Horwood. Active evaluation acquisition for efficient LLM benchmarking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35581–35602. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25bp.html.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Nicholas Lourie, Kyunghyun Cho, and He He. Show your work with confidence: Confidence bands for tuning curves. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3455–3472. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.189. URL https://aclanthology.org/2024.naacl-long.189/.

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

Vivekananda Roy and Evangelos Evangelou. Selection of proposal distributions for multiple importance sampling. *Statistica Sinica*, 34(1):27–46, 2024. doi: 10.5705/ss.202021.0049. URL https://www3.stat.sinica.edu.tw/statistica/j34n1/j34n102/j34n102.html.

Yu. M. Shtar'kov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):175–186, 1987. URL https://www.mathnet.ru/eng/ppi811.

Aad van der Vaart and Jon A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5:192–203, 2011. doi: 10.1214/11-EJS605. URL https://doi.org/10.1214/11-EJS605.

Eric Veach and Leonidas J. Guibas. Optimally combining sampling techniques for Monte Carlo rendering. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 419–428. Association for Computing Machinery, 1995. doi: 10.1145/218380.218498. URL https://doi.org/10.1145/218380.218498.

Wei Xiong, Hanning Zhang, Nan Jiang, and Tong Zhang. An implementation of generative PRM. GitHub repository, 2024. URL https://github.com/RLHFlow/RLHF-Reward-Modeling.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010. PMLR, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.

The appendices provide oracle proofs (Appendix A), finite-corpus proofs (Appendix B), allocation shape and candidate-budget resolution (Appendix C), experimental protocols and results (Appendices D–E), curve illustrations (Appendix F), scaling experiments on simulated and real pools (Appendix G), prediction-assisted estimation (Appendix H), and large-pool audits and budget selection on new inputs (Appendix I).

A Oracle analysis

A.1 Auxiliary oracle results

For $L = 3$, Equation (4) gives $b = (1/3, 1/3, 1)$ and $q^{\mathrm{env}} = (1/5, 1/5, 3/5)$. Envelope puts most queries at the top rank and retains lower-rank labels for smaller budgets. A top-rank label receives weight $5/3$ in the estimate of θ_3, which depends only on the top rank, and weight $5/9$ in the estimate of θ_1, which averages all ranks.

Proposition A.1 (Unbiasedness and shared covariance)**.** *For a fixed full-support allocation q, let $v_r = \mathbb{E}[(Y - g_r)^2 \mid R = r]$ and $h_k = \theta_k - G_k$. Then*

$$\mathbb{E}\widehat{\theta}_k = \theta_k, \qquad \mathrm{Cov}(\widehat{\theta}_k, \widehat{\theta}_\ell) = \frac{1}{T}\left(\sum_r \frac{a_{kr} a_{\ell r} v_r}{q_r} - h_k h_\ell\right). \tag{A.1}$$

For a corpus predictor, replace v_r by its average residual second moment over inputs. If the predictor is fitted using an independent pilot, the identities hold conditionally on the pilot for the remaining estimation labels.

Lemma A.2 (Envelope bounds)**.** *The envelope design uniquely minimizes the maximum importance ratio:*

$$\inf_{q:\, q_r > 0,\, \sum q_r = 1} \max_{k,r} \frac{a_{kr}}{q_r} = B_L, \qquad 1 \le B_L \le H_L \le 1 + \log L, \tag{A.2}$$

where $H_L = \sum_{s=1}^{L} 1/s$. For $q = q^{\mathrm{env}}$, the weighted residual $Z_{tk} = (a_{kR_t}/q_{R_t})(Y_t - g_{R_t})$ satisfies $|Z_{tk}| \le B_L$ and $\mathbb{E}Z_{tk}^2 \le B_L$ for all k.

Proof. We prove the three claims in order.

1. Optimal allocation. For any full-support q, set $C(q) = \max_r b_r/q_r$. Then

$$b_r \le C(q) q_r \quad \text{for every } r, \qquad B_L = \sum_r b_r \le C(q) \sum_r q_r = C(q).$$

The choice $q_r = b_r/B_L$ attains equality. If another q attains it, every inequality $b_r \le B_L q_r$ must be an equality, proving uniqueness.

2. Envelope mass. For fixed k, a_{kr} is nondecreasing in r. Consequently,

$$(L - r + 1) a_{kr} \le \sum_{s=r}^{L} a_{ks} \le 1, \qquad b_r \le \frac{1}{L - r + 1}.$$

Summing over ranks gives $B_L \le H_L \le 1 + \log L$. Every row of a sums to one and is bounded by b, so $B_L \ge 1$.

3. Weighted residuals. Since $Y, g_r \in [0, 1]$ and $a_{kr}/q_r^{\mathrm{env}} \le B_L$,

$$|Z_k| \le B_L, \qquad \mathbb{E}Z_k^2 \le \sum_r \frac{a_{kr}^2}{q_r^{\mathrm{env}}} \le B_L \sum_r a_{kr} = B_L.$$

These bounds hold for every budget k and every tuple of outcome laws on $[0, 1]$. □

Computing the envelope without the full matrix. The adjacent-ratio identity (A.3), derived below, locates each column maximum in closed form. Put $s = L - r + 1$. One maximizing budget is

$$k_r^* = \min\{L, \lfloor (L+1)/s \rfloor\}, \qquad b_r = \frac{\binom{r-1}{k_r^*-1}}{\binom{L}{k_r^*}}.$$

Indeed, the adjacent ratio is at least one exactly when $k \le r/s = (L+1)/s - 1$. If this threshold is an integer, the two neighboring coefficients tie; the displayed choice takes the larger index. The endpoint $r = L$ has its maximum at $k = L$. Thus constructing the allocation requires L coefficient evaluations and one normalization. Computing coefficients from log-binomial values avoids the $L \times L$ matrix and takes time linear in L if each special-function evaluation takes constant time.

A.1.1 Proof of Lemma 3

Proof. The coefficient shape controls distances between the normalized functions.

1. Unimodality of the coefficients. For $k < r$, direct cancellation of adjacent binomial coefficients gives

$$\frac{a_{k+1,r}}{a_{kr}} = \frac{(k+1)(r-k)}{k(L-k)} \ge 1 \quad \Longleftrightarrow \quad k \le \frac{r}{L-r+1}. \tag{A.3}$$

Thus $a_{1r}, \dots, a_{Lr}$ first increases and then decreases, including its zero tail. At each observation (r, y), all f_k have the same sign, $|f_k| \le 1$, and

$$\sum_{k=1}^{L-1} |f_{k+1}(r,y) - f_k(r,y)| \le 2.$$

2. A distance controlled by total variation. Fix any probability measure Q on observations, and define

$$v_j = \mathbb{E}_Q \sum_{k<j} |f_{k+1} - f_k|, \qquad 0 \le v_j \le 2.$$

For $j < \ell$, the same-sign property gives $|f_\ell - f_j| \le 1$. Therefore

$$\begin{aligned} \|f_\ell - f_j\|_{L_2(Q)}^2 &\le \mathbb{E}_Q |f_\ell - f_j| \\ &\le \mathbb{E}_Q \sum_{k=j}^{\ell-1} |f_{k+1} - f_k| = v_\ell - v_j. \end{aligned}$$

3. Covering and second moments. Partition $[0, 2]$ into intervals of length η^2, with a possible endpoint interval. Choose one f_j for each occupied interval of the numbers v_j. The preceding bound places every function within $L_2(Q)$ distance η of a representative. The number of representatives is at most $1 + 2/\eta^2$.

When ranks are drawn from q^{env}, $\rho_r = \mathbb{E}[(Y - g_r)^2 \mid R = r] \le 1$, and

$$\mathbb{E} f_k^2 = \frac{1}{B_L} \sum_r \frac{a_{kr}^2 \rho_r}{b_r} \le \frac{1}{B_L} \sum_r a_{kr} = \frac{1}{B_L}.$$

For a shared corpus predictor, include the pool index in the observation and average the bounded residual second moment over the pools within each rank. Every preceding step remains unchanged. □

A.2 Details of the uniform upper bound

A.2.1 Covariance and conditioning

To prove Proposition A.1, note that for any k, conditioning on R_t in estimator (5) gives

$$\mathbb{E} Z_{tk} = \sum_r q_r \frac{a_{kr}}{q_r} (\mu_r - g_r) = \theta_k - G_k = h_k.$$

Similarly, $\mathbb{E}[Z_{tk} Z_{t\ell}] = \sum_r a_{kr} a_{\ell r} v_r / q_r$. Independence over t proves Proposition A.1. When a fixed corpus implements the oracle, also average over the uniformly drawn pool index. If g is fitted on independent pilot observations, condition on the entire pilot; all identities are then unchanged.

A paid pilot. Suppose a bounded predictor is fitted on p pilot queries and the next T independent queries use Envelope. Conditional on any pilot realization, Theorem 4 holds with the same universal constant, for rank predictions or a shared corpus predictor. Integrating over the pilot gives the same unconditional probability bound. The total cost is $N = p + T$. If $p \le \gamma N$ for a fixed $\gamma < 1$, then $T \ge (1-\gamma)N$, so the minimax order is preserved with a constant depending only on γ. No accuracy assumption on the learned predictor is required. Prediction quality affects the covariance in Proposition A.1.

For binary outcomes and the default $g = 1/2$, the residual second moment is $v_r = 1/4$ at every rank. With a learned predictor, v_r can vary with rank and can guide a learned proposal. Conditional on the pilot, Proposition A.1 still applies to any full-support learned proposal; Theorem 4 applies to the Envelope choice.

A.2.2 The entropy integral

We prove Theorem 4 by bounding the mean and upper tail of the estimation error. Let $\mathcal{F}$ be the finite normalized class of Lemma 3. The constant function one bounds every member of $\mathcal{F}$ in absolute value, so we take it as a measurable envelope. For any finitely supported probability measure Q, the covering number satisfies the first bound in Equation (7), and hence the uniform entropy integral obeys

$$\begin{aligned} J(u,\mathcal{F}) &:= \int_0^u \sup_Q \sqrt{1+\log N(\eta,\mathcal{F},L_2(Q))}\,d\eta \\ &\le \int_0^u \sqrt{1+\log(1+2/\eta^2)}\,d\eta \le C_0\, u\sqrt{\log(e/u)}, \qquad 0<u\le 1. \end{aligned} \tag{A.4}$$

For the last inequality, put $\eta = uv$, bound the logarithm by a constant plus $2\log(1/u) + 2\log(1/v)$, and use $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and $\int_0^1 \sqrt{\log(1/v)}\,dv < \infty$. Thus C_0 is a numerical constant independent of every problem parameter.

Write P_T for the empirical measure of iid observations and $\mathbb{G}_T = \sqrt{T}(P_T - P)$. For $L \ge 2$, $B_L > 1$; use $u = B_L^{-1/2}$. Theorem 2.1 of van der Vaart & Wellner (2011) implies

$$\begin{aligned} \mathbb{E}\,\|\mathbb{G}_T\|_{\mathcal{F}} &\le C_1 J(u,\mathcal{F})\left(1 + \frac{J(u,\mathcal{F})}{u^2\sqrt{T}}\right) \\ &\le C_2\left[B_L^{-1/2}\sqrt{\log(eB_L)} + \frac{\log(eB_L)}{\sqrt{T}}\right]. \end{aligned} \tag{A.5}$$

The theorem places the supremum over Q outside the entropy integral, which is bounded by our J; its bound is nondecreasing in J. The cited theorem is written with a strict second-moment inequality. Applying it at $u' > B_L^{-1/2}$ and decreasing u' proves the displayed non-strict case. The class is finite, so its measurability conditions are automatic. Multiplication by $B_L/\sqrt{T}$ gives

$$\mathbb{E}\left\|\widehat{\theta} - \theta\right\|_\infty \le C\left[\sqrt{\frac{B_L\log(eB_L)}{T}} + \frac{B_L\log(eB_L)}{T}\right]. \tag{A.6}$$

For $L = 1$, a bound for the mean of bounded scalar variables handles this case.

A.2.3 Concentration around the mean

Let $\mathcal{E} = \left\|\widehat{\theta} - \theta\right\|_\infty$ and use the finite centered class

$$\mathcal{H} = \left\{\frac{Z_k - \mathbb{E}Z_k}{2B_L},\ -\frac{Z_k - \mathbb{E}Z_k}{2B_L} : k \le L\right\}.$$

Each function has mean zero, absolute value at most one, and second moment at most $1/(4B_L)$. For

$$V = \sup_{h\in\mathcal{H}} \sum_{t=1}^{T} h(O_t) = \frac{T\mathcal{E}}{2B_L},$$

the variance parameter in Bousquet's inequality is at most $v = T/(4B_L) + 2\mathbb{E}V$. For $x > 0$, with probability at least $1 - e^{-x}$, the inequality gives

$$V \le \mathbb{E}V + \sqrt{2vx} + x/3.$$

Rescaling yields

$$\mathcal{E} \le \mathbb{E}\mathcal{E} + \sqrt{\frac{2B_L x}{T} + \frac{8B_L x \mathbb{E}\mathcal{E}}{T}} + \frac{2B_L x}{3T}. \tag{A.7}$$

Using $2\sqrt{ab} \le a + b$ bounds the term involving the mean by

$$\sqrt{8B_L x \mathbb{E}\mathcal{E}/T} \le \mathbb{E}\mathcal{E}/2 + 4B_L x/T.$$

Insert the bound (A.6), set $x = \log(1/\delta)$, and combine square-root terms. This proves Theorem 4 with a universal constant. For $0 < \varepsilon \le 1$, taking $T \ge C_3 B_L \log(eB_L/\delta)/\varepsilon^2$ makes both terms at most $\varepsilon/2$ after increasing C_3.

A.2.4 Coefficient computation

To tabulate the full matrix stably, initialize $a_{kL} = k/L$ and recurse

$$a_{k,r-1} = a_{kr}\frac{r-k}{r-1}, \qquad r = L, L-1, \dots, 2,$$

with zero below rank k. This costs $O(L^2)$ when all coordinates are required.

A.3 Details of the adaptive lower bound

We prove Theorem 5 with

$$J = 1 + \lfloor \log_8 L \rfloor, \qquad \sup_P \mathbb{E}_P \tau \ge \frac{J \log(J/\delta)}{160\varepsilon^2}.$$

The construction uses the top rank and disjoint intervals below it. Mixtures of candidate budgets isolate successive intervals, whose quality changes can be recovered from the top downward.

A.3.1 Mixtures on separated rank intervals

Put $m = \lfloor \log_8 L \rfloor$ and define $J = m + 1$ disjoint sets

$$\mathcal{B}_0 = \{L\}, \qquad \mathcal{B}_j = \{r : 8^{j-1} \le L - r < 8^j\}, \quad j = 1, \dots, m. \tag{A.8}$$

When $m = 0$, only $\mathcal{B}_0$ is present. For $j \ge 1$, write $d_j = 8^{j-1}$ and choose

$$x_j = (2/3)^{1/d_j}, \qquad p_j = 1 - x_j.$$

Consider the following known linear combination of the target curve:

$$\psi_j = \frac{x_j^L}{2} + \sum_{k=1}^{L} \binom{L}{k} p_j^k x_j^{L-k} \theta_k. \tag{A.9}$$

It is the expected quality obtained by including every rank independently with probability p_j, selecting the highest included rank, and assigning the known value $1/2$ to an empty set. Conditional on its size k, the set of included ranks is a uniformly random k-subset, so the coefficient on rank r is

$$w_{jr} = p_j x_j^{L-r} = \sum_{k=1}^{L} \binom{L}{k} p_j^k x_j^{L-k} a_{kr}, \qquad \sum_r w_{jr} = 1 - x_j^L \le 1.$$

With $w_j(\mathcal{B}) = \sum_{r \in \mathcal{B}} w_{jr}$, the mass on $\mathcal{B}_j$ and a bound on the mass of all farther intervals are

$$w_j(\mathcal{B}_j) = x_j^{d_j} - x_j^{8d_j} = c_0 := \frac{2}{3} - \left(\frac{2}{3}\right)^8, \qquad \sum_{\ell > j} w_j(\mathcal{B}_\ell) \le x_j^{8d_j}. \tag{A.10}$$

Their difference is at least

$$\beta := x_j^{d_j} - 2x_j^{8d_j} = \frac{2}{3} - 2\left(\frac{2}{3}\right)^8 > \frac{5}{9}. \tag{A.11}$$

Also set $\psi_0 = \theta_L$, which puts mass one on $\mathcal{B}_0$.

If every $\widehat{\theta}_k$ has error at most ε, substituting these estimates into the definition (A.9) gives $|\widehat{\psi}_j - \psi_j| \le (1 - x_j^L)\varepsilon \le \varepsilon$. The empty-set term is known and contributes no estimation error. These mixtures are used only in the reduction; the auditing procedure may choose ranks and stop arbitrarily.

The same construction gives the envelope-mass lower bound used in Section 3.3. Since $b_r \ge w_{jr}$ and the intervals are disjoint,

$$B_L \ge 1 + c_0 m \ge c_0 J.$$

Consequently $B_L \ge c_B \log(eL)$ for a universal $c_B > 0$ and every $L \ge 2$.

A.3.2 Reduction to independent signs

Let $\Delta = 9\varepsilon/5 \le 9/40$, and let $s = (s_0, \dots, s_m) \in \{-1, 1\}^J$. Give rank r the Bernoulli mean

$$\mu_r(s) = \begin{cases} 1/2 + \Delta s_j, & r \in \mathcal{B}_j, \\ 1/2, & r \notin \bigcup_{j=0}^m \mathcal{B}_j. \end{cases}$$

First recover s_0 by thresholding $\widehat{\theta}_L$ at $1/2$, which succeeds on the uniform-accuracy event because $\Delta > \varepsilon$. Then process $j = 1, \dots, m$ in order, and threshold

$$\widehat{z}_j = \widehat{\psi}_j - \frac{1}{2} - \Delta \sum_{\ell < j} w_j(\mathcal{B}_\ell)\widehat{s}_\ell$$

at zero. Inductively, all previously recovered signs equal their true values on the uniform-accuracy event, so subtracting their known contributions introduces no additional error. For the corresponding true residual,

$$s_j z_j = \Delta \left[w_j(\mathcal{B}_j) + \sum_{\ell > j} s_j s_\ell w_j(\mathcal{B}_\ell) \right] \ge \Delta\beta > \varepsilon.$$

Since $|\widehat{z}_j - z_j| \le \varepsilon$ on that event, thresholding $\widehat{z}_j$ at zero also recovers s_j. Thus uniform curve accuracy identifies all J signs. It suffices to lower-bound the cost of this simultaneous sign-recovery experiment.

A.3.3 Factorization at a stopping time

For an accurate procedure, put $u = \sup_s \mathbb{E}_s \tau$. The lower bound is immediate if $u = \infty$; hence assume $u < \infty$. Let the transcript contain all internal randomization, including randomness used at the final decision. The policy and stopping rule are functions of that randomness and the observed history, so their likelihood factors do not depend on s conditional on a realized history. The Bernoulli likelihood of observations from $\mathcal{B}_j$ depends only on s_j. Hence an independent uniform prior on the signs has a product posterior conditional on the complete stopped transcript. The argument applies on every event $\{\tau = n\}$ and therefore to their countable union, which has probability one.

Let Λ_j be the stopped log likelihood ratio of sign $+1$ against -1 based on observations from $\mathcal{B}_j$. With $c_\Delta = \log((1/2 + \Delta)/(1/2 - \Delta))$,

$$\Lambda_j = c_\Delta \sum_{t=1}^{\tau} \mathbf{1}\{R_t \in \mathcal{B}_j\}(2Y_t - 1).$$

The posterior error of the more likely sign is $e_j = (1 + \exp|\Lambda_j|)^{-1}$. The largest posterior mass of a full sign vector is $\prod_j (1 - e_j)$. Success probability at least $1 - \delta$ under every sign vector implies Bayes success probability at least $1 - \delta$ under the uniform prior, and thus

$$\mathbb{E}\left[1 - \prod_j (1 - e_j)\right] \le \delta. \tag{A.12}$$

By Markov's inequality, the event $\mathcal{A} = \{\prod_j (1 - e_j) \geq 1 - 2\delta\}$ has probability at least $1/2$. On this event, $\sum_j e_j \leq -\log(1 - 2\delta) \leq 4\delta$. Since $e_j \leq 1/2$, Jensen's inequality gives

$$\sum_j |\Lambda_j| = \sum_j \log \frac{1 - e_j}{e_j} \geq \sum_j \log \frac{1}{2e_j} \geq J \log \frac{J}{8\delta} \quad \text{on } \mathcal{A}. \tag{A.13}$$

A.3.4 Expected information and integrability

The posterior mean of s_j is $\tanh(\Lambda_j/2)$. For $x \geq 0$, $x \tanh(x/2) \geq x - 1$, since $2x/(e^x + 1) \leq 1$. Thus

$$\mathbb{E} \sum_j s_j \Lambda_j \geq \mathbb{E} \sum_j |\Lambda_j| - J \geq \frac{J}{2} \log \frac{J}{8\delta} - J. \tag{A.14}$$

All quantities are integrable because $\sum_j |\Lambda_j| \leq c_\Delta \tau$ and $\mathbb{E}\tau \leq u$.

Under the true signs, a query in any interval has expected signed log likelihood increment

$$d = \mathrm{KL}\big(\mathrm{Bern}(1/2 + \Delta) \,\|\, \mathrm{Bern}(1/2 - \Delta)\big) = 2\Delta \log \frac{1/2 + \Delta}{1/2 - \Delta} \leq 30\varepsilon^2.$$

To verify the last bound, the power series of $\operatorname{arctanh}$ shows that $\Delta^{-1} \log((1 + 2\Delta)/(1 - 2\Delta))$ is increasing. For $\Delta \leq 9/40$, therefore,

$$d \leq \frac{80}{9} \log \frac{29}{11} \Delta^2 = \frac{144}{5} \log \frac{29}{11} \varepsilon^2 < 30\varepsilon^2,$$

where $\log(29/11) < 1$.

The event $\{t \leq \tau\}$ is predictable. The following integrability bound justifies summing the conditional expectations of stopped increments:

$$\sum_t \mathbb{E}[\mathbf{1}\{t \leq \tau\} |\text{increment}|] \leq c_\Delta \mathbb{E}\tau < \infty.$$

Consequently

$$\mathbb{E} \sum_j s_j \Lambda_j = d\, \mathbb{E} \sum_{t \leq \tau} \mathbf{1}\{R_t \in \cup_j \mathcal{B}_j\} \leq du.$$

Combining this upper bound with the lower bound (A.14) and the bound on d gives

$$u \geq \frac{J}{60\varepsilon^2} \left[\log \frac{J}{8\delta} - 2\right]. \tag{A.15}$$

When the right-hand side is nonpositive, the inequality follows from $u \geq 0$.

A.3.5 The confidence term and explicit constant

Compare the all-positive sign vector with the vector obtained by flipping only sign j. The event of outputting the all-positive vector has probability at least $1 - \delta$ under the former and at most δ under the latter. Data processing and the stopped likelihood decomposition (Kaufmann et al., 2016, Lemma 1) give

$$d\mathbb{E}_+ N_j \geq \mathrm{kl}(1 - \delta, \delta) = (1 - 2\delta) \log \frac{1 - \delta}{\delta} \geq \frac{2}{3} \log \frac{1}{\delta},$$

where N_j counts queries in $\mathcal{B}_j$. For the last inequality, write $h = \log(1/\delta) \geq \log 8$ and use $\log(1 - \delta) \geq -1/7$; the divergence is at least $3h/4 - 3/28 \geq 2h/3$. The preceding integrability argument also justifies this stopped decomposition. Summing over the disjoint intervals yields

$$u \geq \frac{J}{45\varepsilon^2} \log \frac{1}{\delta}. \tag{A.16}$$

Put $z = \log(J/\delta)$. If $z \leq 10h/3$, Equation (A.16) gives $u \geq Jz/(150\varepsilon^2)$. Otherwise,

$$\frac{5}{8} z > \frac{25}{12} h \geq \frac{25}{12} \log 8 > \log 8 + 2,$$

and Equation (A.15) gives $u \geq Jz/(160\varepsilon^2)$. This proves Theorem 5 for all $L \geq 2$, including $2 \leq L < 8$, where the single top-rank sign suffices.

Finally, $J = 1 + \lfloor \log_8 L \rfloor$ has order $\log L$, and $\log(J/\delta)$ has order $\log\log L + \log(1/\delta)$ on the stated domain. This proves the lower bound in Theorem 2.

A.3.6 REFERENCE COSTS

For one specified k, sample ranks from a_k and average the outcomes. By Hoeffding's inequality, $O(\varepsilon^{-2}\log(1/\delta))$ labels suffice. For necessity, choose every rank mean equal to $1/2 \pm 2\varepsilon$ and apply binary change of measure; any rank then gives the same Bernoulli experiment.

For uniform iid rank allocation, compare two instances differing only at rank L, with its mean $1/2 \pm 2\varepsilon$. The last target is $\theta_L = \mu_L$. If the procedure stops adaptively with finite expectation, its expected number of top-rank samples is $\mathbb{E}\tau/L$ because the next rank is uniform conditional on the past. Binary change of measure therefore gives the worst-case lower bound $\Omega(L\varepsilon^{-2}\log(1/\delta))$ for every accurate estimator. Under uniform rank allocation, averaging the top-rank subsequence gives a matching upper bound for the single endpoint θ_L.

B FINITE-CORPUS ESTIMATION AND COMPLEXITY

B.1 THE FIXED-QUOTA ESTIMATOR

Proposition B.1 (Finite-corpus covariance). *For $M \geq 2$, write $\mu_r = M^{-1}\sum_i y_{ir}$ and $S_r^2 = (M-1)^{-1}\sum_i (y_{ir}-\mu_r)^2$. Independent uniform subsets of sizes n_r within the ranks give*

$$\mathbb{E}\widehat{\theta}_k = \theta_k, \qquad \mathrm{Cov}(\widehat{\theta}_k, \widehat{\theta}_\ell) = \sum_r a_{kr}a_{\ell r}S_r^2\left(\frac{1}{n_r} - \frac{1}{M}\right). \tag{B.1}$$

For $M = 1$, all quotas equal one and the estimate is exact.

Proof. Within one rank, the inclusion indicators I_i of the pools satisfy $\mathbb{E}I_i = n_r/M$ and $\mathbb{E}I_iI_j = n_r(n_r-1)/(M(M-1))$ for $i \neq j$. Expanding the sample mean gives $\mathbb{E}\overline{Y}_r = \mu_r$ and $\mathrm{Var}(\overline{Y}_r) = S_r^2(1/n_r - 1/M)$. Independence across ranks yields the displayed covariance. □

Quota rules. Rule P is the proportional rule in Algorithm 2. Rule A reserves one label per rank: choose $c \geq 0$ so that $x_r = \min\{M, 1 + cq_r^{\mathrm{env}}\}$ sums to n, then use the same largest-remainder rounding. At $n = L$ set all quotas to one; at $n = ML$ set them to M. These endpoints also specify Algorithm 2 when the scalar multiplier is nonunique. Appendix B.3 gives $\kappa_n \leq B_L/(n-L)$ for rule A and $\kappa_n \leq 2B_L/(n-L)$ for rule P.

A quota example. For $L = 3$, where $q^{\mathrm{env}} = (1/5, 1/5, 3/5)$, a six-label audit with $M \geq 4$ gives additive quotas $(2, 1, 3)$ and proportional quotas $(1, 1, 4)$. Under either rule, the audit estimates the same three targets with estimator (10). The proportional rule puts one additional label at the top rank, whose outcomes determine θ_3. At larger label budgets, the requirement of one label per rank takes a smaller share of the allocation.

B.2 MINIMAX TRANSFER TO FINITE CORPORA

In the finite-corpus experiment, a query can reveal any previously unseen entry of an arbitrary fixed array $y \in [0,1]^{M\times L}$, and all later uses of that entry are free. A procedure must estimate $\theta_k(y) = M^{-1}\sum_{i,r} a_{kr}y_{ir}$ within ε simultaneously for every k, with probability at least $1-\delta$ for every array. Let $m^*_{\mathrm{fin}}(M, L, \varepsilon, \delta)$ be the infimum, over such procedures, of the worst-case expected number of distinct entries revealed. Internal randomization and adaptive stopping at an almost surely finite time are allowed. A census, which reveals all ML entries, returns the exact curve.

Lemma B.2 (Concentration of a randomly generated corpus). *Let all ML entries be independent, with $Y_{ir} \in [0,1]$ and $\mathbb{E}Y_{ir} = \mu_r$. For every $\eta > 0$,*

$$\mathbb{P}\left\{\max_{k\leq L}\left|\sum_r a_{kr}(\overline{Y}_r - \mu_r)\right| > \eta\right\} \leq 2e^{-2M\eta^2}.$$

The bound is uniform in L and all rank laws.

Proof. We express each target as a mixture of upper-rank averages and bound all those averages at once.

1. Decompose the weights. For a nonnegative nondecreasing vector $a_1, \dots, a_L$ summing to one, put $a_0 = 0$ and $c_s = s(a_{L-s+1} - a_{L-s})$ for $s = 1, \dots, L$. Then $c_s \geq 0$, $\sum_s c_s = 1$, and, for every vector x,

$$\sum_{r=1}^{L} a_r x_r = \sum_{s=1}^{L} c_s \left(\frac{1}{s} \sum_{r=L-s+1}^{L} x_r \right).$$

Every row of the selection matrix is nondecreasing in rank, so its weighted error is bounded by the largest absolute upper-rank average.

2. Bound the upper-rank averages. Write $W_r = \sum_{i=1}^{M} (Y_{ir} - \mu_r)$ and $S_s = \sum_{r=L-s+1}^{L} W_r$. Hoeffding's bounded-variable moment bound gives $\mathbb{E}e^{\lambda W_r} \leq e^{M\lambda^2/8}$. By independence across ranks,

$$Z_s = \exp\{4\eta S_s - 2M\eta^2 s\}, \qquad Z_0 = 1,$$

is a nonnegative supermartingale when ranks are exposed from the top. If $S_s > M\eta s$ for any $s \geq 1$, then $Z_s > e^{2M\eta^2 s} \geq e^{2M\eta^2}$. The maximal inequality therefore bounds this event by $e^{-2M\eta^2}$. Apply the same argument to $-S_s$ and combine the two tails. Step 1 transfers the result to every target. □

Proposition B.3 (Finite-to-oracle lower-bound transfer)**.** *For $\eta, \beta > 0$ with $\delta + \beta < 1$, suppose $2e^{-2M\eta^2} \leq \beta$. Then*

$$m^*_{\text{fin}}(M, L, \varepsilon, \delta) \geq m^*(L, \varepsilon + \eta, \delta + \beta).$$

Proof. Take any accurate finite-corpus procedure with worst-case expected cost u. Given a rank oracle for arbitrary laws P_r, simulate a product corpus lazily: the first time the procedure requests an entry at rank r, query P_r and assign that observation to the entry. Repeated requests use its stored value. The unobserved entries can be completed independently in the probability space without querying them. The simulation therefore has the law of a product corpus followed by the finite procedure.

The expected oracle cost is at most u, since it equals the number of distinct entries revealed, and averaging the procedure's conditional cost over random arrays cannot exceed its worst-case cost. Conditional on every completed array, the procedure estimates its finite curve within ε except with probability δ. Lemma B.2 puts that curve within η of the population curve except with probability β. A union bound and the triangle inequality give an $(\varepsilon + \eta, \delta + \beta)$-accurate oracle procedure. Taking the infimum proves the comparison. □

Upper bound by caching. Choose a deterministic number T of iid Envelope queries sufficient for the oracle upper bound at (ε, δ). For each query, independently draw a rank from q^{env} and a pool uniformly from $\{1, \dots, M\}$. Reveal the drawn record's label only on its first visit, and use all T addressed values, including repeats, in estimator (5). Conditional on any fixed array, these draws have the required iid rank-oracle law. Thus the curve guarantee holds and the expected paid cost is exactly

$$\mathbb{E}U_T = M \sum_r \left[1 - (1 - q_r^{\text{env}}/M)^T\right] \leq \min\{T, ML\}.$$

Proof of Corollary 7. Use Proposition B.3 with $\eta = \varepsilon$ and $\beta = \delta$. The oracle lower bound applies at $(2\varepsilon, 2\delta)$, whose arguments lie in Theorem 2's domain. On the stated domain, the rate at $(2\varepsilon, 2\delta)$ differs from the rate at (ε, δ) only by universal factors. The cached construction gives the upper bound, with T of the order of the oracle upper bound. □

Corollary 7 gives a sufficient condition on M for the matching rate. For smaller M, cached queries and a census still give valid upper bounds. The product corpus is a device for the lower-bound proof; the finite-corpus guarantee itself holds for every deterministic array.

B.3 Proof of the fixed-quota error bound

We prove Theorem 6 for every fixed array $y \in [0,1]^{M \times L}$. The quotas are deterministic, and the sampled subsets are independent across ranks. The weights, quotas, and radii are independent of the revealed labels. Let $I = \{r : n_r < M\}$, $B = B_L$, and $\kappa = \kappa_n$. If I is empty, every error and radius is zero; otherwise $\kappa > 0$.

1. Control rank-mean errors. Put $e_r = \overline{Y}_r - \mu_r$. For $1 \le n_r < M$, the Hoeffding–Serfling moment bound (Bardenet & Maillard, 2015, Proposition 2.2) gives

$$\log \mathbb{E}e^{\lambda e_r} \le \frac{\lambda^2}{8n_r}\left(1 - \frac{n_r - 1}{M}\right) \qquad (\lambda \in \mathbb{R}).$$

The cited result for positive λ applies to negative λ by replacing each outcome y_{ir} with $1 - y_{ir}$. For $r \in I$, write $u_r = M - n_r$ and let $\overline{Y}^c_r$ be the mean of the unsampled complement. The sample and complement satisfy

$$e_r = -\frac{u_r}{n_r}(\overline{Y}^c_r - \mu_r).$$

The complement is a uniform subset of size u_r. Applying the same moment bound to it and rescaling gives

$$\log \mathbb{E}e^{\lambda e_r} \le \frac{\lambda^2}{8}\frac{u_r}{n_r^2}\left(1 - \frac{u_r - 1}{M}\right) = \frac{\lambda^2}{8}\frac{(M - n_r)(n_r + 1)}{Mn_r^2}.$$

Taking the smaller moment bound, and treating census ranks deterministically, yields

$$\mathbb{E}e^{\lambda e_r} \le e^{\lambda^2 h_r/8}, \qquad h_r = \min\left\{\frac{M - n_r + 1}{Mn_r}, \frac{(M - n_r)(n_r + 1)}{Mn_r^2}\right\}. \tag{B.2}$$

2. Retain each budget's error scale. For $X_k = \sum_r a_{kr}e_r$, independence across ranks gives

$$\mathbb{E}e^{\lambda X_k} \le \exp(\lambda^2 V_k/8), \qquad V_k = \sum_r a_{kr}^2 h_r.$$

For $V_k > 0$, exponential Markov with $\lambda = 4u/V_k$ and the corresponding negative value gives

$$\mathbb{P}\{|X_k| > u\} \le 2e^{-2u^2/V_k} \qquad (u > 0).$$

If $V_k = 0$, every contributing rank is known and $X_k = 0$. A union bound over the remaining budgets therefore yields

$$\mathbb{P}\{\max_k |X_k| > u\} \le S(u) := \sum_{k:V_k>0} 2e^{-2u^2/V_k}. \tag{B.3}$$

If there are K_+ positive V_k, S is continuous and strictly decreasing on $[0, \infty)$, from $2K_+$ to zero. The equation $S(u_\delta) = \delta$ has a unique positive solution for $0 < \delta < 1$. It follows that $\mathbb{P}\{\max_k |X_k| > u_\delta\} \le \delta$. The case $K_+ = 0$ gives $u_\delta = 0$.

3. Cover the budget family. To account for the relation between budgets, define

$$d(k, \ell)^2 = \sum_r h_r(a_{kr} - a_{\ell r})^2.$$

Multiplication of the moment bounds (B.2) and exponential Markov give

$$\mathbb{P}\{|X_k - X_\ell| > u\} \le 2\exp\{-2u^2/d(k, \ell)^2\}. \tag{B.4}$$

A zero distance means the increment is zero almost surely. Add an index 0 with $X_0 = 0$ and $a_{0r} = 0$. Since $a_{kr} \le b_r$,

$$d(k, 0)^2 \le \sum_r h_r b_r a_{kr} \le \kappa \sum_r a_{kr} = \kappa.$$

Set

$$W = \sum_{r\in I} b_r^2 h_r, \qquad Q(r) = b_r^2 h_r/W, \qquad f_k(r) = a_{kr}/b_r.$$

Then $W \le \kappa B$. At every rank, $f_k(r) \in [0, 1]$ is unimodal in k and has total variation at most two, by the coefficient argument in Lemma 3. The cumulative expected variation of $f_1, \ldots, f_k$ under Q is

$$v_k = \sum_{j<k} \mathbb{E}_Q|f_{j+1} - f_j| \in [0, 2].$$

For $k < \ell$, the bound $|f_\ell - f_k| \le 1$ implies

$$d(k,\ell)^2 = W\mathbb{E}_Q(f_\ell - f_k)^2 \le W(v_\ell - v_k). \tag{B.5}$$

At level $j \ge 1$, divide the variation axis into half-open intervals of length $h^{(j)} = \kappa/(W4^j)$, with endpoints at integer multiples of this length. Choose the first budget in each occupied interval as its representative, denoted $\pi_j(k)$ for the interval containing v_k. These partitions are nested: every parent interval has four children. The number of representatives satisfies

$$N_j \le 1 + 2W4^j/\kappa \le 1 + 2B4^j \le 3B4^j.$$

Put $\pi_0(k) = 0$. At level 1 the squared edge distance is at most κ. For $j \ge 2$, a representative and its parent share a level-$(j-1)$ interval. Equation (B.5) therefore gives

$$d(\pi_j(k), \pi_{j-1}(k))^2 \le \kappa 4^{-(j-1)}.$$

Each level-j representative has one parent, so at most N_j distinct edges require control. Because the index set is finite, sufficiently fine partitions separate all distinct values of v_k. Equal values of v_k imply zero distance and identical X. Every chain is consequently exact after finitely many nonzero increments.

4. Sum the edge bounds. Allocate $\delta_j = \delta 2^{-j}$ to level j. Applying the tail bound (B.4) to all its edges and then to all levels gives, with probability at least $1 - \delta$,

$$\begin{aligned} |X_{\pi_j(k)} - X_{\pi_{j-1}(k)}| &\le \sqrt{\kappa/2}\, 2^{-(j-1)} \sqrt{\log(2N_j/\delta_j)} \\ &\le \sqrt{2\kappa}\, 2^{-j} \sqrt{\log(6B/\delta) + j\log 8} \qquad \text{for every } j,k. \end{aligned}$$

Telescope each chain. Jensen's inequality with weights 2^{-j}, which sum to one and under which j has mean two, yields

$$\begin{aligned} \max_k |X_k| &\le \sqrt{2\kappa} \sum_{j \ge 1} 2^{-j} \sqrt{\log(6B/\delta) + j\log 8} \\ &\le \sqrt{2\kappa\,[\log(6B/\delta) + 2\log 8]} = \sqrt{2\kappa \log(384B/\delta)}. \end{aligned}$$

This gives the covering term in $R_n(\delta)$.

5. Combine the radii. Both u_δ and the covering radius are deterministic functions of the known design. Choose whichever is smaller before observing labels. Its failure probability is at most δ by the corresponding argument; no additional union bound is needed for this choice. Also, $\widehat{\theta}_k$ and θ_k belong to $[0,1]$, so the maximum error is at most one. These facts prove the bound (11), including the minimum with one.

6. Account for quotas and rounding. For additive quotas, $x_r = \min\{M, 1 + cq_r^{\mathrm{env}}\}$ and $n = \sum_r x_r \le L + c$, so $c \ge n - L$. If the rounded quota is unsaturated, then $x_r < M$ and largest-remainder rounding gives

$$n_r \ge \lfloor 1 + cq_r^{\mathrm{env}} \rfloor \ge cq_r^{\mathrm{env}}.$$

Thus $b_r h_r \le b_r/n_r \le B/(n-L)$, where the first inequality uses $h_r \le 1/n_r$. Saturated ranks contribute zero. For proportional quotas, $x_r = \min\{M, \max\{1, cq_r^{\mathrm{env}}\}\}$, and again $c \ge n - L$. At an unsaturated rank, $x_r \ge cq_r^{\mathrm{env}}$ and $\lfloor x_r \rfloor \ge x_r/2$ because $x_r \ge 1$. Hence $b_r h_r \le 2B/(n-L)$. These bounds also cover ranks increased by rounding.

Interpretation. Reserving one label per rank costs L labels, and a census costs ML. The two quota rules attain the rate of Corollary 7 when that corollary's condition on M holds and L is at most a constant multiple of the statistical term. The error bound (11) also applies when these conditions on M and L fail.

Computing the radius and choosing a label budget. Evaluate $S(u)$ using log-sum-exp. Bisection starts with zero and $\sqrt{(\max_k V_k)\log(2K_+/\delta)/2}$, whose tail sum is at most δ; retain and return the upper bracket. At a census return zero. This calculation requires no outcome labels. For a desired accuracy, construct quotas at prospective label budgets and choose any n satisfying $R_n(\delta) \le \varepsilon$ before auditing. Largest-remainder quotas need not be nested as n increases, so the search evaluates the radius at each prospective label budget.

For the chosen design and label budget, the intervals $[\widehat{\theta}_k - R_n(\delta), \widehat{\theta}_k + R_n(\delta)] \cap [0,1]$ cover every curve value jointly with probability at least $1-\delta$. An outcome-independent randomized design satisfies the same guarantee after conditioning on its design choice; outcome-dependent quotas, stopping, or fitted residual estimators need a separate argument.

For example, with $M = 512$, $L = 32$, and proportional Envelope quotas, $R_{512}(0.05) < 0.1040$ and $R_{2048}(0.05) < 0.0422$. Thus 2,048 distinct labels suffice for simultaneous error of at most five percentage points at 95% confidence on every fixed array of this size. Table 13 reports the 95% radii for Envelope quotas and the 95th percentiles of maximum error for the compared designs, each averaged over settings within a collection.

C Allocation shape and audit efficiency

C.1 Other proposals with the same order

The proof of Theorem 4 extends to every proposal q that covers each selection probability within a controlled factor.

Proposition C.1 (A sufficient condition on the proposal)**.** *For a fixed full-support q, define $D(q) = \max_r b_r/q_r$. Algorithm 1, with q in place of Envelope, satisfies Theorem 4 with B_L replaced by $D(q)$. The constant C remains universal.*

Proof. Put $D = D(q)$ and $f_k(r,y) = a_{kr}(y - g_r)/(Dq_r)$. Since $Dq_r \ge b_r$, at each observation the functions have one sign, absolute value at most one, and total variation at most two as k increases. The proof of Lemma 3 therefore gives the same covering bound for every probability measure Q. Under the sampling law,

$$\mathbb{E} f_k^2 \le \frac{1}{D^2}\sum_r \frac{a_{kr}^2}{q_r} \le \frac{1}{D}\sum_r a_{kr} = \frac{1}{D}.$$

The maximal inequality and concentration argument in Appendix A.2 now apply with D throughout. □

Harmonic allocation. Let $s = L - r + 1$ and $q_r^{\mathrm{H}} = 1/(sH_L)$. Lemma A.2 gives $b_r \le 1/s$, so $D(q^{\mathrm{H}}) \le H_L$. Equality holds at the top rank, where $b_L = 1$. Therefore

$$D(q^{\mathrm{H}}) = H_L, \qquad \min_r \frac{q_r^{\mathrm{H}}}{q_r^{\mathrm{env}}} = \frac{B_L}{H_L}.$$

Both H_L and B_L are $\Theta(\log L)$, so Harmonic also attains the minimax order. Envelope minimizes $D(q)$ and thus gives the smallest scale parameter in this bound.

Dyadic allocation. Dyadic gives equal mass to each of the $K = \lfloor \log_2 L \rfloor + 1$ nonempty bins $2^j \le s < 2^{j+1}$ and spreads each bin's mass uniformly over its ranks. If bin j has m_j ranks, its members therefore have probability $1/(Km_j)$. For s in that bin, $m_j \le 2^j \le s$. Hence $b_r/q_r \le Km_j/s \le K$. The top rank attains this bound, so $D(q) = K$, and Proposition C.1 gives the same minimax order. Exact envelope weights, harmonic weights, and equal mass across logarithmic bins thus share the order, and their finite-budget constants can differ.

Grid-mixture. Let $\mathcal{K} = \{1, 2, 4, \ldots, 2^{\lfloor \log_2 L \rfloor}\} \cup \{L\}$ and $q_r^{\mathrm{G}} = |\mathcal{K}|^{-1} \sum_{j \in \mathcal{K}} a_{jr}$. Then $D(q^{\mathrm{G}}) = |\mathcal{K}|$, so Proposition C.1 also gives the minimax order for this allocation. To see this, note that a_{kr}/k

is nonincreasing in k by Equation (A.3). For $2^h \le k < 2^{h+1}$,

$$\sum_{j\in\mathcal{K}} a_{jr} \ge \sum_{i=0}^{h} a_{2^i,r} \ge \frac{2^{h+1}-1}{k} a_{kr} \ge a_{kr}.$$

Hence $D(q^{\mathrm{G}}) \le |\mathcal{K}|$. At rank $r = 1$, only $a_{1,1} = 1/L$ is nonzero, giving equality.

Finite-corpus guarantees for other proposals. Theorem 6 applies to every deterministic quota vector satisfying $1 \le n_r \le M$ and $\sum_r n_r = n$. For a full-support proposal q with $D(q) = \max_r b_r/q_r$, its additive and proportional quota conversions satisfy, respectively,

$$\kappa_n \le \frac{D(q)}{n-L}, \qquad \kappa_n \le \frac{2D(q)}{n-L}, \qquad n > L.$$

To verify this, note that $\sum_r x_r \le L + c$ in either conversion, so $c \ge n - L$. If a rounded quota is unsaturated, additive quotas give $n_r \ge \lfloor 1 + cq_r \rfloor \ge cq_r$, while proportional quotas give $n_r \ge \lfloor x_r \rfloor \ge x_r/2 \ge cq_r/2$. Since $h_r \le 1/n_r$ and $b_r \le D(q)q_r$, the displayed bounds follow. Saturated ranks have $h_r = 0$. Thus Harmonic, Dyadic, and Grid-mixture inherit the fixed-quota rate, including the L-label initialization cost. The logarithm in the covering radius remains $\log(384B_L/\delta)$ because B_L is a property of the target family, independent of the proposal.

The caching construction in Appendix B.2 also works for any fixed q: T independent queries, each at a rank drawn from q and a uniformly drawn pool, incur expected distinct-label cost

$$M \sum_r \left[1 - (1 - q_r/M)^T\right] \le \min\{T, ML\}.$$

Choosing T by Proposition C.1 therefore gives these proposals the same finite minimax upper order without first labeling every rank, and the finite-corpus lower bound applies to them because it permits arbitrary audit procedures.

C.2 Efficiency of power-law allocations

Power(α) samples rank r with probability proportional to $s^{-\alpha}$, where $s = L - r + 1$, and $Z_\alpha = \sum_{s=1}^{L} s^{-\alpha}$ is its normalizer. For estimator (5) with $g = 1/2$ and iid outcomes $Y \mid r \sim \mathrm{Bern}(1/2)$, the endpoint variances are

$$\mathrm{Var}(\widehat{\theta}_L) = \frac{Z_\alpha}{4T}, \qquad \mathrm{Var}(\widehat{\theta}_1) = \frac{Z_\alpha}{4TL^2} \sum_{s=1}^{L} s^\alpha. \tag{C.1}$$

Integral comparisons prove that the first variance is $\Theta(L^{1-\alpha}/T)$ for fixed $0 < \alpha < 1$ and that the second is $\Theta(L^{\alpha-1}/T)$ for fixed $\alpha > 1$, with constants depending on α. Too little mass near the top thus harms θ_L, which depends only on the top rank, and too much harms θ_1, the average over all ranks. At $\alpha = 1$, both variances are $O(\log L/T)$. These formulas motivate the powers $1/2$ and $3/2$ used in Appendix C.3.

C.3 Allocation-shape experiments

Table 4 compares these proposals at matched label budgets using the shared estimator with $g = 1/2$.

Table 4: Allocation-shape ablation under independent queries. Mean maximum errors and Monte Carlo SEs are multiplied by 10^3. The first four columns use $T = 512$; the last two use $T = 2L$. Every proposal uses the shared estimator (5) with $g = 1/2$.

Method	PPE	SummEval	WMT	SAL 256	SAL 1024	SAL 2048
Power ($\alpha = 1/2$)	63.8 (1.2)	39.2 (0.8)	40.8 (0.8)	105.1 (1.9)	77.0 (1.4)	64.8 (1.2)
Grid-mixture	47.1 (0.6)	26.3 (0.4)	27.9 (0.4)	57.8 (0.7)	33.1 (0.4)	24.8 (0.3)
Harmonic	48.2 (0.7)	28.4 (0.5)	29.6 (0.5)	60.0 (0.8)	34.5 (0.5)	25.9 (0.3)
Power ($\alpha = 3/2$)	54.2 (0.8)	28.2 (0.4)	31.2 (0.5)	86.5 (1.4)	61.4 (1.0)	52.9 (0.9)
Dyadic	54.7 (0.8)	35.7 (0.6)	32.8 (0.6)	68.4 (1.0)	39.5 (0.6)	28.7 (0.4)
Envelope	45.2 (0.6)	25.7 (0.4)	26.9 (0.4)	55.3 (0.6)	31.4 (0.4)	23.4 (0.3)

The rate-optimal allocations differ at finite budgets: Envelope has the lowest mean error in these six summaries, followed by Grid-mixture, Harmonic, and Dyadic. The power allocations also show the cost of concentrating too little or too much near the top rank. At $L = 2048$, the errors of Power$(1/2)$ and Power$(3/2)$ are respectively 2.78 and 2.27 times Envelope's.

Constants and observed error. At $L = 32$ and 2048, Envelope's maximum importance ratio B_L is 0.648 and 0.507 times Harmonic's H_L, respectively. For both allocations, the largest second-moment proxy $\max_k \sum_r a_{kr}^2/q_r$, which MMV minimizes, also equals B_L and H_L, respectively, and is attained at $k = L$. Expected maximum error depends on all marginal variances and their correlations, so it need not decrease by the same factor. At $L = 2048$, MMV has maximum ratio 6.088 and maximum second-moment proxy 3.195, compared with Envelope's 4.159 for both. Because the two designs minimize different objectives, Envelope's optimal maximum ratio can coexist with an empirical error comparable to MMV's. Figure 4 illustrates these differences at $L = 32$.

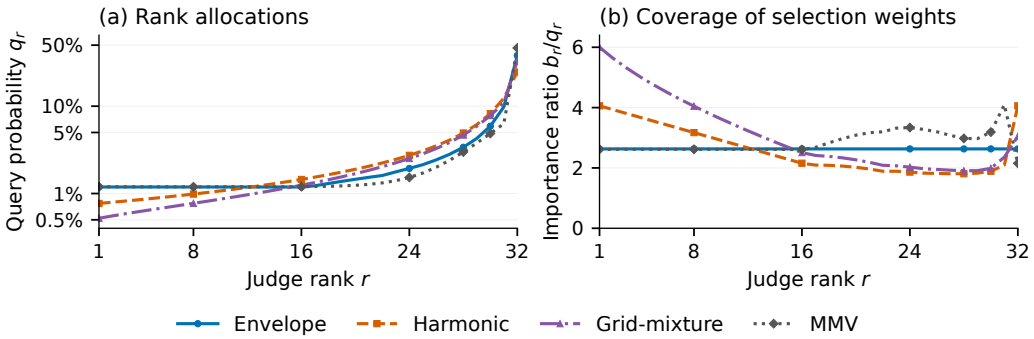

Figure 4: Allocation geometry at $L = 32$. (a) Query probabilities, on a logarithmic vertical scale. (b) Rank-wise importance ratios b_r/q_r, whose maximum is $D(q)$. Envelope equalizes these ratios at $B_{32} = 2.631$. MMV minimizes the maximum second-moment proxy $\max_k \sum_r a_{kr}^2/q_r$, which is 2.155 for MMV and 2.631 for Envelope.

C.4 Resolution of the candidate-budget grid

The geometric grid is $\mathcal{K} = \{1, 2, 4, \ldots\} \cap [1, L]$ together with L. Lemma A.2 and scalar Bernstein concentration give simultaneous grid accuracy using $O(\varepsilon^{-2} B_L \log(|\mathcal{K}|/\delta))$ queries. Since $|\mathcal{K}| = O(\log L)$, this count has the same order as the full-curve upper bound, and Lemma 3 controls the intermediate budgets without increasing that order.

Exact grid values alone do not generally determine those budgets. Consider $L = 4$ and the two monotone rank-mean vectors

$$\mu^+ = (0.25, 0.30, 0.65, 0.80), \qquad \mu^- = (0.15, 0.50, 0.55, 0.80).$$

Both give $(\theta_1, \theta_2, \theta_4) = (1/2, 2/3, 4/5)$, but $\theta_3^+ = 0.7625$ and $\theta_3^- = 0.7375$. With generation cost $c_k = 0.05(k-1)$, their optimal budgets are respectively 3 and 4. Thus even identical grid curves can lead to different decisions, including when rank quality is monotone.

Table 5 evaluates the full curve and the grid from the same Envelope audit. Because the audit estimates every integer budget, the two error metrics differ only in the coordinates over which they take the maximum. Across the six collection summaries, the mean full-curve error is only 1.4%–2.1% larger than the mean grid error. With the allocation and the audit held fixed, this comparison measures the statistical cost of the added resolution.

Table 5: Full-curve and geometric-grid errors from the same Envelope estimates; entries are means (Monte Carlo SEs), multiplied by 10^3. Regret uses $c_k = 0.1(k-1)/(L-1)$. The full choice maximizes estimated utility over all k; the grid choice maximizes it over $\mathcal{K}$. Both are compared with the best true utility over all k. Label budgets and the independent-query protocol match Table 4.

Collection	Full error	Grid error	Full choice regret	Grid choice regret
PPE	45.2 (0.6)	44.5 (0.6)	6.5 (0.3)	6.8 (0.3)
SummEval	25.7 (0.4)	25.1 (0.4)	3.6 (0.2)	4.5 (0.3)
WMT	26.9 (0.4)	26.4 (0.4)	4.3 (0.3)	4.3 (0.3)
SAL 256	55.3 (0.6)	54.5 (0.6)	9.3 (0.4)	9.4 (0.5)
SAL 1024	31.4 (0.4)	30.9 (0.4)	4.1 (0.2)	4.3 (0.2)
SAL 2048	23.4 (0.3)	23.0 (0.3)	2.9 (0.1)	3.0 (0.1)

D Data and experimental protocol

D.1 Outcome arrays and rankings

The PPE data come from the `ppe-result-data` archive released by the PPE authors. The three judges are ArmoRM-Llama3-8B-v0.1, Skywork-Reward-Llama-3.1-8B, and internlm2-1_8b-reward, abbreviated as ArmoRM, Skywork, and InternLM in the figures. From the archive's five `best_of_k` directories, we use the complete scalar-score files, which provide one scalar score for every candidate. Label-selected pairwise comparisons in other files of the archive are outside this experiment. Within each task, pools are ordered by question identifier and generator name. The four tasks other than MBPP+ each include 128 pools from each of Gemma-2-9B-IT, GPT-4o-mini-2024-07-18, Meta-Llama-3-8B-Instruct, and Claude-3-Haiku-20240307, and MBPP+ contains 248 Claude, 158 Llama, 63 GPT, and 38 Gemma pools. The score files of the three judges have identical pool identifiers and identical verification outcomes, and we check both equalities before analysis.

PPE's collection protocol filters pools using correctness fractions and reports best-of-k evaluation as part of its original benchmark (Frick et al., 2025). We use this filtered collection and each task's verification outcome as released, so our curves are conditional on PPE's filter.

For SummEval, the base outcome file is joined to each G-Eval dimension file by document and system identifier. Human scores must be identical across these files, and each document must contain 16 systems. A judge response is parsed when it begins, after optional whitespace, with an integer from 1 through 5 followed by a non-digit or the end of the string. The ranking score of a summary is the mean of its parsed scores, and every summary has at least one parsed response. The numbers of unparsed responses are 0, 81, 7, and 0 for coherence, consistency, fluency, and relevance. The fluency judge prompt uses a 1–3 rubric, and every parsed fluency response is in that range. Human fluency ratings use the released 1–5 scale. Monotone rescaling of a judge score leaves the ranking unchanged. Human outcomes are the released dimension averages transformed by $(y-1)/4$.

In PPE and SummEval, we break score ties with independent uniform keys, generated before any audit from a deterministic hash of the setting name and the master seed, so the ranking depends only on the scores and these keys. Every method and replication in a setting uses the same ordering.

Expert translation outcomes. We join WMT22 expert MQM scores (Freitag et al., 2022) with released GEMBA GPT-4 direct-assessment scores (Kocmi & Federmann, 2023). For each of the two language pairs, English–German and Chinese–English, source-only and reference-based GPT-4 scores give two ranking settings. We include every non-reference system with human scores and both ranking scores. This availability rule excludes one English–German system and four Chinese–English systems, which have no human annotations. We then retain all complete segments across the included systems, which yields 1,315 and 1,875 pools, respectively, each with 14 translations. The preparation code in our anonymous repository records the included systems and segment identifiers.

The expert scores are negative error penalties. We define the audited outcome as the bounded quality $\min\{1, \max\{0, 1+\mathrm{MQM}/25\}\}$, using the 0–25 MQM error scale documented by WMT, and choose this clipping before comparing methods. Scores below -25 occur in 0.049% of the retained English–German records and 0.933% of the Chinese–English records. GPT-4's discrete scores produce many

ties, and, as in the PPE and SummEval protocol, a fixed independent random key determines their order. The two ranking settings of a language pair share the same human records. Each pool is a fixed menu of system outputs, and the estimand averages the outcome of the winner of a random candidate subset within the pool.

Mathematical candidate pools. SAL (Beeching et al., 2024) releases 256, 1,024, and 2,048 Llama-3.2-1B responses per problem for the same 500 MATH500 problems. We rank each response by its last-step PRM score. Equal scores are ordered by the original candidate index, with the first candidate winning, as in the publisher's selection code. The outcome is correctness under the publisher's pinned answer extraction and mathematical checker. Duplicate pairs of extracted answer and reference answer reuse their checked outcome. Strings rejected by the checker's input guard and checker calls that exceed its three-second time limit receive outcome zero. All 500 problem identifiers and candidate counts are checked; the two larger arrays contain 512,000 and 1,024,000 outcomes.

The 1,024-candidate release has 104,205 distinct pairs of extracted answer and reference answer: 104,062 completed checks, 33 timeouts, and 110 strings rejected by the input guard. The 2,048-candidate release has 178,241 pairs: 177,943 completed checks, 64 timeouts, and 234 rejected strings. The response selected by our score and tie order has the same hash as the publisher's selected response on all 500 problems in both releases, and all 9,199 shared response occurrences have consistent labels. Appendix E.3 quantifies the effect of the tie convention on the target curve.

Unequal pool sizes. The experiments retain complete pools of the stated size. For a collection with variable sizes, one can choose L before any label is seen, no larger than the smallest retained pool, and draw a uniform L-subset from each pool before ranking, independently of outcome labels. The theory then applies to those fixed reduced pools. Auditing all original variable-size pools would use size-dependent weights and a correspondingly different target family.

D.2 Allocation objectives and implementation

All proposals are computed from the selection matrix. The numerical objectives use fixed unit proxies for residual second moments in independent-query MMV and GW, and for rank variances in finite MMV. Independent-query audits use estimator (5) with $g = 1/2$, and distinct-label audits use estimator (10). Uniform and Target-mixture have proposals $1/L$ and $L^{-1}\sum_k a_{kr}$, respectively, and Grid-mixture takes the same average over the geometric grid $\mathcal{K}$. These mixture proposals use the randomized balance-heuristic form of multiple importance sampling (Veach & Guibas, 1995; Elvira et al., 2019). Envelope is q^{env} in Equation (4). With $s = L - r + 1$, Power(α) uses $q_r = s^{-\alpha}/\sum_{u=1}^{L} u^{-\alpha}$ for $\alpha \in \{1/2, 1, 3/2\}$; $\alpha = 1$ is Harmonic. Dyadic splits s into the nonempty bins $2^j \le s < 2^{j+1}$, assigns equal probability to each bin, and samples uniformly within it. Its last bin can be truncated by L.

Numerical objectives. MMV minimizes the largest marginal second-moment proxy,

$$\min_{q\in\Delta_L} \max_k \sum_r \frac{a_{kr}^2}{q_r}.$$

For distinct labels, Finite MMV solves the continuous quota problem

$$\min_{1\le x_r\le M,\,\sum_r x_r=n} \max_k \sum_r a_{kr}^2(1/x_r - 1/M). \tag{D.1}$$

Finite MMV uses the same largest-remainder rounding as Algorithm 2, and in distinct-label audits every other primary proposal uses proportional quotas (rule P). As an additional control for the quota conversion, additive Envelope applies additive quotas (rule A) to the same Envelope proposal.

Solving MMV. For the independent-query objective, let $C_{kr} = a_{kr}^2$ and introduce a distribution p over target budgets. The dual is

$$\max_{p\in\Delta_L} \left(\sum_r \sqrt{\sum_k p_k C_{kr}}\right)^2, \qquad q_r(p) = \frac{\sqrt{\sum_k p_k C_{kr}}}{\sum_s \sqrt{\sum_k p_k C_{ks}}}.$$

For finite quotas, put $w_r = \sum_k p_k C_{kr}$ and $\mathcal{P} = \{r : w_r > 0\}$. The inner problem minimizes $\sum_r w_r(1/x_r - 1/M)$ over $1 \le x_r \le M$ and $\sum_r x_r = n$. If $n \le M|\mathcal{P}| + L - |\mathcal{P}|$, a solution is $x_r = \min\{M, \max\{1, c\sqrt{w_r}\}\}$, with $c \ge 0$ chosen to match the budget. Otherwise, set $x_r = M$ on $\mathcal{P}$ and distribute the remaining budget among the zero-weight ranks within $[1, M]$. At $n = L$ or $n = ML$, set all quotas to 1 or M, respectively. We use pairwise conditional-gradient steps, followed by full correction of the active target mixture. The largest target variance is a primal upper bound U, and the mixture value is a dual lower bound D. We report the relative gap $\max\{0, U - D\}/\max\{U, 10^{-12}\}$ and request a tolerance of 10^{-6}. Integer rounding is applied afterward, so the gap certifies the continuous design before rounding.

Gaussian-width design. For 4,096 fixed independent $G_i \sim N(0, I_L)$, Gaussian-width (GW) minimizes

$$\frac{1}{4096\beta} \sum_{i=1}^{4096} \log \sum_{k=1}^{L} \sum_{s \in \{-1,1\}} \exp\Big(\beta s\, a_k^\top \operatorname{diag}(q)^{-1/2} G_i\Big), \qquad \beta = 30.$$

This objective is a smooth proxy for the expected largest absolute Gaussian error. We minimize it from two starts, Envelope and MMV, by L-BFGS in softmax coordinates with analytic gradients, at most 500 iterations per start, objective tolerance 10^{-12}, and gradient tolerance 10^{-7}. The solution with the lower training objective is the GW proposal, and an independent set of 32,768 Gaussian draws evaluates its unsmoothed width. To keep the matrix products sparse, the optimization omits coefficients below 10^{-16}; validation and audit estimation use the full matrix. The solver diagnostics record termination and discarded row mass. This design is a sample-average heuristic motivated by Gaussian-width analysis (Katz-Samuels et al., 2020).

Connection to multi-policy evaluation. Viewed as a single-state, single-step problem of multi-policy evaluation, an audit has ranks as actions and the known distributions a_k as target policies. With all action distributions available, the objective in Chen et al. (2025, Equation 4) is then the oracle MMV objective, and our MMV comparator solves it over all distributions q. The additional structure studied here is the relation between the target policies across k.

Allocation construction cost. The explicit allocations in Appendix C can be constructed from their formulas without numerical optimization. Envelope uses L log-binomial evaluations and a normalization, with linear memory. Table 6 reports construction times for Envelope and the numerical designs. Each timed CPU job has eight allocated cores. For Envelope we report the median time of 21 constructions, and for MMV and GW the times of their completed solver runs at the tolerances above. The numerical timings start from the weight matrix and exclude its construction. Estimating the curve after labels arrive is common to the methods and is outside these setup timings. For a fixed pool size, the oracle proposals can be reused across tasks and audits; finite MMV additionally depends on the corpus size and label budget.

Table 6: Proposal construction times, in milliseconds for Envelope and in seconds for the numerical designs. Both MMV runs have relative primal–dual gaps below 10^{-6}.

L	Envelope (ms)	MMV (s)	GW (s)
1,024	0.045	0.343	178.0
2,048	0.077	0.575	709.9

The same implementation constructs Envelope in 2.93 ms at $L = 10^5$ and 30.62 ms at $L = 10^6$. At these pool sizes, we rely on the theoretical bounds for accuracy.

D.3 Budgets, randomization, and reporting

Every setting, label budget, and protocol receives 1,000 audit replications. For $L \le 32$, the label budgets are $128, 256, 512, 1024$; for $L = 256$, they are $512, 1024, 2048, 8192$; and for $L = 1024$ and 2048, they are $2L, 4L, 8L$. In each of the 26 settings of Table 2, we compare nine proposals under both protocols: Uniform, Target-mixture, Envelope, Harmonic, Power$(1/2)$, Power$(3/2)$, Dyadic,

MMV, and GW. Additive Envelope is an additional fixed-quota control. Grid-mixture uses the budgets in Tables 2 and 4, with the same paired draws. The budgets and proposals in the nine-proposal comparison were fixed before the allocation-shape experiment; the underlying corpora had already been studied.

Independent-query methods share the pool indices and the uniform variables used to draw ranks. Repeated queries of a record count toward T, and we also track the number of distinct records they reveal. Fixed-quota methods share independent random permutations of the records within each rank and take the first n_r records, so each pays for exactly n labels. Within-setting comparisons are therefore paired. Stable SHA-256 offsets from master seed 20260915 index each setting, budget, and protocol.

The primary error metric is the maximum error $\max_k |\widehat{\theta}_k - \theta_k|$. We report its mean and Monte Carlo standard error (SE), with paired SEs in tables of differences. Collection summaries average the setting means and, conservatively, their within-setting SEs. These SEs describe audit randomization conditional on the released data. Settings share records across judges, and the SAL releases share problems. Full labels are used only for reference curves and evaluation metrics.

All arrays are fixed before sampling. At $\varepsilon = \delta = 0.05$, the sufficient condition of Corollary 7 is $M \geq 738$, which the WMT arrays satisfy. Theorem 6 applies to every array, including PPE ($M = 507$–512), SummEval ($M = 100$), and SAL ($M = 500$).

These corpus experiments use released responses and scores and run on CPUs. Candidate checking uses Python 3.11; audit simulations use Python 3.14, NumPy 2.3.5, and SciPy 1.18.1 on compute nodes with eight allocated cores and at most 48 GB of memory. Our anonymous code repository includes acquisition scripts, pinned source revisions, numerical design diagnostics, and machine-readable results for every setting and budget.

E ALLOCATION BUDGETS AND SENSITIVITY

Tables 7–12 compare Uniform, Target-mixture, MMV, GW, and Envelope at every tested label budget. Entries are collection summaries of the mean maximum error, multiplied by 10^3, with 1,000 replications per setting and budget. In the independent-query columns, every draw counts toward T; in the quota columns, each distinct label counts once toward n. Quotas are proportional except for additive Envelope, marked (A), and MMV optimizes its finite-quota objective before rounding. Dashes mark the independent-query entries of additive Envelope, because additive quotas apply only to distinct-label audits. Table 4 compares allocation shapes, and our anonymous code repository includes all tested shapes and budgets, with Monte Carlo SEs.

The additive-quota rows of Tables 7–12 isolate the effect of the quota conversion: at $n = 2L$, reserving one label per rank consumes half the budget. At $L = 2048$ and $n = 2L$, Envelope's mean error is 0.0232 with additive quotas and 0.0195 with proportional quotas.

Table 7: PPE: mean maximum error ($\times 10^3$) across label budgets.

	Independent queries (T)				Distinct labels (n)			
Method	128	256	512	1024	128	256	512	1024
Uniform	213.8	151.3	107.3	75.5	214.6	151.4	105.7	72.6
Target-mixture	139.8	98.8	70.1	50.0	94.9	75.8	61.2	46.4
MMV	93.9	66.2	46.6	33.2	90.9	62.1	41.7	27.4
GW	91.5	64.6	45.3	32.3	89.3	60.9	41.4	26.2
Envelope (A)	–	–	–	–	90.7	61.8	41.4	26.2
Envelope	91.2	64.6	45.2	32.2	89.5	61.0	41.3	26.2

Table 8: SummEval: mean maximum error ($\times 10^3$) across label budgets.

	Independent queries (T)				Distinct labels (n)			
Method	128	256	512	1024	128	256	512	1024
Uniform	125.5	90.7	61.8	44.2	29.1	19.9	12.6	6.4
Target-mixture	67.0	47.6	33.5	23.7	35.1	26.4	14.2	5.7
MMV	53.0	37.6	26.5	18.9	20.9	13.4	7.8	3.8
GW	51.2	36.2	25.8	18.3	22.4	14.5	8.1	3.9
Envelope (A)	–	–	–	–	20.9	14.2	8.0	3.8
Envelope	51.1	36.2	25.7	18.3	22.1	14.5	8.0	3.8

Table 9: WMT: mean maximum error ($\times 10^3$) across label budgets.

	Independent queries (T)				Distinct labels (n)			
Method	128	256	512	1024	128	256	512	1024
Uniform	124.5	89.2	63.5	44.2	34.7	25.4	17.9	12.4
Target-mixture	73.9	52.9	37.8	26.5	29.4	23.5	15.4	11.2
MMV	56.6	39.6	27.9	20.0	22.3	15.8	10.9	7.5
GW	54.2	38.4	26.9	19.2	21.9	15.5	10.8	7.4
Envelope (A)	–	–	–	–	21.6	15.4	10.8	7.4
Envelope	54.0	38.3	26.9	19.2	22.0	15.5	10.8	7.4

Table 10: SAL 256: mean maximum error ($\times 10^3$) across label budgets.

	Independent queries (T)				Distinct labels (n)			
Method	512	1024	2048	8192	512	1024	2048	8192
Uniform	301.4	214.2	151.3	76.4	303.0	219.2	148.1	73.0
Target-mixture	174.7	129.1	91.6	45.6	60.4	43.6	31.7	15.9
MMV	56.9	40.8	28.2	14.5	55.3	35.4	22.9	10.0
GW	54.6	39.3	28.3	14.1	55.3	34.4	21.3	8.1
Envelope (A)	–	–	–	–	62.3	37.0	21.7	8.1
Envelope	55.3	39.6	28.1	14.1	55.1	34.1	21.5	8.0

Table 11: SAL 1024: mean maximum error ($\times 10^3$) across label budgets.

	Independent queries (T)			Distinct labels (n)		
Method	2048	4096	8192	2048	4096	8192
Uniform	287.0	207.7	154.1	304.8	207.8	148.0
Target-mixture	183.6	124.3	90.9	38.0	24.3	16.9
MMV	32.0	22.7	16.1	28.9	17.8	12.1
GW	31.8	22.6	16.1	28.9	16.5	10.9
Envelope (A)	–	–	–	33.0	17.5	10.7
Envelope	31.4	22.4	15.8	28.4	16.1	10.8

Table 12: SAL 2048: mean maximum error ($\times 10^3$) across label budgets.

	Independent queries (T)			Distinct labels (n)		
Method	4096	8192	16384	4096	8192	16384
Uniform	294.0	207.7	153.3	299.4	210.2	151.4
Target-mixture	183.5	125.7	92.6	27.7	17.1	12.0
MMV	23.3	16.8	11.9	21.1	12.9	8.9
GW	23.8	17.5	12.1	19.7	11.7	7.7
Envelope (A)	–	–	–	23.2	12.6	7.5
Envelope	23.4	16.7	11.6	19.5	11.5	7.5

E.1 TAIL ERROR AND SIMULTANEOUS BOUNDS

Our guarantees bound the largest error over the entire curve in one audit, so Table 13 reports the 95th percentile of this error, together with the design-based radius $R_n(0.05)$ for Envelope quotas. We compute each setting's percentile and radius before averaging over the settings of a collection. Theorem 6 guarantees coverage for each fixed setting; the table summarizes these settings. The ordering of the methods is similar to that in the mean-error comparison.

For these designs, $u_{0.05}$ is smaller than the covering term in Theorem 6, so the scalar equation gives the reported finite-sample radii. The covering argument establishes the order of the sufficient label cost as the target family grows.

Table 13: The 95th percentile of the maximum error ($\times 10^3$), averaged over settings, each with 1,000 audits. Label budgets are 512 for $L \leq 256$ and $2L$ otherwise. The final row gives Theorem 6's 95% radii for Envelope quotas, averaged over settings and rounded upward. Quantiles use the reference curve; radii use only the known design.

Independent queries (T)

Method	PPE	SummEval	WMT	SAL 256	SAL 1024	SAL 2048
Uniform	243.6	143.7	145.5	696.0	708.0	686.0
Target-mixture	145.0	68.7	79.7	371.3	420.5	402.4
MMV	80.8	46.3	49.2	94.6	53.1	37.3
GW	79.2	47.0	50.3	90.4	52.1	38.7
Envelope	79.7	46.4	49.6	93.8	52.0	37.9

Distinct labels (n)

Method	PPE	SummEval	WMT	SAL 256	SAL 1024	SAL 2048
Whole-pool	240.1	28.2	42.1	555.1	550.1	564.0
Uniform	240.0	26.9	40.0	554.0	542.0	564.0
Target-mixture	125.1	32.5	33.1	108.1	72.0	53.1
MMV	72.3	15.3	19.7	93.4	46.6	34.4
GW	71.6	17.7	19.5	94.3	49.7	34.7
Envelope	71.2	17.7	19.6	94.4	49.7	33.4
Envelope: 95% radius	104.0	64.0	99.7	153.0	84.3	58.3

E.2 SETTING-LEVEL UNCERTAINTY

Table 14 gives setting-level errors and paired comparisons of Envelope with MMV and GW under both protocols, all at the label budgets of Tables 2 and 4. The collection tables report the average within-setting Monte Carlo SE. By the covariance bound $|\operatorname{Cov}(X, Y)| \leq \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$, this average upper-bounds the SE of the collection mean even if its settings are dependent, and it measures audit randomization conditional on these corpora. The repository includes complete per-setting summaries at every budget, including distinct-record counts.

Table 14: Setting-level comparisons with optimized designs. E gives Envelope's mean maximum error; differences give means ± 1.96 paired Monte Carlo SE, all multiplied by 10^3. Positive differences favor Envelope. The label budget is 512 for $L \leq 256$ and $2L$ for larger pools. Each interval is a marginal Monte Carlo interval for a single comparison.

	Independent queries			Distinct labels		
Setting	E	MMV − E	GW − E	E	MMV − E	GW − E
GPQA / Armo	46.0	$+0.5 \pm 1.1$	$+0.1 \pm 0.6$	41.9	$+0.6 \pm 0.6$	-0.1 ± 0.3
GPQA / Sky	44.7	$+1.6 \pm 1.1$	$+0.3 \pm 0.6$	40.7	$+0.9 \pm 0.6$	$+0.1 \pm 0.3$
GPQA / Intern	45.5	$+1.3 \pm 1.1$	-0.2 ± 0.6	41.1	$+0.5 \pm 0.6$	$+0.1 \pm 0.3$
IFEVAL / Armo	45.6	$+1.4 \pm 1.1$	$+0.2 \pm 0.6$	41.8	-0.0 ± 0.6	-0.0 ± 0.3
IFEVAL / Sky	44.5	$+1.1 \pm 1.1$	$+0.1 \pm 0.6$	41.8	$+0.3 \pm 0.6$	-0.1 ± 0.3
IFEVAL / Intern	44.6	$+2.5 \pm 1.1$	$+0.1 \pm 0.6$	42.1	$+0.1 \pm 0.6$	-0.2 ± 0.3
MATH / Armo	45.4	$+0.8 \pm 1.1$	-0.1 ± 0.5	40.9	$+0.7 \pm 0.5$	-0.1 ± 0.3
MATH / Sky	45.6	$+1.3 \pm 1.1$	-0.4 ± 0.6	40.0	$+0.5 \pm 0.5$	$+0.2 \pm 0.3$
MATH / Intern	44.3	$+1.3 \pm 1.1$	$+0.0 \pm 0.6$	40.8	$+0.2 \pm 0.5$	$+0.0 \pm 0.3$
MBPP+ / Armo	45.3	$+1.5 \pm 1.2$	$+0.3 \pm 0.6$	41.8	$+0.1 \pm 0.6$	-0.0 ± 0.3
MBPP+ / Sky	45.9	$+1.0 \pm 1.1$	$+0.6 \pm 0.6$	40.4	$+0.6 \pm 0.6$	$+0.1 \pm 0.3$
MBPP+ / Intern	45.9	$+2.0 \pm 1.2$	$+0.0 \pm 0.6$	41.9	$+0.9 \pm 0.6$	-0.0 ± 0.3
MMLU-Pro / Armo	45.3	$+1.4 \pm 1.1$	-0.1 ± 0.6	41.9	$+0.0 \pm 0.5$	$+0.1 \pm 0.3$
MMLU-Pro / Sky	45.6	$+0.5 \pm 1.1$	$+0.2 \pm 0.6$	41.2	$+0.4 \pm 0.5$	-0.1 ± 0.3
MMLU-Pro / Intern	44.3	$+2.0 \pm 1.1$	-0.1 ± 0.6	41.8	$+0.1 \pm 0.6$	$+0.4 \pm 0.3$
SAL 1024	31.4	$+0.6 \pm 0.7$	$+0.4 \pm 0.6$	28.4	$+0.5 \pm 0.5$	$+0.5 \pm 0.3$
SAL 2048	23.4	-0.1 ± 0.5	$+0.4 \pm 0.5$	19.5	$+1.6 \pm 0.4$	$+0.3 \pm 0.2$
SAL 256	55.3	$+1.6 \pm 1.3$	-0.7 ± 1.0	55.1	$+0.3 \pm 0.7$	$+0.3 \pm 0.3$
SummEval / coherence	23.3	$+0.4 \pm 0.5$	$+0.1 \pm 0.2$	10.0	$+0.4 \pm 0.3$	$+0.0 \pm 0.1$
SummEval / consistency	29.2	$+1.3 \pm 0.6$	$+0.2 \pm 0.2$	7.6	-0.9 ± 0.2	$+0.1 \pm 0.1$
SummEval / fluency	28.0	$+1.4 \pm 0.6$	$+0.2 \pm 0.2$	6.6	-0.5 ± 0.2	$+0.1 \pm 0.1$
SummEval / relevance	22.2	$+0.4 \pm 0.5$	$+0.1 \pm 0.2$	7.7	$+0.1 \pm 0.2$	$+0.0 \pm 0.1$
WMT en-de refA	26.8	$+0.9 \pm 0.6$	$+0.1 \pm 0.2$	7.2	$+0.0 \pm 0.1$	-0.0 ± 0.1
WMT en-de src	27.1	$+0.7 \pm 0.6$	-0.2 ± 0.2	7.5	$+0.0 \pm 0.1$	-0.0 ± 0.1
WMT zh-en refA	26.6	$+1.3 \pm 0.6$	$+0.1 \pm 0.2$	14.3	-0.1 ± 0.2	$+0.0 \pm 0.1$
WMT zh-en src	27.2	$+1.1 \pm 0.6$	$+0.0 \pm 0.2$	14.4	$+0.3 \pm 0.2$	$+0.1 \pm 0.1$

E.3 Sensitivity to score ties

Our target breaks score ties by a fixed order, and averaging over a uniformly random order gives a second well-defined target. For a tied block occupying ranks $u, \ldots, v$ in one pool, each response has average selection weight

$$\overline{a}_{k,[u,v]} = \frac{1}{v-u+1} \sum_{r=u}^{v} a_{kr}.$$

This follows because each response is equally likely to occupy every position in the block. Replacing the outcomes within each tied block by their average and applying a_k therefore computes the target averaged over uniform tie-breaking exactly. We use this identity only to measure target sensitivity (Table 15); the main comparisons use the fixed tie order. The additional audits below test three other fixed tie orders.

The tie order leaves θ_1 unchanged but can change θ_L, because it decides which member of a large highest-score block is selected. Computed from complete corpora, Table 15 separates the choice of selection rule from the sampling error of an audit.

Auditing other fixed tie orders. For each SAL release, we generate three independent random orderings within every equal-score block, using random keys independent of outcomes. Each ordering is frozen before auditing and defines its own reference curve. We then repeat the comparison of Envelope, Harmonic, MMV, and GW with 1,000 audits at a label budget of $2L$ under each protocol (Table 16) to test whether allocation performance is stable when the selection rule changes.

Table 15: SAL target sensitivity to tie-breaking. The fixed target orders ties by original candidate index, as in the publisher's selection code, and the uniform target averages over all orders within each tied block. Adjacent ties is the percentage of neighboring rank pairs with equal scores, averaged over pools; $\Delta\theta_k$ is the difference between the fixed and uniform targets. All entries are exact corpus quantities.

L	Adjacent ties (%)	Fixed θ_L	Uniform θ_L	$\max_k \lvert\Delta\theta_k\rvert$
256	80.2	0.4460	0.4394	0.0073
1024	93.9	0.4580	0.4467	0.0113
2048	96.7	0.4360	0.4475	0.0115

Table 16: Mean maximum error ($\times 10^3$) under three randomly generated tie orders. Rows give the SAL release, the seed of the tie order, and the protocol (query: independent queries; quota: distinct labels). Within a row, all methods share the reference curve and paired audits.

Setting / seed / protocol	Envelope	Harmonic	MMV	GW
256 / 0 / query	55.88	61.96	57.96	56.31
256 / 0 / quota	54.50	60.36	54.88	54.36
256 / 1 / query	56.22	61.24	56.59	56.35
256 / 1 / quota	54.12	60.20	54.45	54.16
256 / 2 / query	55.90	61.75	57.53	55.49
256 / 2 / quota	55.60	61.33	55.80	55.77
1024 / 0 / query	30.64	34.00	31.21	31.59
1024 / 0 / quota	28.56	30.95	28.73	28.64
1024 / 1 / query	31.36	34.60	31.56	31.84
1024 / 1 / quota	28.65	31.40	29.07	28.82
1024 / 2 / query	30.74	33.78	31.72	31.64
1024 / 2 / quota	28.39	31.34	28.84	28.54
2048 / 0 / query	23.12	25.12	23.41	23.71
2048 / 0 / quota	19.63	20.50	20.49	19.73
2048 / 1 / query	23.13	25.47	23.43	23.86
2048 / 1 / quota	19.57	20.15	20.40	19.58
2048 / 2 / query	23.14	25.98	23.66	23.99
2048 / 2 / quota	19.81	20.40	20.45	19.85

Under each ordering, Envelope, MMV, and GW give comparable errors, and Envelope has lower error than Harmonic. Paired error differences and Monte Carlo SEs for all comparisons are in the repository.

E.4 Whole-pool audits and input-opening costs

Whole-pool samples $h = \lfloor n/L \rfloor$ complete pools uniformly without replacement. If $e = n - hL > 0$, it samples e ranks uniformly in one additional distinct pool. Each rank mean averages its h or $h + 1$ observed outcomes, and estimator (10) combines the rank means into the curve estimate. Conditional on the chosen remainder ranks, the observed pools are uniform for each rank, so each rank mean is unbiased. The procedure pays for $hL + e = n$ distinct labels. When $e = 0$, its curve estimate is the average of the full-label curves of the h sampled pools. The audit simulation keeps the dependence between ranks within each sampled pool. Whole-pool's random draws are separate from those shared by the rank-stratified comparisons.

When $n = hL$, let S be the finite covariance matrix of ranked outcomes across pools. Then

$$\mathrm{Var}(\widehat{\theta}_k^{\mathrm{pool}}) = \left(\frac{1}{h} - \frac{1}{M}\right) a_k^\top S a_k.$$

The off-diagonal entries of S capture the dependence between outcomes from the same input. Across the 26 settings of Table 2, Whole-pool has lower mean error than Envelope on SummEval consistency (0.0066 versus 0.0076) and higher error in the other 25 settings. At $n = 512$, it allocates 32 labels to each of the two lowest ranks, compared with Envelope's 19. Its advantage in this case comes from allocating more labels to the variable lower ranks.

Reading a new input can have a fixed cost shared by its candidate labels. We charge $C = n + \alpha U$, where n is the number of distinct labels and U is the number of distinct inputs opened. Labels can be grouped by input, so an input is charged once per audit. We compare proportional Envelope quotas with Whole-pool on all four SummEval dimensions and four WMT rankings, at an expected cost of 512 and $\alpha \in \{0, 0.5, 1, 5\}$.

For independent rank subsets, $\mathbb{E}U = M[1 - \prod_r (1 - n_r/M)]$; Whole-pool opens $\lceil n/L \rceil$ inputs, including a possible partial final pool. We choose the label budgets before any label is seen, randomizing between adjacent budgets so that the expected cost is exactly 512. Because the comparison fixes only the expected cost, realized costs can exceed 512. Each setting uses 1,000 audits.

Table 17: Mean maximum error ($\times 10^3$) at equal expected cost $C = n + \alpha U$. Parentheses give average within-setting Monte Carlo SEs. Each collection includes all four settings.

Collection	Method	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 5$
SummEval	Envelope	8.1 (0.2)	8.8 (0.2)	9.8 (0.2)	23.0 (0.4)
SummEval	Whole-pool	14.4 (0.2)	14.8 (0.2)	15.1 (0.2)	17.4 (0.3)
WMT	Envelope	10.8 (0.1)	13.3 (0.2)	15.5 (0.2)	26.6 (0.4)
WMT	Whole-pool	20.7 (0.4)	21.0 (0.4)	21.3 (0.4)	24.1 (0.4)

In Table 17, Envelope has lower collection-average error for each tested $\alpha \leq 1$, and Whole-pool has lower average error at $\alpha = 5$, when opening an input costs five labels. Rank allocation and input grouping thus address different cost components, so their relative value depends on the annotation workflow.

F Quality curves and audit reconstructions

F.1 Reference curves across tasks and judges

Figure 5 shows the reference curves of all PPE tasks and judges, each computed from all verification labels in its fixed corpus. The three judges share responses and outcomes within each task, so their curves start at the same value at $k = 1$, and larger budgets expose differences in their rankings. Because the selection weights at $L = 32$ (Figure 1(c)) are the same for every judge, the curves differ

only through the outcome means along each ranking. Figure 4 compares the label allocations at this pool size.

A four-candidate example explains how a peak can arise. Outcomes $(0, 0, 1, 0)$, ordered from lowest to highest judge score, give $(\theta_1, \theta_2, \theta_3, \theta_4) = (1/4, 1/3, 1/4, 0)$ by Equation (2). Two candidates give the highest quality: larger subsets more often include the incorrect top-ranked response.

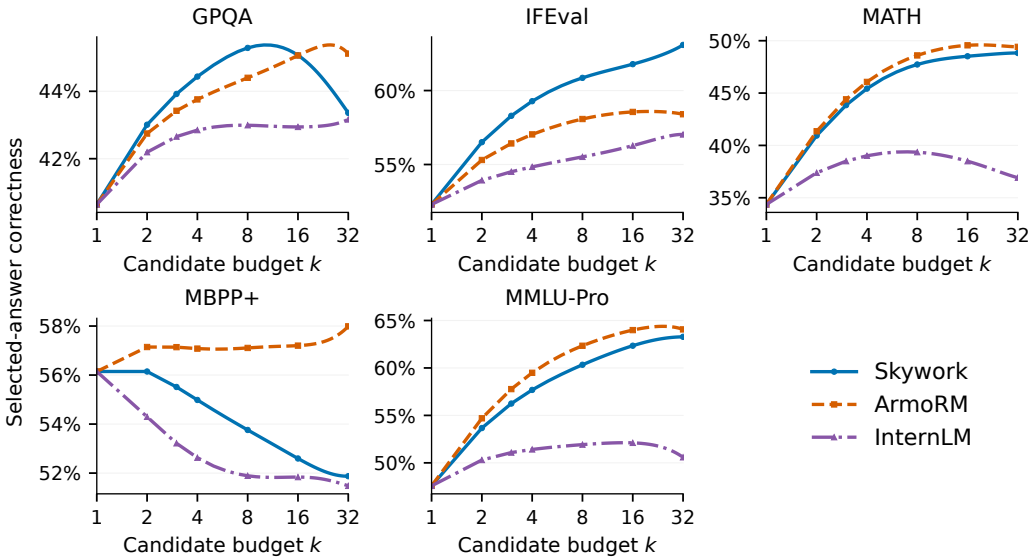

Figure 5: All 15 PPE reference curves, with panel-specific vertical scales. Every integer $k \in \{1, \ldots, 32\}$ is evaluated from complete labels. Lines join adjacent budgets on logarithmic horizontal axes; markers highlight $k = 1, 2, 3, 4, 8, 16, 32$. Judge names abbreviate the models in Appendix D.

F.2 Reconstruction protocol

Figure 2 uses three cases chosen for the shapes of their reference curves: the rising curve of Skywork on MATH, the interior peak of InternLM on MATH, and the declining curve of Skywork on MBPP+. The first two hold the response pools fixed while changing the judge; the first and third use the same judge across tasks. Bands show the pointwise 2.5th–97.5th percentiles over 1,000 audits conditional on the corpus. They describe variation at each candidate budget; Theorem 6 provides a simultaneous guarantee. The repository provides curve-reconstruction summaries under both sampling protocols and the sampled record indices for the first replication.

G Scaling and independent-outcome simulations

G.1 Full-curve scaling

We simulate Envelope audits with $Y \mid r \sim \mathrm{Bern}(1/2)$, so the true quality is $1/2$ at every candidate budget. For each L, we run 2,000 independent audits under each of two schedules for the label budget, $T = \lceil 200 B_L \rceil$ and $T = \lceil 200 B_L \log(eB_L) \rceil$. A multinomial draw of the counts in the $2L$ categories formed by rank and outcome implements T iid oracle queries, and every candidate budget $k \le L$ is evaluated. With $T = \lceil 200 B_L \rceil$, mean maximum error increases across the displayed pool sizes; adding the factor $\log(eB_L)$ keeps it approximately stable (Figure 6). This comparison illustrates the query growth in Theorem 4 over the tested range, and Theorem 5 shows that this growth is necessary in the worst case.

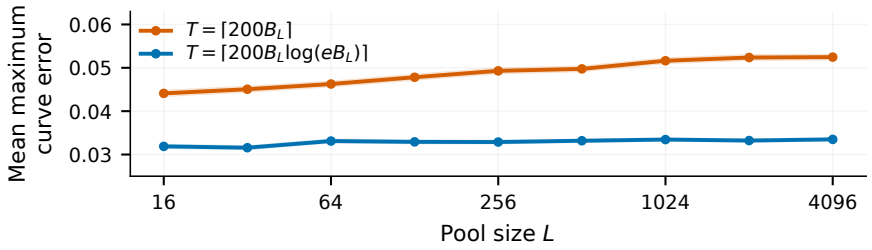

Figure 6: Envelope full-curve error under two query schedules, with fresh $\mathrm{Bern}(1/2)$ outcomes. Each point averages the maximum error over all $k \le L$ in 2,000 audits; shading gives ± 1.96 Monte Carlo SEs of that mean.

G.2 Fixed label budgets on nested real pools

We use the released $L = 2048$ SAL pools for both Llama 1B and 3B. For each input, an independent random candidate permutation supplies nested subsets of size $32, 128, 512, 2048$, chosen without inspecting outcomes or scores. Each subset inherits the original score ordering and tie rule, and its complete labels define its own reference curve. At each size, Uniform, Target-mixture, Grid-mixture, Harmonic, and Envelope spend the same fixed label budget of $T \in \{256, 512, 1024\}$ queries, which may revisit records, and use the shared estimator with $g = 1/2$. We run 1,000 paired audits for each comparison.

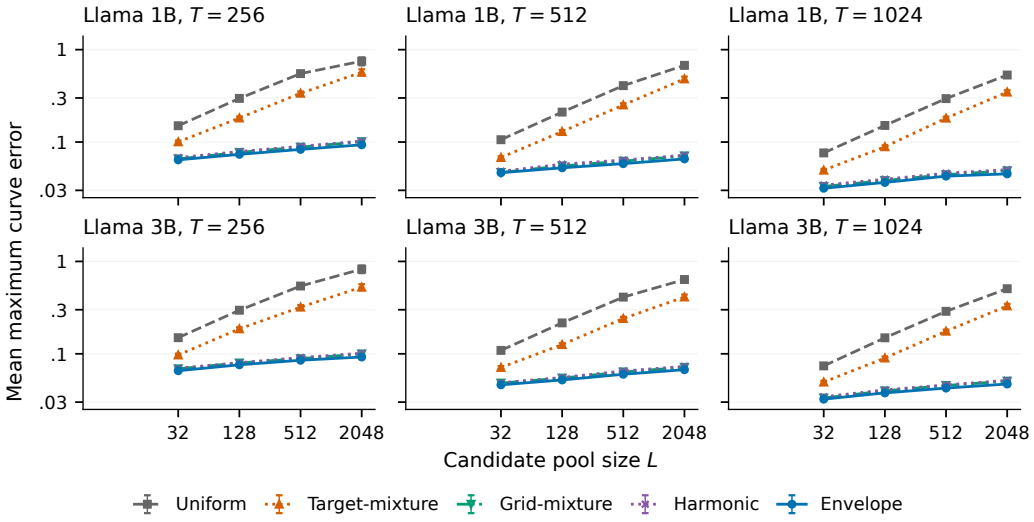

Figure 7: Full-curve error at fixed label budgets as candidate pools grow. Axes are logarithmic; bars give ± 1.96 Monte Carlo SEs over audit randomization conditional on the reduced pools. Envelope, Harmonic, and Grid-mixture maintain much smaller errors than Uniform and Target-mixture.

Envelope, Harmonic, and Grid-mixture, the three rate-optimal allocations, maintain much smaller errors than Uniform and Target-mixture as the pools grow (Figure 7). For example, at $T = 512$, Envelope's mean error changes from about 0.047 to 0.067 across $L = 32$ to 2048, while Uniform's grows from about 0.11 to 0.64–0.68. Because this comparison holds the label budget fixed, including when $T < L$, it illustrates the value of covering ranks across scales (Section 3.3). Theorem 2 supplies the worst-case complexity characterization.

H PREDICTION-ASSISTED ESTIMATION

H.1 PREDICTION AND ALLOCATION AS SEPARATE CHOICES

A score predictor estimates each record's outcome from its judge score; an allocation determines which residuals to measure. We compare these choices under the same total label budget across all 34 settings. Every setting uses 1,000 paired replications, with $T = 512$ for $L \leq 32$ and $T = 2L$ otherwise. This experiment uses its own random draws, so its mean errors for Envelope with $g = 1/2$ differ slightly from those of the same settings in Appendix E and Table 19.

Predictors and costs. For a fixed record-level predictor g_{ir}, the shared estimator (5) becomes

$$G_k = \frac{1}{M}\sum_{i,r} a_{kr} g_{ir}, \qquad \widehat{\theta}_k = G_k + \frac{1}{T}\sum_{t=1}^{T} \frac{a_{kR_t}}{q_{R_t}}(y_{I_t R_t} - g_{I_t R_t}).$$

The raw-score predictions are judge scores mapped to $[0, 1]$ without trusted labels: SummEval's judge ratings are rescaled from the rubric of the judge prompt (1–3 for fluency, 1–5 otherwise), GEMBA scores are divided by 100, and SAL and RLHFlow use the released probability scores. Uniform + raw scores and Envelope + raw scores use these same predictions with their respective rank allocations. PPE reward scores have no corresponding probability scale, so we omit both raw-score methods for PPE. The estimator above applies the prediction-plus-correction principle underlying prediction-powered inference and its stratified variant (Angelopoulos et al., 2023a; Fisch et al., 2024) to the vector of rank-weighted means, so it remains unbiased when the scores are miscalibrated.

Isotonic plug-in fits an increasing least-squares map from judge score to quality using T uniformly sampled records and returns G_k without a correction. The fitted map is linearly interpolated between its knots and held constant beyond its observed score range. Envelope + fitted scores uses the same fitting rule on a pilot of $p = T/8$ uniformly sampled records and spends the remaining $T - p$ queries, drawn from Envelope, on the residual correction. Every training label counts toward the label budget T, and no evaluation labels select the fit, the pilot fraction, or the score map. Conditional on the pilot, the corrected estimator is unbiased, and the covering proof in Appendix A.2.1 applies to these bounded record-level predictions.

Table 18: Prediction and allocation at the same total label budget T. Mean maximum error ($\times 10^3$), with average within-setting Monte Carlo SEs. Columns average 15, 4, 4, 4, and 7 settings. Raw-score corrections fit no parameters, and fitted-score corrections count their pilot labels toward T. Dashes indicate that raw-score methods are omitted for PPE because its reward scores have no probability scale.

Estimator	PPE	SummEval	WMT	SAL	RLHFlow
Envelope, $g = 1/2$	45.5 (0.6)	25.8 (0.4)	27.0 (0.4)	33.4 (0.4)	30.2 (0.3)
Uniform + raw scores	–	21.3 (0.4)	17.1 (0.3)	433.6 (9.2)	268.9 (6.4)
Envelope + raw scores	–	16.7 (0.3)	10.7 (0.1)	43.3 (0.5)	26.1 (0.3)
Isotonic plug-in	69.5 (0.9)	12.0 (0.2)	13.8 (0.1)	269.9 (0.5)	210.3 (0.3)
Envelope + fitted scores	49.8 (0.7)	12.9 (0.2)	11.5 (0.2)	32.6 (0.4)	25.7 (0.3)

Raw scores reduce Envelope's mean error on SummEval, WMT, and RLHFlow but increase it on SAL (Table 18). Changing Uniform to Envelope with the same raw predictions improves every collection summary. Isotonic plug-in is effective for the human-rating collections, but its fitted curve has substantially larger error on mathematical pools; the importance correction removes the need for a correct score-to-quality model. Prediction and rank allocation therefore address different sources of error. An additional experiment in the repository uses paid pilot labels to fit both predictions and residual-based allocations, with results at every pilot fraction tested.

I LARGE-POOL COMPARISONS AND BUDGET SELECTION

I.1 CANDIDATE POOLS AND MATCHED COMPARISONS

The large-pool experiments use nine configurations of task, generator, and judge (Table 3). SAL supplies Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct responses on the same 500 MATH500 questions, with $L = 2048$. Both sets of responses are ranked by the last-step score of the released RLHFlow Llama3.1-8B PRM trained on DeepSeek data. We apply the same frozen Qwen answer parser and checker to both generators. As in Appendix D, strings rejected by the checker's input guard and calls that exceed the three-second time limit receive outcome zero.

RLHFlow (Xiong et al., 2024) supplies seven configurations with $L = 1024$: DeepSeek-Math-7B-Instruct on MATH500 and GSM8K (Cobbe et al., 2021), each scored by two ORMs; MetaMath-finetuned Mistral-7B on MATH500 scored by a PRM and an ORM; and the same Mistral generator on GSM8K scored by that ORM. MATH500 has 500 questions and GSM8K has 1,319. The two ORMs are Llama3.1-8B-ORM-Deepseek-Data and Llama3.1-8B-ORM-Mistral-Data; the additional PRM is Llama3.1-8B-PRM-Mistral-Data. Thus the four large-pool judges share a Llama3.1-8B base, with different training data or supervision objectives. The reported Mistral generator is the publisher's MetaMath finetune; its precise generating checkpoint is not specified in the release.

Each RLHFlow record contains the full response list, binary outcome labels, and response scores or step scores. We use the released outcome labels for the primary analysis. The PRM score is the mean across steps, following the publisher's description; ORMs supply scalar scores. Higher scores receive higher ranks, and ties favor the response with the smallest original index, as in SAL. A sensitivity analysis in Appendix I.5 ranks responses by the last-step PRM score, a change of selection rule that keeps the checkpoint and responses fixed.

We compare six matched pairs (Figure 8): the 1B and 3B SAL generators with a fixed PRM; the two ORMs on DeepSeek's MATH500 responses; the two ORMs on its GSM8K responses; the PRM and the ORM on Mistral's MATH500 responses; and Mistral and DeepSeek with the fixed Mistral-data ORM on each task. Within each shared response pool, candidate texts, their original ordering, and outcome labels are identical across judges. All MATH500 configurations use the same question identifiers, and both GSM8K generators use the same 1,319 questions, so the configurations share their questions across judges and candidate counts.

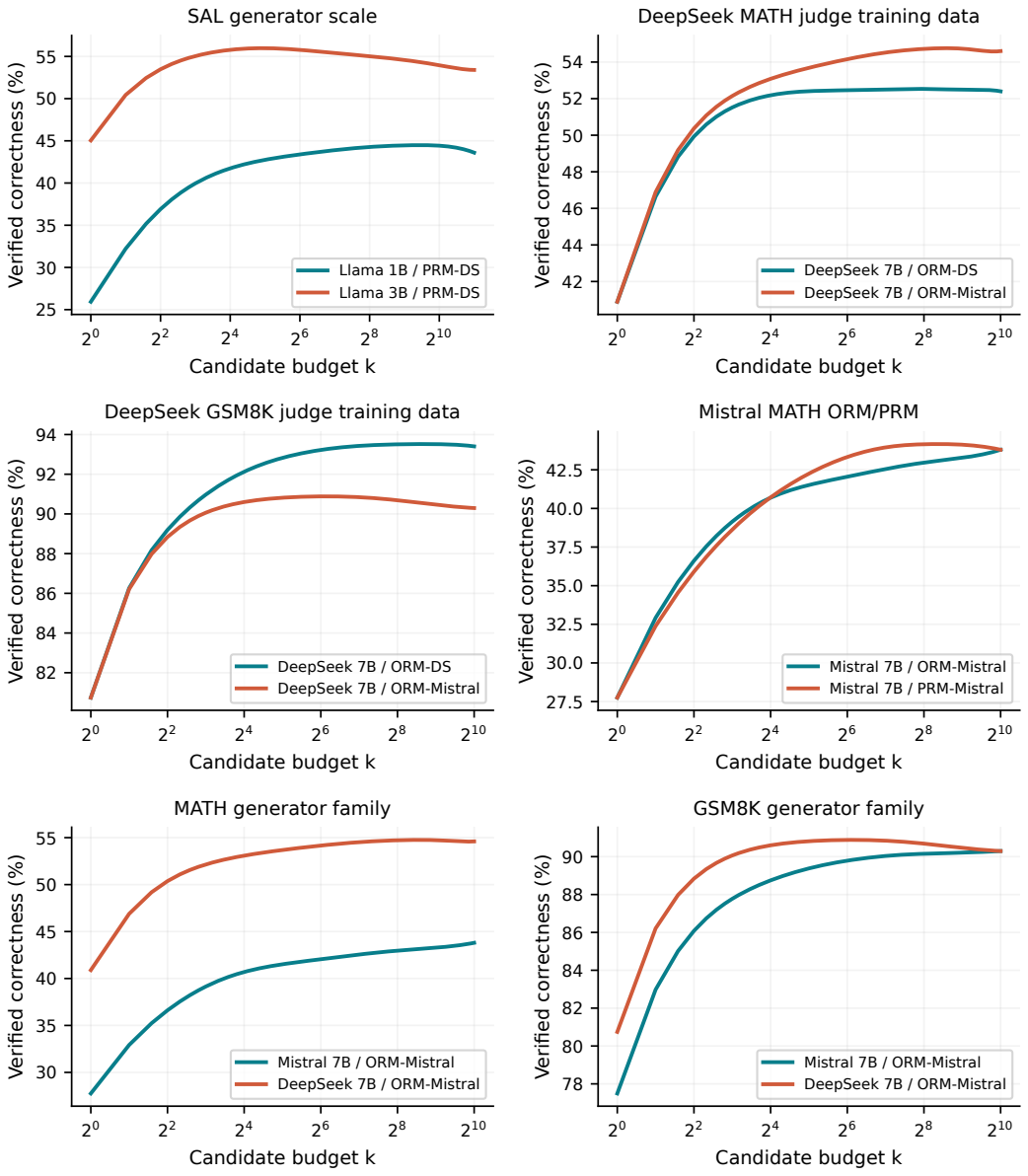

Figure 8: The six prespecified matched comparisons of generators or judges. Complete outcome labels determine each curve. Each comparison isolates a change of generator under a fixed scoring rule or a change of scoring rule on identical responses. In the comparison of the PRM with the ORM, the PRM's scoring rule includes its step-score aggregation, the mean over steps.

I.2 Fixed-corpus audit protocol and results

Every configuration uses 1,000 paired audit replications at label budgets $2L$, $4L$, and $8L$. Configuration 1 (SAL 1B) is also the SAL 2048 setting of Table 2; here we audit it again with the paired randomization used for the other configurations. Independent-query audits count every query toward the budget, including repeats. Distinct-label audits pay once per label and share random within-rank permutations across the compared allocations. MMV quotas are optimized for $M = 500$ on MATH500 and $M = 1319$ on GSM8K, with relative primal–dual tolerance 10^{-6}; other proposals use proportional bounded quotas. Because every budget is a multiple of L, Whole-pool samples n/L complete question pools without replacement.

Tables 19 and 20 give the errors of the main methods at all three budgets. Results for Dyadic, Power$(1/2)$, and Power$(3/2)$, together with 95th-percentile errors and paired differences, are in the repository.

Table 19: Independent-query mean maximum errors ($\times 10^3$), with Monte Carlo SEs from 1,000 audits. IDs follow Table 3.

ID	T/L	Uniform	Target-mixture	Harmonic	MMV	GW	Envelope
1	2	304.7 (6.8)	177.0 (3.6)	25.4 (0.3)	23.6 (0.3)	24.0 (0.3)	23.1 (0.2)
2	2	301.6 (6.3)	172.0 (3.5)	25.9 (0.3)	24.4 (0.3)	24.3 (0.3)	23.5 (0.3)
3	2	302.7 (6.8)	181.6 (3.5)	33.8 (0.4)	31.8 (0.3)	32.0 (0.4)	31.4 (0.3)
4	2	296.7 (6.8)	184.9 (3.7)	34.3 (0.4)	31.9 (0.3)	32.2 (0.4)	31.5 (0.3)
5	2	300.0 (6.8)	178.3 (3.5)	33.3 (0.4)	31.8 (0.3)	31.4 (0.3)	30.9 (0.3)
6	2	294.0 (6.4)	180.4 (3.6)	33.9 (0.4)	31.6 (0.3)	31.2 (0.3)	30.8 (0.3)
7	2	302.5 (6.5)	179.4 (3.6)	31.8 (0.4)	28.4 (0.3)	28.4 (0.3)	28.1 (0.3)
8	2	299.5 (6.5)	176.0 (3.5)	31.7 (0.4)	28.6 (0.3)	28.8 (0.3)	28.4 (0.3)
9	2	301.4 (6.5)	172.6 (3.5)	32.0 (0.4)	29.2 (0.3)	29.0 (0.3)	28.7 (0.3)
1	4	207.4 (4.5)	126.3 (2.5)	18.1 (0.2)	16.6 (0.2)	16.9 (0.2)	16.2 (0.2)
2	4	215.1 (4.5)	125.5 (2.5)	18.3 (0.2)	17.0 (0.2)	17.2 (0.2)	16.8 (0.2)
3	4	211.3 (4.6)	130.3 (2.6)	24.3 (0.3)	22.2 (0.2)	22.6 (0.2)	22.1 (0.2)
4	4	208.9 (4.6)	131.4 (2.6)	24.0 (0.3)	22.4 (0.2)	22.3 (0.2)	22.0 (0.2)
5	4	215.3 (4.5)	128.8 (2.5)	24.7 (0.3)	22.0 (0.2)	22.3 (0.2)	21.9 (0.2)
6	4	209.3 (4.4)	128.5 (2.6)	24.7 (0.3)	22.6 (0.2)	22.3 (0.2)	21.9 (0.2)
7	4	210.1 (4.5)	130.8 (2.5)	22.8 (0.3)	20.5 (0.2)	20.3 (0.2)	20.0 (0.2)
8	4	207.9 (4.5)	129.8 (2.5)	22.7 (0.3)	20.7 (0.2)	20.5 (0.2)	20.0 (0.2)
9	4	207.4 (4.3)	130.3 (2.5)	22.7 (0.3)	21.0 (0.2)	21.0 (0.2)	20.5 (0.2)
1	8	159.4 (3.3)	91.1 (1.9)	12.4 (0.2)	11.7 (0.1)	11.9 (0.1)	11.6 (0.1)
2	8	158.5 (3.2)	92.7 (1.8)	12.4 (0.2)	11.7 (0.1)	11.9 (0.1)	11.6 (0.1)
3	8	149.8 (3.1)	88.7 (1.7)	17.5 (0.2)	16.1 (0.2)	16.1 (0.2)	16.1 (0.2)
4	8	148.0 (3.1)	89.7 (1.7)	17.3 (0.2)	15.9 (0.2)	16.0 (0.2)	15.7 (0.2)
5	8	153.6 (3.2)	89.7 (1.7)	17.2 (0.2)	15.9 (0.2)	15.8 (0.2)	15.5 (0.2)
6	8	151.0 (3.1)	89.8 (1.7)	17.1 (0.2)	16.0 (0.2)	15.8 (0.2)	15.8 (0.2)
7	8	154.2 (3.2)	91.3 (1.8)	15.7 (0.2)	14.3 (0.2)	14.2 (0.2)	14.1 (0.2)
8	8	150.7 (3.2)	92.0 (1.7)	16.0 (0.2)	14.6 (0.2)	14.6 (0.2)	14.3 (0.2)
9	8	151.4 (3.1)	90.8 (1.7)	15.9 (0.2)	14.7 (0.2)	14.4 (0.2)	14.3 (0.2)

Table 20: Distinct-label mean maximum errors ($\times 10^3$), with Monte Carlo SEs from 1,000 audits. IDs follow Table 3.

ID	n/L	Uniform	Target-mixture	Harmonic	MMV	GW	Envelope	Whole-pool
1	2	295.9 (6.0)	27.8 (0.4)	20.5 (0.2)	20.6 (0.2)	19.8 (0.2)	19.7 (0.2)	347.6 (5.2)
2	2	301.7 (6.3)	28.3 (0.4)	21.1 (0.2)	21.2 (0.2)	20.6 (0.2)	20.5 (0.2)	340.3 (5.9)
3	2	299.1 (6.3)	37.9 (0.6)	32.0 (0.4)	29.0 (0.3)	29.1 (0.4)	28.8 (0.3)	336.8 (5.8)
4	2	292.7 (6.1)	38.0 (0.6)	32.1 (0.4)	29.5 (0.3)	28.6 (0.3)	28.7 (0.3)	338.1 (5.8)
5	2	311.5 (6.2)	37.2 (0.6)	31.5 (0.4)	28.6 (0.3)	28.4 (0.4)	28.3 (0.3)	352.6 (5.1)
6	2	312.4 (6.2)	36.0 (0.5)	31.9 (0.4)	29.1 (0.3)	28.3 (0.4)	28.2 (0.3)	345.5 (5.1)
7	2	112.7 (3.9)	21.8 (0.3)	18.7 (0.2)	18.1 (0.2)	18.1 (0.2)	18.1 (0.2)	202.0 (4.2)
8	2	157.3 (4.1)	23.7 (0.3)	21.2 (0.3)	19.6 (0.2)	19.7 (0.2)	19.6 (0.2)	221.5 (4.6)
9	2	152.6 (4.1)	25.5 (0.3)	22.0 (0.3)	20.7 (0.2)	20.6 (0.2)	20.6 (0.2)	246.3 (4.3)
1	4	212.4 (4.2)	17.7 (0.3)	11.5 (0.1)	13.0 (0.1)	11.9 (0.1)	11.6 (0.1)	243.7 (3.7)
2	4	215.2 (4.4)	17.8 (0.3)	12.3 (0.2)	13.4 (0.1)	12.5 (0.2)	12.2 (0.2)	255.5 (4.0)
3	4	218.9 (4.3)	24.5 (0.4)	17.4 (0.2)	18.7 (0.2)	17.5 (0.2)	17.3 (0.2)	251.9 (4.0)
4	4	215.8 (4.5)	24.7 (0.4)	17.6 (0.2)	19.1 (0.2)	17.8 (0.2)	17.7 (0.2)	255.8 (4.0)
5	4	211.2 (4.4)	23.7 (0.4)	16.6 (0.2)	18.2 (0.2)	16.7 (0.2)	16.6 (0.2)	259.6 (3.9)
6	4	214.1 (4.2)	23.2 (0.3)	16.6 (0.2)	18.1 (0.2)	16.9 (0.2)	16.6 (0.2)	256.8 (3.8)
7	4	99.9 (2.2)	17.5 (0.2)	13.0 (0.2)	12.4 (0.2)	12.7 (0.2)	12.8 (0.2)	150.8 (2.5)
8	4	128.4 (2.2)	18.0 (0.3)	13.5 (0.2)	12.9 (0.2)	12.9 (0.2)	12.9 (0.2)	161.1 (2.5)
9	4	131.8 (2.4)	19.6 (0.3)	14.7 (0.2)	13.8 (0.2)	14.0 (0.2)	14.0 (0.2)	176.5 (2.6)
1	8	148.0 (3.0)	12.1 (0.2)	7.5 (0.1)	8.7 (0.1)	7.8 (0.1)	7.5 (0.1)	175.4 (2.9)
2	8	151.5 (3.0)	13.4 (0.2)	8.3 (0.1)	9.0 (0.1)	8.5 (0.1)	8.1 (0.1)	180.5 (2.9)
3	8	151.5 (3.1)	17.7 (0.3)	11.5 (0.2)	12.5 (0.1)	11.2 (0.1)	11.1 (0.1)	183.5 (2.9)
4	8	153.9 (3.2)	17.7 (0.3)	11.4 (0.2)	12.2 (0.1)	11.3 (0.2)	11.1 (0.1)	178.3 (2.9)
5	8	150.9 (3.0)	16.7 (0.3)	10.9 (0.1)	12.3 (0.1)	10.9 (0.1)	10.9 (0.1)	181.2 (2.9)
6	8	154.3 (3.1)	16.7 (0.3)	10.8 (0.1)	12.2 (0.1)	10.9 (0.1)	10.9 (0.1)	183.5 (3.0)
7	8	78.4 (1.4)	13.9 (0.2)	9.2 (0.1)	8.0 (0.1)	8.6 (0.1)	8.1 (0.1)	108.4 (1.8)
8	8	92.5 (2.0)	14.0 (0.2)	9.5 (0.1)	8.5 (0.1)	8.8 (0.1)	8.5 (0.1)	117.5 (1.9)
9	8	91.2 (1.9)	15.0 (0.2)	9.9 (0.1)	8.9 (0.1)	9.4 (0.1)	9.0 (0.1)	128.3 (2.0)

I.3 From an audited curve to a population budget

Fix a generator, its decoding configuration, a response-wise judge, and the bounded outcome definition. Let independent, identically distributed audit-set pools be indexed by $i = 1, \ldots, M_{\mathrm{A}}$, with ranked outcomes Y_{ir}. For each budget, define

$$Z_i(k) = \sum_r a_{kr} Y_{ir}, \quad \theta_{\mathrm{A}}(k) = M_{\mathrm{A}}^{-1} \sum_i Z_i(k), \quad Q(k) = \mathbb{E} Z_i(k).$$

Each $Z_i(k)$ lies in $[0, 1]$. Dependence between ranks within one pool is permitted. Hoeffding's inequality and a union bound over budgets give

$$\mathbb{P}\{\|\theta_{\mathrm{A}} - Q\|_\infty > e_g\} \le \delta_g, \qquad e_g = \sqrt{\frac{\log(2L/\delta_g)}{2M_{\mathrm{A}}}}. \tag{I.1}$$

Suppose an audit satisfies $\mathbb{P}_{\mathrm{audit}}\{\|\widehat{\theta} - \theta_{\mathrm{A}}\|_\infty > e_a \mid \text{corpus}\} \le \delta_a$ for every fixed corpus. The radius e_a may depend on the corpus if this conditional guarantee holds. Integrating the audit failure probability and combining the two events yields $\|\widehat{\theta} - Q\|_\infty \le e_a + e_g$ with probability at least $1 - \delta_a - \delta_g$.

For a prespecified cost $c(k)$, let $\widehat{k}$ maximize $\widehat{\theta}(k) - c(k)$ within tolerance τ, and let k^* maximize $Q(k) - c(k)$. Adding and subtracting the two estimated utilities gives

$$Q(k^*) - c(k^*) - [Q(\widehat{k}) - c(\widehat{k})] \le 2(e_a + e_g) + \tau.$$

The middle estimated-utility difference is at most τ, and the two remaining differences are each bounded by $e_a + e_g$. The experiments use $\tau = 10^{-12}$. This bound separates the label-query error from the error due to sampling audit inputs.

For a new input on which deployment generates only k responses, $Q(k)$ is the expected quality of the selected response, provided that a uniform k-subset of the L-response generation law agrees with the size-k generation law and that selection is consistent under restriction. Conditionally iid responses and a fixed response-wise scoring rule are sufficient. For human-rated system menus (SummEval and WMT) and correctness-filtered collections (PPE), the target remains the curve of the collection itself.

I.4 Audit and test split protocol

The budget-selection study uses the nine configurations in Table 3. For every configuration, we draw 100 random splits of the inputs into audit and test sets and run 20 audits of each of Uniform, Target-mixture, Whole-pool, Harmonic, Grid-mixture, MMV, GW, and Envelope on every split. MATH500 uses 250 audit and 250 test inputs; GSM8K uses 659 and 660. Matched configurations use the same splits of their shared inputs, and all responses to an input remain on the same side of the split. Rank allocations share their audit randomization; Whole-pool draws complete pools within the same splits.

Distinct-label budgets are $2L$ and $8L$. For each $\rho \in \{0, .025, .05, .1, .2\}$, the audit chooses the smallest maximizer of $\widehat{\theta}_{\mathrm{A}}(k) - \rho(k-1)/(L-1)$ within tolerance 10^{-12}. The primary price $\rho = .1$ penalizes using all L candidates by ten percentage points of accuracy. Table 23 gives Envelope's regret at $n = 8L$ for every price. Equation (3) and the population argument above apply to any specified cost vector, including nonlinear costs.

Only labels from the audit set enter the decision, and complete test outcomes then give the expected quality of the chosen k over random k-subsets of the test pools. The full-label reference queries all $M_{\mathrm{A}}L$ audit-set labels to choose its budget. Table 22 reports the mean selected budget $\bar{k}$, which averages the integer choices over splits and replications, together with test quality, test utility regret R, and the full-label reference regret R_{full}. Their difference $\Delta = R - R_{\mathrm{full}}$ measures the audit's extra loss.

For each split,

$$R_{\mathrm{test}}(\widehat{k}) \le 2\|\widehat{\theta}_{\mathrm{A}} - \theta_{\mathrm{A}}\|_\infty + 2\|\theta_{\mathrm{A}} - \theta_{\mathrm{test}}\|_\infty + 10^{-12}.$$

For standard errors, we first average the 20 audits within each split and then measure Monte Carlo precision across splits, conditional on the released inputs. These retrospective comparisons assess budget selection on new inputs within each collection, and the assumptions in Appendix I.3 give them a population interpretation.

Table 21 compares all eight methods at both label budgets. The repository also reports each method's difference from Envelope, computed within each split before the Monte Carlo SE is calculated, and complete results at all five prices. Envelope and GW yield similar test utility, and Envelope has slightly lower mean regret than MMV at $8L$. Harmonic and Grid-mixture have lower mean regret than Envelope in all nine configurations at $8L$, by 0.006–0.017 and 0.007–0.021 percentage points, respectively. Regret evaluates the chosen budget; full-curve error also measures inaccuracies at other candidate counts.

Table 21: Mean test utility regret (percentage points) at $\rho = .1$, with Monte Carlo SEs in parentheses. All methods use the stated label budget on the same audit sets. IDs match Table 3; the full-label reference appears in Table 22.

ID	n/L	Uniform	Target-mixture	Whole-pool	Envelope
1	2	3.377 (0.099)	0.447 (0.019)	4.563 (0.115)	0.405 (0.020)
2	2	4.622 (0.123)	0.501 (0.020)	4.002 (0.100)	0.398 (0.018)
3	2	3.391 (0.097)	0.633 (0.029)	3.446 (0.094)	0.602 (0.028)
4	2	3.282 (0.089)	0.534 (0.025)	3.570 (0.101)	0.529 (0.025)
5	2	3.013 (0.084)	0.516 (0.025)	4.701 (0.122)	0.503 (0.026)
6	2	3.429 (0.094)	0.556 (0.019)	4.823 (0.124)	0.455 (0.019)
7	2	1.696 (0.052)	0.311 (0.007)	2.325 (0.072)	0.231 (0.008)
8	2	3.218 (0.088)	0.353 (0.011)	2.301 (0.071)	0.267 (0.008)
9	2	3.388 (0.079)	0.358 (0.013)	3.728 (0.087)	0.303 (0.014)
1	8	3.005 (0.109)	0.185 (0.020)	2.571 (0.102)	0.197 (0.019)
2	8	3.576 (0.128)	0.197 (0.016)	2.245 (0.083)	0.158 (0.014)
3	8	3.010 (0.104)	0.405 (0.025)	2.062 (0.078)	0.353 (0.027)
4	8	2.805 (0.081)	0.339 (0.028)	2.200 (0.067)	0.261 (0.023)
5	8	2.602 (0.076)	0.342 (0.035)	2.429 (0.101)	0.291 (0.026)
6	8	2.859 (0.099)	0.140 (0.015)	2.126 (0.076)	0.154 (0.014)
7	8	0.795 (0.030)	0.135 (0.007)	0.970 (0.043)	0.064 (0.004)
8	8	1.622 (0.076)	0.156 (0.005)	1.081 (0.041)	0.102 (0.005)
9	8	2.149 (0.078)	0.192 (0.013)	1.549 (0.051)	0.109 (0.009)

ID	n/L	Harmonic	Grid-mixture	MMV	GW	Envelope
1	2	0.402 (0.021)	0.404 (0.021)	0.426 (0.022)	0.405 (0.020)	0.405 (0.020)
2	2	0.396 (0.018)	0.397 (0.018)	0.395 (0.017)	0.401 (0.018)	0.398 (0.018)
3	2	0.582 (0.026)	0.596 (0.028)	0.608 (0.029)	0.576 (0.025)	0.602 (0.028)
4	2	0.538 (0.026)	0.537 (0.026)	0.551 (0.027)	0.525 (0.025)	0.529 (0.025)
5	2	0.518 (0.028)	0.520 (0.029)	0.516 (0.028)	0.513 (0.028)	0.503 (0.026)
6	2	0.451 (0.020)	0.456 (0.020)	0.455 (0.021)	0.453 (0.020)	0.455 (0.019)
7	2	0.211 (0.007)	0.219 (0.007)	0.223 (0.007)	0.237 (0.008)	0.231 (0.008)
8	2	0.249 (0.007)	0.245 (0.007)	0.255 (0.007)	0.269 (0.008)	0.267 (0.008)
9	2	0.278 (0.013)	0.292 (0.014)	0.294 (0.013)	0.307 (0.014)	0.303 (0.014)
1	8	0.187 (0.018)	0.184 (0.018)	0.225 (0.018)	0.197 (0.018)	0.197 (0.019)
2	8	0.149 (0.014)	0.146 (0.013)	0.169 (0.014)	0.161 (0.014)	0.158 (0.014)
3	8	0.347 (0.027)	0.340 (0.027)	0.376 (0.026)	0.353 (0.026)	0.353 (0.027)
4	8	0.244 (0.023)	0.240 (0.023)	0.281 (0.022)	0.259 (0.022)	0.261 (0.023)
5	8	0.285 (0.027)	0.280 (0.027)	0.307 (0.025)	0.290 (0.026)	0.291 (0.026)
6	8	0.144 (0.014)	0.144 (0.015)	0.170 (0.014)	0.152 (0.014)	0.154 (0.014)
7	8	0.058 (0.003)	0.056 (0.003)	0.066 (0.004)	0.063 (0.003)	0.064 (0.004)
8	8	0.096 (0.005)	0.095 (0.005)	0.103 (0.005)	0.101 (0.005)	0.102 (0.005)
9	8	0.098 (0.007)	0.098 (0.007)	0.109 (0.008)	0.108 (0.008)	0.109 (0.009)

Table 22: Envelope on test inputs, $\rho = .1$. Test quality is a percentage; R, R_{full}, and Δ are percentage points. Parentheses give Monte Carlo SEs for the paired difference. IDs match Table 3.

ID	n/L	$\bar{k}$	Test quality	R	R_{full}	Δ (MC SE)
1	2	147.5	43.31	0.405	0.151	0.254 (0.016)
1	8	137.8	43.47	0.197	0.151	0.046 (0.005)
2	2	41.6	55.40	0.398	0.092	0.306 (0.014)
2	8	28.7	55.58	0.158	0.092	0.066 (0.004)
3	2	52.9	51.95	0.602	0.274	0.328 (0.020)
3	8	36.5	52.03	0.353	0.274	0.079 (0.012)
4	2	74.8	53.66	0.529	0.140	0.389 (0.017)
4	8	63.5	53.82	0.261	0.140	0.121 (0.009)
5	2	89.2	41.63	0.503	0.218	0.285 (0.022)
5	8	75.5	41.71	0.291	0.218	0.073 (0.010)
6	2	100.6	42.99	0.455	0.102	0.352 (0.016)
6	8	92.1	43.21	0.154	0.102	0.052 (0.004)
7	2	60.9	93.03	0.231	0.019	0.212 (0.008)
7	8	54.1	93.13	0.064	0.019	0.045 (0.002)
8	2	40.0	90.67	0.267	0.039	0.228 (0.008)
8	8	29.5	90.73	0.102	0.039	0.063 (0.004)
9	2	64.3	89.55	0.303	0.049	0.254 (0.011)
9	8	56.8	89.67	0.109	0.049	0.060 (0.005)

Table 23: Envelope's mean test utility regret in percentage points at $n = 8L$ for each price. With $\rho = 0$, the budget is chosen for quality alone. All columns evaluate the same audits, each at its own price.

ID	$\rho = 0$	$\rho = .025$	$\rho = .05$	$\rho = .1$	$\rho = .2$
1	1.071	0.510	0.296	0.197	0.160
2	0.427	0.195	0.177	0.158	0.141
3	1.084	0.532	0.434	0.353	0.273
4	0.837	0.418	0.303	0.261	0.217
5	0.447	0.530	0.339	0.291	0.237
6	0.762	0.249	0.189	0.154	0.134
7	0.237	0.088	0.073	0.064	0.053
8	0.258	0.141	0.120	0.102	0.082
9	0.268	0.137	0.120	0.109	0.096

I.5 Outcome labels and score aggregation

We checked 400 released labels using 20 questions selected with a fixed random seed and five fixed candidate positions in each of the four RLHFlow response collections, one for each combination of task and generator. Two DeepSeek labels on GSM8K disagreed with the frozen Qwen checker: both responses gave the reference answer of 12 percent. We therefore rechecked all labels of both GSM8K generators against the canonical test answers and kept the released labels for the primary results. Rechecking changed 5,422 of 1,350,656 DeepSeek labels and 3,729 of 1,350,656 Mistral labels. Because the alternative checker also has parsing and timeout limitations, we regard these changes as differences between two labeling protocols. We repeated the audits of all three GSM8K configurations under the rechecked labels at both label budgets, $2L$ and $8L$ (Table 24). Envelope's mean curve error changes by at most 0.049 percentage points in these six comparisons. Evaluating both budget choices under the rechecked labels changes mean test utility by at most 0.013 percentage points at $\rho = .1$.

Table 24: GSM8K label sensitivity for Envelope. Curve errors are multiplied by 10^3. The last column evaluates both audited budget choices under rechecked test labels and gives their mean utility difference in percentage points at $\rho = .1$.

ID	Labels	Released-label error	Rechecked error	Budget utility change
7	2048	18.08	18.06	-0.0003
7	8192	8.14	8.14	0.0006
8	2048	19.57	19.08	0.0123
8	8192	8.46	8.29	0.0105
9	2048	20.62	20.77	0.0059
9	8192	8.97	9.04	0.0039

Table 25: PRM aggregation sensitivity for Mistral on MATH500. Distinct-label mean maximum errors ($\times 10^3$); both rules use the same checkpoint, responses, labels, and fixed tie convention.

Labels	Method	Mean-step PRM	Last-step PRM
2048	Envelope	28.25	28.70
2048	MMV	29.06	28.88
2048	GW	28.26	28.58
4096	Envelope	16.63	16.46
4096	MMV	18.10	18.09
4096	GW	16.92	16.61
8192	Envelope	10.86	10.68
8192	MMV	12.23	12.15
8192	GW	10.86	10.81

The PRM's step-score aggregation also changes the quality curve. For the same Mistral PRM, replacing the step mean with the last step changes Envelope's mean selected budget from 92.1 to 43.5 and its mean test quality from 43.21% to 39.95% at $n = 8L$, $\rho = .1$. These quantities therefore depend on the complete scoring rule as well as the reward-model checkpoint. Full-curve audit errors remain close under the two rules (Table 25).