How Far Can Shallow Evaluation See? The Limits of Best-of-K Prediction

Anonymous authors
Paper under double-blind review

Abstract

Can evaluations with few candidates per prompt predict Best-of-K success at larger budgets? Under a common rank-correctness response, complete shallow, score-ordered joint label distributions leave the same worst-case ambiguity as rank marginals alone. We characterize this gap exactly at every finite deployment budget and construct collisions attaining it. In a slope-bounded monotone response class, worst whole-curve ambiguity scales inversely with the square of evaluation depth; smooth monotonicity without a common slope bound can still permit nearly maximal ambiguity as deployment grows at fixed evaluation depth. With a common response and arbitrary unknown atomless group-specific score distributions of unrestricted complexity, numerical scores do not improve minimax risk. By contrast, one unknown conditional score distribution shared across prompts enables unbiased pooling despite heterogeneous responses. For declared small uniform departures from sharing, we derive risk bounds that match on the logarithmic scale. Within-prompt comparisons reduce mismatch bias, but noise limits the correction order finite data support. At equal total cost, two designs can both be consistent yet differ in whether they eventually meet a shrinking error tolerance. Together, these results separate irreducible information loss from finite-sample error and show why prediction range and consistency do not determine attainable precision.

1 Introduction

Best-of-K selects the highest-scoring candidate, not any correct answer in a sample. Larger budgets emphasize the upper score tail: measuring shallow behavior precisely need not establish deeper selection performance. Against this background (Dorner et al., 2026; Mukherjee et al., 2026), we ask which deeper performance claims shallow evidence justifies. This requires separating three questions: what do the records identify, what structure makes deeper behavior predictable, and what precision can finite data support?

We first remove sampling uncertainty entirely. Suppose the complete shallow distribution is known: the probability of every correctness pattern after score sorting, not just accuracy at each rank. Assume independent candidate sampling and a common *response*: correctness probability as a function of within-prompt score percentile. Even this complete joint information leaves a worst-case gap exactly as large as rank accuracies alone permit. We determine that gap at every finite deployment budget (Figure 1).

The difficulty is attaining the rank-based bound while matching every joint pattern. One alternating response construction does both: the same switches make all shallow pattern probabilities agree and maximize the deeper-performance difference. Thus the gap is not merely an example of failed extrapolation; it is the largest one compatible with the marginal constraints. Repetition at the same depth cannot separate this pair.

Response structure changes what this ambiguity means. We characterize all weighted response averages determined class-wide by shallow records. Smooth monotonicity alone does not enlarge this set and can still leave nearly maximal ambiguity as deployment grows. A common positive slope cap at most one instead makes whole-curve ambiguity proportional to the cap and inverse-square in evaluation depth. Quantitative regularity can therefore improve approximation without enlarging the class-wide identifiable set.

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

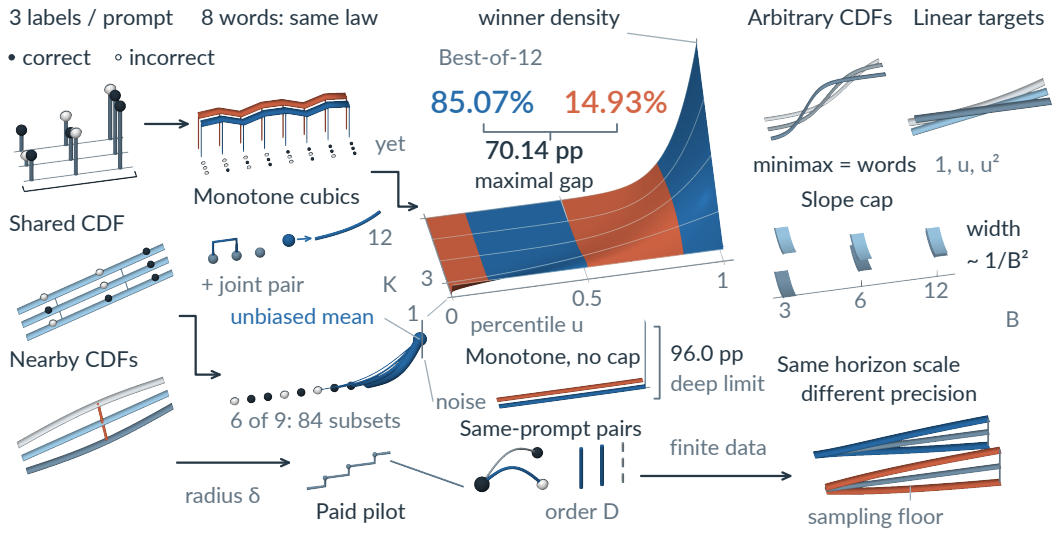


Figure 1: **From shallow agreement to conditional recovery.** The central example matches every three-label word probability while attaining the maximal Best-of-12 gap. Other components show distinct assumptions, not stages of one method: a common slope bound controls ambiguity, a monotone cubic restriction makes a joint count decisive, and score sharing permits pooling. The thin-tail gap is a deep limit; ribbons and other graphical devices are schematic, not empirical curves or numerical risk bounds.

This does not make joint labels dispensable: for monotone cubics at depth three, adding one joint-label count to the final rank means enables consistent recovery of the deeper curve; the means alone do not.

A different route changes score comparability rather than response shape. With a common response and arbitrary unknown continuous fixed-cohort score distributions of unrestricted complexity, numerical scores give no smaller minimax risk than ordered labels. Under a separate assumption—one unknown conditional score distribution shared across prompts—pooling is unbiased for the original deployment target despite heterogeneous responses, up to the pooled sample size. The distinction is not whether scores are recorded, but whether the model supplies the comparability needed to combine them.

Recovery being possible does not settle its attainable precision. For declared small uniform departures from sharing, we characterize risk on the logarithmic scale. Within-prompt comparisons estimate mismatch corrections without true percentiles, but noise limits the order the data support. Two equal-cost designs can consequently both be consistent, yet only one eventually meets a shrinking error tolerance. Prediction range and consistency alone therefore do not determine whether a precision requirement can be met.

2 Related Work

Performance prediction and identification. Dorner et al. (2026) express Best-of-N accuracy through initial accuracy and the verifier's receiver operating characteristic. Their conference-version Proposition 6 gives arbitrarily close finite-budget curves with opposite infinite-budget limits. Our question instead fixes a finite deployment target and requires exactly identical complete shallow records: how much can the target still differ? Mukherjee et al. (2026) connect generator coverage and verifier errors to sampling-policy sub-optimality. Gaussian reward-tail forecasting (Li et al., 2026) predicts maximum reward rather than selected-answer correctness.

From collision existence to exact ambiguity. In random utility models over a fixed item universe, Chierichetti et al. (2024) establish a square-root slate-size requirement for constant-error prediction.

Their Theorem 14 constructs binary-string distributions with identical expected subsequence counts but sharply different first-bit probabilities. Thus finite-order agreement can already conceal different choices. In our independently sampled percentile model, two fixed responses match every shallow ordered-label probability and attain the maximal gap at every finite deployment budget. Our construction couples classical ordered cancellation (Yang & Liu, 2008) with extremal interpolation (Pinkus, 1989): the same switches satisfy the joint constraints and attain the approximation bound.

Structure, exact identification and approximation. Word-limit theory supplies the observation law and finite identification of particular piecewise-polynomial responses (Hàn et al., 2021); path-signature inversion reconstructs monotone paths as depth grows (Chang et al., 2017). We instead fix depth and characterize linear response summaries identifiable throughout a class, while allowing individual responses to be uniquely identified. Shape can reduce the remaining ambiguity without changing this class-wide identification boundary. The cubic example asks a separate question: what recovery is lost by retaining only final rank means rather than complete joint records?

Numerical recovery and finite-data precision. Our minimax reduction builds on labeled-order Ramsey reductions (Diakonikolas et al., 2024), retaining randomized estimates and cross-cohort comparisons under expected loss. Established subset-winner weights give within-prompt unbiasedness even with distinct selection and evaluation scores (Nakano et al., 2021, Appendix I). We justify their cross-prompt use for the original deployment target with heterogeneous responses and quantify prompt dependence in the risk. Under approximate sharing, functional-estimation duality (Polyanskiy & Wu, 2021) supplies the lower-bound route. Recovery must additionally estimate an unknown score distribution with a paid pilot and control dependent within-prompt comparisons: these statistical costs limit the usable correction order and hence attainable precision.

3 Exact limits of shallow records

3.1 Deployment target and complete ordered labels

Given a prompt $X = x$ on a standard Borel space with regular conditional laws, candidates (S_i, G_i) are iid, with score S_i and target $G_i \in [0,1]$. Independent uniform keys break ties. For K fixed before observing data, let I_K select the maximum and $\theta_K = \mathbb{E}G_{I_K}$. With conditional CDF F_x and independent $V \sim \mathrm{Unif}[0,1]$, the randomized percentile $U = F_x(S^-) + V\{F_x(S) - F_x(S^-)\}$ is conditionally uniform. Set $\eta_x(u) = \mathbb{E}[G \mid U = u, X = x]$ and $\bar{\eta} = \mathbb{E}_X \eta_X$. A candidate at percentile u wins against $K-1$ others with probability u^{K-1}, giving

$$\theta_K = \int_0^1 f_K(u)\bar{\eta}(u)\,du, \qquad f_K(u) = Ku^{K-1}. \tag{1}$$

Appendix A.1 includes atoms. The word and numerical models below use binary labels and atomless scores; true percentiles remain unobserved. The homogeneous word experiment has one common measurable response η, iid uniform latent ranks and conditionally independent Bernoulli labels:

$$p_B(\eta)_w = B! \int_{0<u_1<\cdots<u_B<1} \prod_{i=1}^{B} \eta(u_i)^{w_i}(1-\eta(u_i))^{1-w_i}\,du_{1:B}, \quad w \in \{0,1\}^B. \tag{2}$$

This is the established word-limit law and factorial-scaled path signature (Hàn et al., 2021; Chang et al., 2017). Write dist_1 for $L^1[0,1]$ distance, $\mathcal{P}_d$ for degree-at-most-d polynomials, U_B for the Chebyshev polynomial of the second kind, and $\Omega^{\mathrm{sig}}_{B,K} = \sup_{p_B(\eta)=p_B(\widetilde{\eta})} |\theta_K(\eta) - \theta_K(\widetilde{\eta})|$.

Theorem 3.1 (Exact word ambiguity)**.** *For all integers $B, K \geq 1$,*

$$\Omega^{\mathrm{sig}}_{B,K} = \Delta_{B,K} := \mathrm{dist}_1(Ku^{K-1}, \mathcal{P}_{B-1}). \tag{3}$$

For $K > B$, $h_B = \operatorname{sgn} U_B(2u-1)$ and $\eta_t = 1/2 + th_B$, $|t| \leq 1/2$, obey

$$p_m(\eta_t) = p_m(\eta_{-t})\ (m \leq B), \qquad |\theta_K(\eta_t) - \theta_K(\eta_{-t})| = 2|t|\Delta_{B,K}. \tag{4}$$

The exact-record minimax radius is $\Delta_{B,K}/2$, with fixed-error horizon $\Theta_\epsilon(B^2)$ for $0 < \epsilon < 1/2$.

One pair: word cancellation and extremality. Rank densities span $\mathcal{P}_{B-1}$, so equal words imply equal rank means and the approximation bound $\Delta_{B,K}$. Attainment requires more. Expanding the word difference at $1/2 \pm t h_B$ leaves odd numbers r of h_B factors; integrating other positions gives a polynomial q of total degree at most $B - r$. Ordered cancellation (Yang & Liu, 2008, Eqs. (9)–(10)) gives

$$\int_{0<u_1<\cdots<u_r<1} q(u_1,\ldots,u_r) \prod_{i=1}^{r} h_B(u_i)\, du_{1:r} = 0 \quad (r \text{ odd},\ \deg q \le B - r).$$

Thus every word agrees. For $K > B$, interpolation of Ku^{K-1} at the same switches has residual sign h_B by its positive Bth derivative. Orthogonality to $\mathcal{P}_{B-1}$ certifies L^1 extremality (Pinkus, 1989): classical cancellation and interpolation jointly attain the full-word bound. Appendix A.2 proves both steps and gives the finite-target sum and horizon (Figure 2).

Exact saturation is target-specific. For a general kernel $g \in L^1$, a word collision attains $\mathrm{dist}_1(g, \mathcal{P}_{B-1})$ exactly when it signs a best approximation residual; even positive densities can fail this criterion (Appendix A.5.2).

3.2 Finite replication and numerical scores

For a model $\mathcal{H}$, define the expected absolute risk $R_{\mathcal{H}}(K) = \inf_{\widehat{\theta}_K} \sup_{P\in\mathcal{H}} \mathbb{E}_P|\widehat{\theta}_K - \theta_K(P)|$, allowing randomized estimators. The fixed-cohort experiment has g known groups of b candidates, $S_{aj} = F_a^{-1}(U_{aj})$, iid uniform ranks and one common Bernoulli response η. The unknown atomless F_a may be unrelated. Let $R^{\mathrm{num}}_{b,g}$ retain all scores, labels, identities and cross-cohort comparisons, and $R^{\mathrm{word}}_{b,g}$ retain only ordered words.

Theorem 3.2 (Numerical-data minimax reduction)**.** *For all finite integers $b, g, K \ge 1$,*

$$R^{\mathrm{num}}_{b,g}(K) = R^{\mathrm{word}}_{b,g}(K). \tag{5}$$

Equality also holds for any fixed finite target list under expected maximum absolute loss and arbitrary randomized estimators. CDFs may be restricted to smooth positive densities on $(0,1)$, without fixed complexity or derivative bounds. Both conclusions hold with responses restricted to any nonempty class having uniformly bounded total variation.

For unrestricted responses and $K > b$, finite replication gives

$$c_1 e^{-C_1 Q/K} \le R^{\mathrm{num}}_{b,g}(K) = R^{\mathrm{word}}_{b,g}(K) \le C_2 e^{-c_2 Q/K}, \qquad Q = b\min\{b, \log(1+g)\}, \tag{6}$$

with universal positive constants. At fixed b, K, both risks tend to $\Delta_{b,K}/2$ as $g \to \infty$.

Numerical reduction and finite replication. Sorting gives one inequality; conversely, label-aware product Ramsey selects score cells with nearly constant quantized output laws for every word and cohort tuple (Diakonikolas et al., 2015; 2024; Correa et al., 2019). The CDFs and induced word rule precede the response supremum; cell averaging and smoothing remove the error (Appendix A.4). For (6), squared polynomial projection on the last $2r+1$ rank densities reduces bias with coefficient norm at most $(3/2)9^r$, restricting r to order $\min\{b, \log g\}$. An order-r collision in a tail of mass $r/(5b)$ requires more than r tail observations to be distinguished (Appendix A.3).

Minimax equality is not experiment equivalence or pointwise irrelevance of scores. Unrestricted cohorts may bring new CDFs; repeated observations from a fixed small population can instead support calibration. Linear cost imposes no depth cap: available depth-K acquisition must enter resource comparisons.

Two separate structural routes follow: restrict responses while retaining ordered labels (Section 4), or allow heterogeneous responses but constrain cross-prompt score distributions (Section 5).

4 What response structure changes

We first keep the ordered-label experiment and ask whether its limitations depend on the oscillatory witnesses in Theorem 3.1. Smooth monotonicity leaves the identifiable linear targets unchanged;

a common slope cap instead controls the remaining ambiguity. For $0 < \kappa < 1/2$, $L > 0$, let $\mathcal{H}_{\kappa,L} = \{\eta \in C^\infty[0,1] : \kappa < \eta < 1-\kappa,\ 0 < \eta' < L\}$ and let $\mathcal{H}_\uparrow$ contain all nondecreasing $[0,1]$-valued responses. For $g \in L^1$, set $T_g(\eta) = \int g\eta$. The space $\mathcal{O}_B^{\mathrm{hom}}(\mathcal{H})$ consists of kernels whose targets are constant on every fibre of p_B in $\mathcal{H}$. For populations supported on $\mathcal{H}$, $\mathcal{O}_B^{\mathrm{mix}}$ instead requires $\mathbb{E}_\nu p_B = \mathbb{E}_{\nu'} p_B$ to imply $\mathbb{E}_\nu T_g = \mathbb{E}_{\nu'} T_g$.

Theorem 4.1 (Shape-stable classification)**.** *For every $B \geq 1$, $0 < \kappa < 1/2$ and $L > 0$,*

$$\mathcal{O}_B^{\mathrm{hom}}(\mathcal{H}) = \mathcal{O}_B^{\mathrm{mix}}(\mathcal{H}) = \mathcal{P}_{B-1}, \qquad \mathcal{H} \in \{\mathcal{H}_{\kappa,L}, \mathcal{H}_\uparrow\}, \tag{7}$$

in L^1 modulo equality almost everywhere. The unrestricted spaces are also $\mathcal{P}_{B-1}$. Every $g \notin \mathcal{P}_{B-1}$ admits an exact word collision in every fixed $\mathcal{H}_{\kappa,L}$ with unequal T_g, even with arbitrary nonlinear decoding.

Shape-preserving collisions. At a maximal-rank smooth response with strict shape slack, implicit-function compensation fixes independent word coordinates along an invisible direction; maximal rank fixes the rest. Using exponential responses with two incommensurate rate lattices restricts identifiable kernels to $\mathcal{P}_{B-1}$ (Appendix A.5). This is class-wide: individual fibres can be singletons, as for constant responses at depth three (Hàn et al., 2021, Remark 26).

4.1 Quantitative regularity and whole-curve ambiguity

Let $\mathcal{M}_L = \{0 \leq \eta \leq 1 : \eta \text{ absolutely continuous},\ 0 \leq \eta' \leq L\}$, $0 < L \leq 1$. Define $W_B^{\mathrm{word,curve}}(L)$ as the largest $\sup_{K \geq B} |\theta_K(\eta) - \theta_K(\widetilde{\eta})|$ over equal-word pairs in $\mathcal{M}_L$; $W_B^{\mathrm{mar,curve}}$ uses equal rank means, and $W_{B,K}^{\mathrm{mar}}(L)$ is its fixed-K counterpart.

Theorem 4.2 (A quantitative shape boundary)**.** *For every $B \geq 1$, $0 < L \leq 1$ and a universal $c > 0$,*

$$cL/B^2 \leq W_B^{\mathrm{word,curve}}(L) \leq W_B^{\mathrm{mar,curve}}(L) = L\lambda_B, \qquad (B+1)^{-2} \leq \lambda_B \leq 4/B^2, \tag{8}$$

where $\lambda_B = \min\{\|P\|_1 : P \in \mathcal{P}_B, P(0) = 0, P(1) = 1\}$. One fixed response pair attains every marginal target width for $K \geq B$, which increases to $L\lambda_B$. It need not match full words. The lower order holds for smooth strictly increasing responses in $(1/8, 7/8)$ with slope below L, and for exact mixed-word laws of populations supported on $\mathcal{M}_L$. In contrast, for $\mathcal{I}_\kappa = \{\eta \in C^\infty[0,1] : \kappa < \eta < 1-\kappa, \eta' > 0\}$, $0 \leq \kappa < 1/2$,

$$\lim_{K\to\infty} \sup_{\substack{\eta,\widetilde{\eta} \in \mathcal{I}_\kappa \\ p_B(\eta) = p_B(\widetilde{\eta})}} |\theta_K(\eta) - \theta_K(\widetilde{\eta})| = 1 - 2\kappa \qquad (B \textit{ fixed}). \tag{9}$$

Integration preserves cancellation under a slope cap. For $Jh(u) = \int_0^u h$, the pair $1/4 + Lu/2 \pm (L/4)Jh_{2B}$ has slopes in $[L/4, 3L/4]$. Expanding primitives reduces word differences to depth-$2B$ cancellation, with gap $L\Delta_{2B,K+1}/[2(K+1)]$, of order L/B^2 at $K = 2(2B+1)^2 - 1$. Integration by parts gives $W_{B,K}^{\mathrm{mar}}(L) = L \inf_{P \in \mathcal{P}_B :\, P(0)=0,\, P(1)=1} \|u^K - P\|_1$. Canonical interpolation attains these marginal widths with one pair (Pinkus, 1989; Dryanov & Petrov, 2012), not necessarily matching words. Its depth-two instance gives a finite-sample guarantee from empirical rank means, not pointwise method dominance (Appendix A.6.4). A thin-tail ramp with interior compensation proves the no-cap boundary, with no B-uniform threshold (Appendix A.6).

4.2 Retaining joint labels can determine consistency

The unrestricted width equality does not justify discarding dependence under response restrictions. In the full monotone class, the complete-word width at $(B, K) = (4, 16)$ is strictly between zero and the marginal benchmark (Appendix A.7). A complementary question is whether retaining joint labels can determine statistical consistency. For common responses in $\mathcal{C}_3 = \{\eta \in \mathcal{P}_3 : 0 \leq \eta \leq 1, \eta' \geq 0\}$, compare M iid depth-three words with only their final mean vector $\bar{W}$, not running or block means. Both cost $3M$ candidates and labels. Use minimax expected $\sup_{4 \leq k \leq T} |\widehat{\theta}_k - \theta_k|$ for $T \in \{4, 5, \ldots\} \cup \{\infty\}$. Let $\mathcal{C}_{3,\gamma}$ impose $c_1 = 3\int_0^1 u(1-u)\eta'(u)\,du \geq \gamma$, $0 < \gamma \leq 1/4$. Theorem C.1 gives

$$R_{M,T}^{\mathrm{word}}(\mathcal{C}_3) \asymp M^{-1/4}, \qquad R_{M,T}^{\mathrm{word}}(\mathcal{C}_{3,\gamma}) \asymp_\gamma M^{-1/2}, \qquad \liminf_M R_{M,T}^{\mathrm{mean}}(\mathcal{C}_{3,\gamma}) > 0. \tag{10}$$

Three means plus $\widehat{A} = M^{-1}\sum_i W_1^{(i)}W_2^{(i)}$ attain both full-word rates, with constants independent of M, T.

The three rank means recover the first three shifted-Legendre coefficients. A joint moment determines the remaining cubic coefficient through a quadratic identity, with monotonicity selecting its root. The inverse's square-root singularity explains the two rates; bounded target coefficients make coefficient error control the whole curve.

The converse must compare the laws of the actual empirical means, not just their expectations. Two responses share means but not covariances; a lattice local limit with tail control bounds the limiting total variation strictly below one. Their target gap then leaves a positive risk floor. Appendix C gives the inverse and proves both rates and the empirical-mean lower bound. This concerns cubics, not general monotonicity or sufficiency of four counters; depth-four means already identify cubics.

5 Recovery and precision under score sharing

We return to the numerical-score question of Theorem 3.2, now studying comparability rather than response regularity. Instead of a common response with unrelated score CDFs, this model allows heterogeneous responses and constrains the CDFs (Table 1). There are M iid prompt draws, with observed states on a standard Borel space and unknown law, and B conditionally iid binary-labeled candidates per draw, $N = MB$ in total. Responses η_x are arbitrary; exact sharing requires one unknown atomless score CDF F. Deployment draws all K candidates at one prompt. Every candidate and label, including a pilot, is charged.

Table 1: Different commonality assumptions. Both numerical experiments retain scores, labels and identities; conditional percentiles are unobserved.

Experiment	Response	Conditional score CDFs
Fixed cohorts	Common η	Arbitrary unrelated F_a
Iid prompts	Arbitrary η_x	Shared F, or within declared δ of unknown F_0

5.1 When cross-prompt pooling preserves the target

Sort all N scores and retain $G_{(j)}$. For $K \le N$, set

$$\widehat{\theta}_K = \sum_{j=K}^{N} w_{j,K} G_{(j)}, \qquad w_{j,K} = \frac{\binom{j-1}{K-1}}{\binom{N}{K}}. \tag{11}$$

These established subset-winner weights average all K-subsets with one sort and a weighted sum (Nakano et al., 2021, Appendix I). Their within-prompt unbiasedness does not itself justify crossing prompts.

Theorem 5.1 (Shared unknown CDF, heterogeneous prompts)**.** *For $K \le N$,* (11) *is unbiased. With $a_{N,K} = \sum_j w_{j,K}^2$ and $c_{M,B} = (B-1)/(N-1)$ for $N > 1$,*

$$\mathrm{Var}(\widehat{\theta}_K) \le \frac{K}{4N} + \frac{c_{M,B}}{4}(1 - a_{N,K}) \le \frac{1}{4M} + \frac{K}{4MB}. \tag{12}$$

For $N = 1$, variance is at most $1/4$. Including the unknown prompt law, for every $K \ge 1$,

$$R^{\mathrm{shared}}(B, M, K) \asymp \min\{1, \sqrt{M^{-1} + K/(MB)}\}. \tag{13}$$

If the response is also common, risk is $\asymp \min\{1, \sqrt{K/N}\}$ and variance is at most $K/(4N)$ for $K \le N$. For $K > N$, use the constant $1/2$. These guarantees are pointwise in K.

Why heterogeneity changes variance, not unbiasedness. Common F makes the percentiles iid uniform independently of prompt states. Conditional on a subset's ranks, its winner has mean $\bar{\eta}(\max U)$, even when prompt blocks repeat; integrating gives the original target. For second moments, let Q_r be ordered percentiles, $C_\eta(u, v) = \mathrm{Cov}_X(\eta_X(u), \eta_X(v))$, and T_{iid} the same statistic

with iid Bernoulli response $\bar{\eta}$. Put $w_r = w_{r,K}$, zero below K. Uniform assignment of ranks to blocks gives exactly

$$\mathrm{Var}(\widehat{\theta}_K) - \mathrm{Var}(T_{\mathrm{iid}}) = 2c_{M,B} \sum_{r<s} w_r w_s \mathbb{E} C_\eta(Q_r, Q_s). \tag{14}$$

The iid U-statistic has variance at most $K/(4N)$ (Hoeffding, 1948); $C_\eta \le 1/4$ gives (12). Rare-tail responses and observed prompt difficulty separately give the matching floors (Appendix B.1). The covariance need not be positive.

This distinguishes performance from precision. With fair $X \in \{-1, 1\}$ and $\eta_X(u) = .3 + .4u + aX$, $0 \le a \le .3$, the marginal score–label law and all deployment targets are independent of a. Yet variance increases by $a^2 c_{M,B}(1 - a_{N,K})$. Figure 5 tests this prediction, with simultaneous Monte Carlo intervals (Appendix D.2.3).

When pooling changes the target. For $H = \mathbb{E}F_X$, mixed-prompt selection has worst bias over responses $\Xi_K = (K/2)\mathbb{E}|H(S)^{K-1} - F_X(S)^{K-1}|$. For $K \ge 2$ it vanishes exactly under sharing. The classical average discrepancy $\mathcal{D}_{\mathrm{cal}} = \mathbb{E}\int (F_X - H)^2 dH$ (Chatterjee, 2021) bounds this bias locally sharply. Along K-dependent score families, it can tend to zero with nonvanishing bias. Repeated-score comparisons identify the average, not a uniform radius (Proposition B.1). A lower confidence bound can refute a proposed radius, not validate it.

5.2 Approximate sharing and the error-uniform risk

Let $\mathfrak{P}_\delta$ contain the iid-prompt laws with $\operatorname{ess\,sup}_x \|F_x - F_0\|_\infty \le \delta$ for an unknown continuous F_0. Write R_δ for full numerical-data risk in this class. The radius is declared, not inferred from an agreement test; neither F_0 nor conditional percentiles are observed. For $K \le MB$, numerical pooling obeys

$$\mathbb{E}|\widehat{\theta}_K - \theta_K| \le \tfrac{1}{2}\sqrt{M^{-1} + K/(MB)} + \min\{1, 2K\delta\}. \tag{15}$$

First-order bias is not an information lower bound: within-prompt comparisons can estimate corrections, but their noise limits the usable comparison order even at large depth. Define

$$r_* \log(1 + r_*/(B\delta)) = \log(eM), \qquad D = \min\{B, \max\{1, r_*\}\}, \tag{16}$$

$E_* = D\log(1 + D/(K\delta))$ and $\rho = \min\{1, \sqrt{M^{-1} + K/(MB)}\}$. Here D is the effective order supported by the data, potentially much smaller than depth B; E_* governs mismatch resolution and ρ is the sampling floor.

Theorem 5.2 (Error-uniform logarithmic risk)**.** *There are universal $c, C > 0$ such that, for all integers $B, M \ge 1$, $K > B$ and $0 < \delta \le 1/4$,*

$$c\max\{\rho, e^{-CE_*}\} \le R_\delta(B, M, K) \le \min\{1/2, C[\rho + e^{-cE_*}]\}. \tag{17}$$

Equivalently, $\log(1/R_\delta) \asymp 1 + \min\{\log(1/\rho), E_*\}$ with universal constants. At $\delta = 0$, $R_0 \asymp \rho$; for $1 \le K \le B$, $R_\delta \asymp M^{-1/2}$. These are pointwise guarantees.

The sampling floor and data-supported order jointly constrain precision. Total sample size and the infinite-replication ceiling alone need not determine finite-data prediction range (Section 5.4). The match is logarithmic, not $R_\delta \asymp \rho + e^{-E_*}$; it provides neither calibrated confidence intervals nor a universally recommended correction order.

5.3 Why finite data limit the correction order

The challenge is to estimate mismatch without conditional percentiles while controlling dependent comparisons. The construction establishes rate attainability, not a fixed-order recommendation.

Recovering mismatch from observed comparisons. Split prompts into a paid pilot and n analysis draws. The pilot estimates $H = \mathbb{E}F_X$. Given the pilot and a labeled anchor at score s from prompt X, products of ℓ distinct same-prompt comparisons $\mathbf{1}\{S_q \le s\} - \widehat{H}(s)$ have mean $(F_X(s) - \widehat{H}(s))^\ell$: mismatch powers are estimable without true ranks. For larger mismatch, gate estimation to the winning tail. Equation (76) specifies a pilot threshold $\widehat{c}$ and deterministic tail bound

A. For a best degree-$(d-1)$ $L^1[1-A,1]$ approximant $P(1-Ax)=\sum_{r<d} a_r x^r$ to f_K, $1\le d\le B$, let $C_b=\#\{q\neq b: S_q>S_b\}$. Use

$$Z_P=\frac{1}{B}\sum_b G_b\mathbf{1}\{S_b>\widehat{c}\}\sum_{r<d}\frac{a_r}{A^r}\frac{\binom{C_b}{r}}{\binom{B-1}{r}}. \tag{18}$$

Given its anchor, the binomial ratios make the inner mean $P(F_X(S_b))$. The coefficients are computable without CDFs; average over the n analysis prompts and clip only the final average.

Dependence limits usable order. For $k=r+1$, two independently uniform size-k subsets of a pool, overlapping in Z observations, have union-tail probability at most A^{2k-Z} on the good pilot event, conditional on the prompt. The bound $\mathbb{E}A^{-Z}\le(1+k/(BA))^k$ controls variance without assuming independent anchors. In the non-dense coarse branch, $A\asymp\delta$; balancing the resulting $O(d\log(1+d/(B\delta)))$ amplification against $\log M$ gives D, and the approximation residual gives E_*. Centered comparisons cover smaller $K\delta/D$; dense and bounded regimes use rank-additive estimation, pooling or the midpoint. Appendices B.5–B.6 supply the normalization, pilot-failure control and complete case split, including certificate (81). Section 6 tests related mechanisms and finite-order implementations, not this full construction.

The converse keeps scores and identities. In a two-band submodel, within-band positions are uniform and parameter-free: given a prompt's band count, numerical values add no information. Moment matching and functional-estimation duality (Polyanskiy & Wu, 2021) give the gap. Assign the two priors to a sufficiently large finite set of named prompts and freeze it before sampling. Repeated identities and target fluctuation have vanishing cost at each fixed experiment; identities remain observed and known prompts keep their types. Optimizing concealment order gives e^{-CE_*}; exact sharing gives ρ (Appendices B.3–B.6).

5.4 Why prediction range does not determine precision

The risk bound answers more than how far one can predict. We first extract its fixed-error horizon, then separate that horizon and mere consistency from attainable precision. Define the pointwise-retuned horizon $\mathcal{C}_\epsilon=\sup\big(\{0\}\cup\{T\in\mathbb{N}:\max_{1\le k\le T}R_\delta(B,M,k)\le\epsilon\}\big)$. For $0<\epsilon<1/8$, once $M\ge M_\epsilon$, uniformly in $B\ge1$ and $0<\delta\le1/4$,

$$\mathcal{C}_\epsilon=\Theta_\epsilon(\min\{MB,D^2/\delta\}). \tag{19}$$

At $\delta=0$ the scale is MB. Different targets may use different estimators; this is not simultaneous coverage of one curve. For $M=B^2,\delta=B^{-1}$, the horizon is $\Theta_\epsilon[B(\log B/\log\log B)^2]$, although both the sample count and infinite-replication ceiling have scale B^3.

A correct horizon still does not encode smaller-error precision. For $B=n^2,\delta=n^{-2},M=(n+1)^n$ and $K=\lfloor n^3/\log n\rfloor$, $D\sim n$. Writing $Z=\min\{MB,D^2/\delta\}$ gives $Z/K\asymp n\log n$ but $R_\delta=\exp\{-\Theta(n\log\log n)\}$, since $E_*\sim n\log\log n$ while $\log(1/\rho)\sim n\log n/2$. Thus replacing the residual by $e^{-\Theta(Z/K)}$ contradicts the lower bound. For any positive threshold s, the exact inversion is $E_*\ge s\iff K\le D/\{\delta[\exp(s/D)-1]\}$. For error ϵ, necessary and sufficient s are different constant multiples of $\log(1/\epsilon)$; the sampling floor must also be small (Appendix B.7).

Consistency likewise loses a precision distinction. For each fixed integer $B\ge1$, all $M\ge1$, $K\ge2B+2$ and $0\le K\delta\le1/16$, Theorem B.6 gives

$$R_\delta(B,M,K)\asymp_B\min\left\{1,\min_{1\le j\le B}\left[(K\delta)^j+\sqrt{K^j/M}\right]\right\}. \tag{20}$$

Unlike (17), these constants may depend on B. At equal cost $N=2K^3$ and $\delta=K^{-5/4}$, this yields $R_b:=R_\delta(b,N/b,K)\asymp K^{-b/4}$ for $b=1,2$. Both vanish, but only depth two eventually meets $K^{-3/8}$. The local condition requires $K\ge65536$, hence $N\ge2^{49}$; uninstantiated constants may delay the crossing further. A separate declared three-band family at $K=8,\delta=1/128,N=2^{20}$ has finite prior-averaged threshold-decision bounds: balanced error above .4999 for every depth-one rule, but below .021 on either side for one depth-two counting rule, with both designs knowing the families (Appendix D.2.7). Neither is a saving over unrestricted depth-K acquisition.

6 EMPIRICAL EVIDENCE AND APPLICABILITY

The studies test mechanisms and applicability, not the full minimax construction. Figure 5 summarizes controlled evidence; Appendix D gives protocols and uncertainty.

Shallow evidence and deeper rankings. Across 17 scorers on the label-filtered sets of Frick et al. (2025), pairwise accuracy tracks Best-of-16 (Kendall $\tau = .71$–1). A pooled-pair/four-rank-mean bootstrap screen selects no unconstrained ordering; it has no coverage guarantee and excludes complete depth-three/four words. Monotonicity narrows uncertainty-widened plug-in intervals, but the depth-eight test detects a $.05$ top-tail drop in only a quarter of simulations. Non-rejection cannot license shape (Appendix D.1).

Recovery under controlled structure. On the cubic at $M = 2^{24}$, the joint inverse has mean grid-maximum error $.0205$ versus $.125$ for the three-mean quadratic fit, which wins at small samples and on constant/affine controls. Under exact sharing, pooled error follows the $M^{-1/2}$ trend across five heterogeneous populations; simple rules win the bit control. At a fixed deployment curve, covariance predictions lie within simultaneous Monte Carlo intervals for the variance changes. Known-marginal-CDF calculations show successive bias cancellation, but pooling beats the best tested correction order in measured RMSE in 129/140 local-regime cells. The grid does not isolate D from depth (Appendix D.2).

Sharing assumptions on natural scores. Qwen/Mistral MATH-500 banks use disjoint protocol/reference pools (Hendrycks et al., 2021; Lightman et al., 2024; Yang et al., 2024; Jiang et al., 2023; Zhang et al., 2025). Sharing-radius lower bounds exclude the analyzed regime under nine scorers (Appendix D.3.5). Paid prompt-specific calibration improves pooling in 12/12 contrasts with raw depth-four or equal-cost raw pooling across both systems. Reassigned, common and random-map controls support prompt-specific information, not a unique fitted-CDF benefit. Direct evaluation and rank fits are more accurate in all six positive-calibration cells; calibration does not establish sharing (Appendix D.3.3).

Prospective applicability test. For 128 further problems, depth-four predictions (512 candidates and labels) precede 64 reference candidates per prompt. Sharing-radius lower confidence bounds exclude the regime of Theorem 5.2; the ungated correction also differs from its full construction (Appendix D.3.7). Clipped process-scored pooling/order-two predictions have absolute Best-of-16 discrepancies $.19864/.80136$, versus $.02011/.00057$ for plateau/inverse-budget extrapolation (Table 21). Plateau extends the depth-four winner mean; inverse-budget extrapolation fits depth-one/four means linearly in $1/K$. Second-verifier plateau/inverse discrepancies are $.03339/.03143$. Reference halves differ by $.02480$; point discrepancies do not certify accuracy.

7 DISCUSSION AND LIMITATIONS

We characterize the maximal disagreement in deeper Best-of-K performance compatible with complete shallow ordered-label agreement. Quantitative response regularity and score comparability change these limits in distinct models, separating exact identification from useful approximation. Finite-data recovery introduces a further constraint through the correction order the data support: consistency and prediction range alone do not determine attainable precision.

The analysis concerns fixed-budget selection from conditionally iid candidates. Numerical-risk equality is a worst-case statement under a common response and unrestricted unknown continuous score distributions, not a claim that scores are uninformative in every system. Recovery under score sharing allows heterogeneous responses but assumes a declared small uniform bound on departures from sharing; using a common scorer does not establish this bound. These guarantees apply separately to each deployment budget, and the logarithmic risk bounds are not calibrated confidence intervals. The natural-score studies fall outside the analyzed sharing regime and favor simple extrapolation or direct evaluation over correction. The contribution is a basis for interpreting deeper performance claims through the information retained, the structure assumed and the precision required. More samples can reduce estimation error; resolving an exact collision requires richer observations or stronger assumptions.

AI USE STATEMENT

We used generative AI, including language-model coding agents, for theoretical modeling; formulating claims and hypotheses; developing proof ingredients and drafting proofs; experimental design and parameter selection; method implementation and code development; benchmark-data cleaning; result interpretation; literature search, novelty checks and brainstorming; manuscript drafting, editing and reference formatting; and scientific figures. Translation and qualitative or thematic analysis were not applicable. Simulated populations were sampled programmatically, not generated by an LLM. Locally generated or publicly released language-model answers and scorer outputs were experimental observations.

We reviewed all AI-assisted work and take responsibility for the final text, claims and artifacts. Checks included exact-arithmetic polynomial and interval certificates for finite constructions and numerical checks of selected formulas and constants. Appendices provide proofs and distinguish numerical illustrations from certified constructions. Experimental results use saved outputs; Appendix D documents controls and uncertainty calculations. AI-assisted adversarial checks informed corrections without replacing author review.

ETHICS STATEMENT

This work uses public mathematics, knowledge-question-answering and programming benchmarks with their released labels. Newly generated model responses are evaluated automatically. No human participants were recruited, and no new human annotations or personal data were collected.

REPRODUCIBILITY STATEMENT

Appendices A, B, and C provide the assumptions and proofs of the theoretical results. Appendix A.7 gives exact-arithmetic certificates for the finite collision and the unique marginal-maximizing response pair. Appendix D documents the experimental data sources, model configurations, sampling protocols, estimators, candidate and label budgets, and uncertainty calculations, with references to the public benchmarks and model releases used. We will publicly release the experimental code upon acceptance.

REFERENCES

Carlos Améndola, Peter Friz, and Bernd Sturmfels. Varieties of signature tensors. *Forum of Mathematics, Sigma*, 7:e10, 2019. doi: 10.1017/fms.2019.3.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. URL `https://arxiv.org/abs/2108.07732`.

Jiawei Chang, Nick Duffield, Hao Ni, and Weijun Xu. Signature inversion for monotone paths. *Electronic Communications in Probability*, 22(42):1–11, 2017. doi: 10.1214/17-ECP70.

Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021. doi: 10.1080/01621459.2020.1758115.

Flavio Chierichetti, Mirko Giacchini, Ravi Kumar, Alessandro Panconesi, and Andrew Tomkins. Tight bounds for learning RUMs from small slates. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105864–105886, 2024. doi: 10.52202/079017-3358.

Laura Colmenarejo and Rosa Preiß. Signatures of paths transformed by polynomial maps. *Beiträge zur Algebra und Geometrie / Contributions to Algebra and Geometry*, 61(4):695–717, 2020. doi: 10.1007/s13366-020-00493-9.

José Correa, Paul Dütting, Felix Fischer, and Kevin Schewior. Prophet inequalities for I.I.D. random variables from an unknown distribution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 3–17, 2019. doi: 10.1145/3328526.3329627.

Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 1183–1202, 2015. doi: 10.1109/FOCS.2015.76.

Ilias Diakonikolas, Daniel M. Kane, and Sihan Liu. Testing closeness of multivariate distributions via Ramsey theory. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 340–347, 2024. doi: 10.1145/3618260.3649657. Proof locators refer to arXiv:2311.13154.

Florian E. Dorner, Yatong Chen, André F. Cruz, and Fanny Yang. ROC-n-reroll: How verifier imperfection affects test-time scaling. In *International Conference on Learning Representations*, pp. 150895–150936, 2026. URL https://openreview.net/forum?id=3Gy5mmyuxn.

Dimiter Dryanov and Petar Petrov. Canonical sets of best L_1-approximation. *Journal of Function Spaces and Applications*, 2012:435945, 2012. doi: 10.1155/2012/435945.

EvalPlus Team. EvalPlus v0.2.0. Software and dataset release, 2023. URL https://github.com/evalplus/evalplus/releases/tag/v0.2.0.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. In *International Conference on Learning Representations*, pp. 18128–18163, 2025. URL https://openreview.net/forum?id=cbttLtO94Q.

Hiệp Hàn, Marcos Kiwi, and Matías Pavez-Signé. Quasi-random words and limits of word sequences. *European Journal of Combinatorics*, 98:103403, 2021. doi: 10.1016/j.ejc.2021.103403.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Junoh Kang, Kiseop Lee, and Bohyung Han. Probabilistic signature inversion: Learning conditional distributions from truncated signatures. *arXiv preprint arXiv:2606.15332*, 2026. URL https://arxiv.org/abs/2606.15332.

Chris A. J. Klaassen and Robert M. Mnatsakanov. Optimal rate estimation of the mixing distribution in Poisson mixture models via Laplace inversion. *Transactions of A. Razmadze Mathematical Institute*, 174(2):175–188, 2020. URL https://rmi.tsu.ge/transactions/TRMI-volumes/174-2/v174(2)-6.pdf.

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

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, pp. 39578–39601, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-Reward-V2: Scaling preference data curation via human-AI synergy. *arXiv preprint arXiv:2507.01352*, 2025. URL https://arxiv.org/abs/2507.01352.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Arpan Mukherjee, Marcello Bullo, Debabrota Basu, and Deniz Gündüz. Test-time verification via optimal transport: Coverage, ROC, & sub-optimality. In *International Conference on Learning Representations*, pp. 149012–149048, 2026. URL https://openreview.net/forum?id=BBDhQJh6GB.

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

D. J. Newman and T. J. Rivlin. Approximation of monomials by lower degree polynomials. *Aequationes Mathematicae*, 14(3):451–455, 1976. doi: 10.1007/BF01835995.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016. doi: 10.1073/pnas.1607774113.

Max Pfeffer, Anna Seigal, and Bernd Sturmfels. Learning paths from signature tensors. *SIAM Journal on Matrix Analysis and Applications*, 40(2):394–416, 2019. doi: 10.1137/18M1212331.

Allan M. Pinkus. *On L^1-Approximation*, volume 93 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989. doi: 10.1017/CBO9780511526497.

Yury Polyanskiy and Yihong Wu. Dualizing Le Cam's method for functional estimation, with applications to estimating the unseens. *arXiv preprint arXiv:1902.05616*, 2021. URL https://arxiv.org/abs/1902.05616.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

James Robins, Lingling Li, Eric Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 2 of *Institute of Mathematical Statistics Collections*, pp. 335–421. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2008. doi: 10.1214/193940307000000527.

Arvind K. Saibaba. Approximating monomials using Chebyshev polynomials. *arXiv preprint arXiv:2101.06818*, 2021. URL https://arxiv.org/abs/2101.06818.

Götz S. Uhrig. Keeping a quantum bit alive by optimized π-pulse sequences. *Physical Review Letters*, 98(10):100504, 2007. doi: 10.1103/PhysRevLett.98.100504.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024. doi: 10.52202/079017-3018.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Wen Yang and Ren-Bao Liu. Universality of Uhrig dynamical decoupling for suppressing qubit pure dephasing and relaxation. *Physical Review Letters*, 101(18):180403, 2008. doi: 10.1103/PhysRevLett.101.180403. Equation locators refer to arXiv:0807.4058v1.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, 2025. doi: 10.18653/v1/2025.findings-acl.547.

A INFORMATION LIMITS FOR ORDERED AND NUMERICAL RECORDS

Exact words and their numerical counterparts use a common response; Appendix B treats the separate heterogeneous score-sharing model.

A.1 RANK TARGETS AND POLYNOMIAL RECOVERY

Proof of the rank identity (1). Conditional on $X = x$, break ties independently using $V \sim \mathrm{Unif}[0,1]$. Write $F_x(s^-) = \Pr(S' < s \mid x)$ and $q_x(s) = \Pr(S' = s \mid x)$. Then $U = F_x(S^-) + Vq_x(S)$ is uniform: at a generalized t-quantile with CDF limits $F^- \le t \le F$, its lower-tail probability is $F^- + (F - F^-)(t - F^-)/(F - F^-) = t$; zero atoms and endpoints follow by limits. A fresh candidate is lexicographically below (S, V) with probability U. Candidate 1 therefore wins among K conditionally iid draws with probability U_1^{K-1}. Exchangeability gives $\theta_K = K\mathbb{E}[G_1 U_1^{K-1}]$.

Polynomial-marginal ambiguity. Put $V = \mathcal{P}_{B-1}$ and $T\eta = (\int f_q \eta)_{q=1}^B$. The largest target diameter over fibres of T is

$$\sup_z \operatorname{diam}\{\langle f_K, \eta\rangle : T\eta = z,\ 0 \le \eta \le 1\} = \min_{v \in V} \|f_K - v\|_1 = \max_{\substack{\|h\|_\infty \le 1 \\ h \perp V}} \int f_K h. \tag{21}$$

A fibre difference is bounded by one and annihilates V; conversely every such h is realized by $(1 \pm h)/2$ in the central fibre. Pairing against $v \in V$ proves the upper bound. For $d = \mathrm{dist}_1(f_K, V) > 0$, extend $\Lambda(v + cf_K) = cd$ from $V + \mathbb{R}f_K$ by Hahn–Banach. Its norm-one L^∞ representative annihilates V and attains d. The case $d = 0$ is immediate. A bounded minimizing sequence in finite-dimensional V gives primal attainment. The response box is weak-star compact (test its bounds against nonnegative L^1 functions), so finite linear fibres have attained interval endpoints. The central interval is $[(1-d)/2, (1+d)/2]$.

An interpolation certificate with a common residual sign. For integer $N \ge 1$ and real $k > N$, write $\Delta_{N,k} = \mathrm{dist}_1(kx^{k-1}, \mathcal{P}_{N-1})$ and let

$$x_j = \sin^2 \frac{j\pi}{2(N+1)}, \qquad \sigma_N(x) = \operatorname{sgn} U_N(2x-1).$$

The interpolant of kx^{k-1} at these N nodes has residual sign σ_N, since its Nth derivative is positive on $(0,1)$. Under $2x - 1 = \cos\vartheta$, a polynomial of degree below N times the Jacobian is a combination of $\sin(j\vartheta)$, $j \le N$, orthogonal to the square wave with frequencies $(2r+1)(N+1)$. Hence $\sigma_N \perp \mathcal{P}_{N-1}$ and

$$\Delta_{N,k} = \int_0^1 kx^{k-1}\sigma_N(x)\,dx = 1 + 2\sum_{j=1}^{N} (-1)^{N-j+1} x_j^k. \tag{22}$$

For integer $1 \le k \le N$, interpolation reproduces the target and the residual is zero. In particular, every nonzero residual above has the same sign; this is the property needed for positive mixtures of deployment kernels.

For a useful upper bound, set $x = \cos^2 y$. In absolute value the integral pairs $g_k(y) = 2k\cos^{2k-1} y \sin y$ with $\operatorname{sgn}\sin(2(N+1)y)$. The latter has a primitive bounded by $\pi/(2(N+1))$. The function g_k is unimodal, absolutely continuous, zero at both endpoints, and has maximum $\sqrt{2k}(1 - 1/(2k))^{k-1/2} \le \sqrt{2k}$. Integration by parts gives

$$\Delta_{N,k} \le \frac{\pi\sqrt{2k}}{N+1}. \tag{23}$$

A.2 EXACT SATURATION OF THE COMPLETE-WORD AMBIGUITY

The upper width follows from marginals. Equality requires a pair that matches every joint word probability and attains that width.

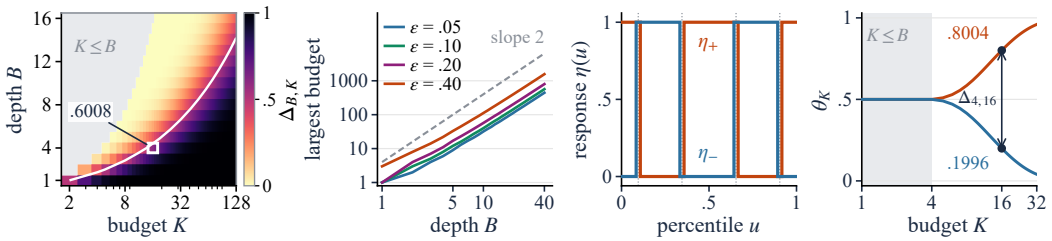


Figure 2: **Exact ambiguity and its attaining pair.** Left to right: closed-form $\Delta_{B,K}$ (grey: identified region $K \le B$; white: $\Delta_{B,K} = 1/2$); horizon $\max\{K : \Delta_{B,K}/2 \le \epsilon\}$ (dashed: asymptotic slope two; secant slopes over $B = 4$–40 are 1.85–1.87); the depth-four collision $\eta_\pm = \frac{1}{2} \pm \frac{1}{2} \operatorname{sgn} U_4(2u-1)$, with coincident jumps offset for visibility and marked by dotted lines; deployment values, equal for $K \le 4$ and separated by exactly $\Delta_{4,K}$ thereafter.

For a measurable $0 \le \eta \le 1$, define the absolutely continuous path

$$Z_\eta(t) = \left(\int_0^t (1-\eta), \int_0^t \eta \right), \qquad S_w(Z_\eta) = \int_{u_1 < \cdots < u_r} dZ_\eta^{w_1}(u_1) \cdots dZ_\eta^{w_r}(u_r).$$

Then $p_r(\eta)[w] = r! S_w(Z_\eta)$ and $S_\varnothing = 1$. This representation is classical (Hàn et al., 2021; Chang et al., 2017). The path is coordinatewise nondecreasing for every bounded response; this does not require η itself to increase. Uniformly subsampling r candidates from a depth-B pool shows that p_B determines all levels through $r = B$; no independence of these coordinates is assumed.

Lemma A.1 (Ordered cancellation, Yang & Liu, 2008)**.** *Let $h_B = \operatorname{sgn} U_B(2u-1)$, with switches $x_j = \sin^2(j\pi/(2B+2))$. For odd n and nonnegative integers p_i with $n + \sum_i p_i \le B$,*

$$\int_{u_1 < \cdots < u_n} \prod_{i=1}^{n} h_B(u_i) u_i^{p_i} \, du_{1:n} = 0. \tag{24}$$

The monomial can be replaced by any polynomial of total degree at most $B - n$.

Specialization of the cited cancellation theorem. In Yang & Liu (2008, Eqs. (9)–(10)), take pulse order $N = B$ and duration $T = 1$. Their switching function F_B has the same nodes x_j and equals $(-1)^B h_B$ almost everywhere. Their ordered scalar integral $F_{p_1,\ldots,p_n}$ is therefore $(-1)^{Bn}$ times (24), with the same integration order. Their cancellation condition is precisely odd n and $n + \sum_i p_i \le B$. Linearity gives the stated polynomial extension. This use needs no regularity of the response η. □

The one-point case and the leading coefficient 4^B of $U_B(2u-1)$ give

$$\int u^j h_B = 0 \ (j < B), \qquad m_B := \int u^B h_B = 4^{-B}, \tag{25}$$

since $\int_0^1 |U_B(2u-1)| du = \frac{1}{2} \int_0^\pi |\sin((B+1)t)| dt = 1$.

Lemma A.2 (Primitive transfer of complete-word collisions)**.** *Let $s \ge 0$, $B \ge 1$ be integers, $n = (s+1)B$, and c a polynomial of degree at most s. Write $J^0 h = h$, $J^{r+1} h(u) = \int_0^u J^r h(t) dt$. Whenever $\eta_\pm = c \pm a J^s h_n$ lie in $[0,1]$,*

$$p_m(\eta_+) = p_m(\eta_-) \ (m \le B), \qquad |\theta_K(\eta_+) - \theta_K(\eta_-)| = \frac{2|a| \Delta_{n,K+s}}{(K+1)\cdots(K+s)}.$$

The empty product is one. The gap may vanish when $K + s \le n$.

Proof. Subtracting the word expansions leaves an odd number of $J^s h_n$ factors. Express each primitive by s nested variables, one carrying h_n, and each monomial u^q of a baseline factor by $q!$ times a q-fold constant integral. At most $(s+1)m \le n$ variables occur. Partition the partial-order integration domain into its finitely many linear extensions. In each resulting simplex, integrate the

variables not carrying h_n: they give a gap polynomial of degree at most the number integrated out. There are an odd number of remaining wave factors, so Lemma A.1 annihilates every term. This is cancellation of the whole word law, not just its marginals.

The moments of h_n below order n vanish, so $J^r h_n(0) = J^r h_n(1) = 0$ for $1 \le r \le s$. For $s \ge 1$, integration by parts gives

$$\int_0^1 K u^{K-1} J^s h_n(u) du = \frac{(-1)^s}{(K+1)\cdots(K+s-1)} \int_0^1 u^{K+s-1} h_n(u) du.$$

The common-sign formula (22) gives the displayed target gap. The case $s = 0$ is immediate from that same formula. Degree lifting of this kind has a general signature interpretation (Colmenarejo & Preiß, 2020, Theorem 3.9); the quantitative target gap and its shape-compatible use here are the additional conclusions. □

Proof of exact ambiguity in Theorem 3.1. Apply Lemma A.2 with $s = 0$, $c = 1/2$, $a = t$. It gives every original word-law equality and the target gap $2|t|\Delta_{B,K}$. Conversely, equal depth-B word laws give equal rank marginals with densities

$$b_{j,B}(u) = \frac{B! u^{j-1}(1-u)^{B-j}}{(j-1)!(B-j)!}, \qquad 1 \le j \le B, \tag{26}$$

which span $\mathcal{P}_{B-1}$. Every collision difference is bounded by one and annihilates this space. Equation (21) supplies the upper width, attained by $t = 1/2$. For $K \le B$ both widths vanish. Compactness, proved in Appendix A.5, ensures attained exact-fibre endpoints, whose midpoints give minimax radius $\Delta_{B,K}/2$.

The exact-record horizon follows from the same explicit residual. The upper bound (23) gives success below a sufficiently small ϵ-dependent multiple of B^2, with all $K \le B$ exactly identified. Reversing the sum in (22) gives

$$\Delta_{B,K} = 1 + 2\sum_{j=1}^{B} (-1)^j \cos^{2K} \frac{j\pi}{2(B+1)} \ge 1 - 2\exp\{-K/(B+1)^2\}.$$

The inequality follows from an alternating decreasing sequence and $\sin(\pi/(2(B+1))) \ge 1/(B+1)$. Thus for each fixed $\epsilon < 1/2$, a sufficiently large multiple of B^2 has radius above ϵ. Depth $B+1$ first separates the exhibited pair: its top-label gap is $2t(B+1)4^{-B} \ne 0$ for $t \ne 0$.

The depth-three example in Figure 1. The switches are $(2-\sqrt{2})/4$, $1/2$, and $(2+\sqrt{2})/4$. Taking the response values $0, 1, 0, 1$ on the four successive intervals and their complements gives identical depth-three word laws by the preceding proof. Equation (22) gives $\Delta_{3,12} = 45967/65536$, with success probabilities $111503/131072$ and $19569/131072$. The resulting two-system selection regret is at least $\Delta_{3,12}/2 = 45967/131072$ in one of the indistinguishable worlds.

Not all fibres are wide. For $F(u) = \int_0^u \eta$ and $W \sim p_3(\eta)$, ordered integration gives

$$\mathbb{E}[(W_1 - c)(W_2 - c)] = 3\int_0^1 (F(u) - cu)^2 \, du. \tag{27}$$

Indeed $\mathbb{E}W_1W_2 = 3\int F^2$ and $\mathbb{E}(W_1 + W_2) = 6\int uF$. Consequently $p_3(\eta) = p_3(c)$ forces $\eta = c$ almost everywhere (Hàn et al., 2021, Remark 26). A nontrivial collision pair does not share its constant midpoint's law.

A.3 Statistical recovery in the same complete-word experiment

One homogeneous measurable binary response generates M independent words; their ranks remain hidden. Put $S = \min\{B, \log(1+M)\}$ and $Q = BS$. Equation (6) combines Theorem 3.2 with the bound proved here, using $(b, g) = (B, M)$. Both sides are established in the same homogeneous word experiment.

Proof of the word-sampling bound. For $K > B$, we establish universal constants in

$$c_1 e^{-C_1 Q/K} \le R^{\mathrm{word}}_{B,M}(K) \le C_2 e^{-c_2 Q/K}. \tag{28}$$

1. Approximate the square root of the target density. Under $f_B(u)du$, the variable $z = (1-u)/u$ has law $\nu_B(dz) = B(1+z)^{-B-1}dz$. Project $h(z) = \sqrt{K/B}(1+z)^{-(K-B)/2}$ onto $\mathcal{P}_R$ in $L^2(\nu_B)$, where $0 \le 2R \le B-1$, obtaining A. Multiplication by u^R turns this into projection of $\sqrt{K/B}u^\lambda$ under $Bu^{A_0-1}du$, with $A_0 = B - 2R$ and $\lambda = R + (K-B)/2$. The Gram matrix has Cauchy entries $(A_0 + j + k)^{-1}$. Its augmented determinant ratio, multiplied by $K = A_0 + 2\lambda$, gives

$$\|h - A\|_2^2 = \delta_R^2, \qquad \delta_R = \prod_{j=0}^{R} \frac{K - B + 2j}{K + B - 2j}. \tag{29}$$

Normalize $v = A/\sqrt{1-\delta_R^2}$. Since $b_{B-r,B}/f_B = \binom{B-1}{r} z^r$, the density $p = f_B v^2$ lies in the span of the last $2R+1$ rank densities. Pythagoras and Cauchy–Schwarz applied to $|h^2 - v^2|$ yield

$$\|f_K - p\|_1 \le 2\sqrt{2(1 - \sqrt{1-\delta_R^2})}. \tag{30}$$

2. Control the score coefficients by orthogonal projection. For $0 \le j \le R$, use the Rodrigues polynomials

$$p_j(z) = \frac{(1+z)^{B+1}}{j!} \frac{d^j}{dz^j}[z^j (1+z)^{j-B-1}].$$

They have degree j, $p_j(0) = 1$, and are orthogonal to $\mathcal{P}_{j-1}$ under ν_B: j integrations by parts prove this, with vanishing boundary terms since $B > 2j$. A sign-change argument gives j simple positive roots, so their coefficients alternate. The same integrations by parts give

$$\langle h, p_j \rangle_{\nu_B} = \frac{(-1)^j B}{j!} \int_0^\infty h^{(j)}(z) z^j (1+z)^{j-B-1} dz > 0,$$

because h is completely monotone. Hence its projection $A = \sum_{j \le R} \langle h, p_j \rangle p_j / \|p_j\|_2^2$ also has alternating coefficients. Writing $p = \sum_j \beta_j b_{j,B}$, use $\int z^\ell \nu_B(dz) = \binom{B-1}{\ell}^{-1}$ and the sign of each coefficient of A^2 to obtain

$$\|\beta\|_1 = \frac{\|A(-z)\|_2^2}{\|A(z)\|_2^2}. \tag{31}$$

For independent $X \sim \mathrm{Exp}(1)$ and $T \sim \mathrm{Gamma}(B,1)$, $X/T \sim \nu_B$. Exponential orthonormal Laguerre polynomials obey $L_n(-x) = \sum_{k=0}^n (-1)^k 2^{n-k} \binom{n}{k} L_k(x)$. Through degree R, absolute column and row sums are at most 3^R and $(3/2)3^R$, so the squared reflection operator norm is at most $(3/2)9^R$. Apply Parseval to $A(x/t)$ and integrate over T. Thus

$$\|\beta\|_1 \le (3/2)9^R. \tag{32}$$

Averaging $Z = \sum_j \beta_j W_j$ over the independent words and clipping gives

$$R^{\mathrm{word}}_{B,M}(K) \le \sqrt{2(1 - \sqrt{1-\delta_R^2})} + \frac{3\,9^R}{4\sqrt{M}}. \tag{33}$$

Bias is half the L^1 distance because both densities integrate to one, and the score range is at most $\|\beta\|_1$. For $M \ge 2$, take $D_0 = \min(B, \log M)$ and $\kappa = (8 \log 9)^{-1}$. If $D_0 \ge 1/\kappa$, choose $R = \lfloor \kappa D_0 \rfloor$. Then $2R \le B - 1$,

$$\delta_R \le e^{-(R+1)(B-R)/K} \le e^{-\kappa(1-\kappa)BD_0/K}, \qquad \frac{3\,9^R}{4\sqrt{M}} \le \frac{3}{4} e^{-3BD_0/(8K)}.$$

Use $S/2 \le D_0 \le S$. When D_0 is smaller or $M = 1$, Q/K is bounded and the midpoint suffices. This also proves the upper bound for heterogeneous responses: the conditional mean is linear in each response and the score range is uniform over words.

3. Quantify a homogeneous collision in a short tail. For $\ell = r+1$, piecewise integration of (22) and a 2ℓth roots-of-unity filter yield

$$\Delta_{r,n} = 2\ell \sum_{m\in\mathbb{Z}} \Pr\{X_n = n + (2m+1)\ell\}, \qquad X_n \sim \mathrm{Bin}(2n, 1/2). \tag{34}$$

Both sides vanish for $0 \le n \le r$. For $n \ge 1$, expand $1 + 2\sum_{j=1}^{\ell-1}(-1)^j \cos^{2n}(j\pi/(2\ell))$; the filter retains exactly the displayed residues. Two-sided Stirling gives $\Pr(X_n = n+d) \ge cn^{-1/2}e^{-Cd^2/n}$ for $0 \le d \le n$. At $d = n$ use 4^{-n}; otherwise use $(1+x)\log(1+x)+(1-x)\log(1-x) \le 2(\log 2)x^2$ on $[0,1]$. For $n \le \ell^2$ retain $d = \ell$; for $n \ge \ell^2$ retain the at least $c\sqrt{n}/\ell$ odd multiples with $|d| \le \sqrt{n}$. Hence

$$\Delta_{r,n} \ge c_0 e^{-C_0(r+1)^2/n}, \qquad n > r \ge 1. \tag{35}$$

Take responses equal to $1/2$ below $1-a$ and equal to $(1\pm h_r((u-1+a)/a))/2$ above it. Conditioning on the deployment tail count gives target gap

$$\Gamma_{K,r,a} = \mathbb{E}\Delta_{r,N}, \qquad N \sim \mathrm{Bin}(K, a). \tag{36}$$

Let $\gamma = 1/5$, $a = \gamma r/B$, $1 \le r < B < K$, and $\lambda = Ka$. If $\lambda \ge 4\ell$, Chernoff gives $\Pr(N \ge \lambda/2) \ge 1-e^{-1}$ and gap at least $c_0 e^{-8C_0 Br/(\gamma K)}$. If $\lambda < 4\ell$, retain $N = \ell$. Since $\lambda/\ell > \gamma/2$, $\binom{K}{\ell} \ge (K/\ell)^\ell$, and $\log(1-a) \ge -a/(1-a)$, this event has probability at least $(\gamma/2)^\ell e^{-\lambda/(1-\gamma)} \ge e^{-C_\gamma \ell}$. Also $\Delta_{r,\ell} = 4\ell 4^{-\ell}$ and $\ell \le 16Br/(\gamma K)$. Thus in both cases

$$\Gamma_{K,r,\gamma r/B} \ge ce^{-CBr/K}. \tag{37}$$

4. Choose the hiding order for the observed sample. Write $L_0 = \log(1+M)$. If $B \le 4L_0$, use the exact collision $r = B, a = 1$ and $(B+1)^2 \le 16BS$. Otherwise choose $r = \lceil 5L_0/(4I_\gamma)\rceil < 3L_0 < B$, where $I_\gamma = \log(1/\gamma) - 1 + \gamma$. Conditional on at most r sampled tail candidates, the two tail laws agree and outside labels are common fair coins. Hence one-word TV is at most $\alpha = \Pr\{\mathrm{Bin}(B,a) > r\} \le e^{-rI_\gamma}$, and

$$\mathrm{TV}(P_+^{\otimes M}, P_-^{\otimes M}) \le M\alpha \le \frac{M}{(1+M)^{5/4}} \le \frac{4}{5^{5/4}} < 1.$$

The absolute-loss common-mass bound is half the target gap times $1 - \mathrm{TV}$. Together with (37), this proves the lower bound in the same homogeneous experiment, without importing a population-mixture converse.

A.4 Why unrestricted numerical observations reduce to words

We prove Theorem 3.2 for a fixed finite target list $\mathbf{K} = (K_1, \dots, K_d)$, $T(\eta) = (\theta_{K_1}, \dots, \theta_{K_d})$, with loss $\ell(t, T) = \|t - T\|_\infty$. Rules may be measurable and randomized. Clipping to $[0,1]^d$ cannot increase risk. The proof extends classical order reductions (Diakonikolas et al., 2015; 2024).

Lemma A.3 (Coarsening and product homogeneity)**.** *Let η_n average η on n equal rank intervals. Uniformly over measurable $0 \le \eta \le 1$,*

$$\mathrm{TV}(p_b(\eta_n), p_b(\eta)) \le \binom{b}{2}/n, \qquad \|T(\eta_n) - T(\eta)\|_\infty \le \max_{k\in\mathbf{K}} \binom{k}{2}/n. \tag{38}$$

For finite integers c, g, b, n with $n \ge b$, there are finite ordered sets H_a such that every c-coloring of $\prod_{a=1}^g \binom{H_a}{b}$ is constant on $\prod_a \binom{V_a}{b}$ for some $V_a \subseteq H_a$ with $|V_a| = n$.

Proof. For r independent uniform ranks, the probability of a repeated cell is at most $\binom{r}{2}/n$. Otherwise the cells determine order; integrating independent within-cell positions gives independent labels with the cell-average parameters. Take $r = b$, or $r = k$ and retain the winning label. No uniform approximation of η is assumed.

For product homogeneity, $g = 1$ is finite hypergraph Ramsey. Inductively choose H_g for c colors and target size n. Color each input of the first $g-1$ factors by its entire function on $\binom{H_g}{b}$, using at most $c^{\binom{|H_g|}{b}}$ colors. Induction makes this function constant on the first selected factors; a final one-set Ramsey application supplies V_g. □

Proof of Theorem 3.2. Sorting gives $R^{\mathrm{num}} \le R^{\mathrm{word}}$. For the converse fix a clipped numerical rule A. Randomly permute sample identities within each cohort; exchangeability preserves risk, and this randomization stays inside the rule on sorted observations. Round each output within $h > 0$ to a finite grid of size D_0, and partition the output-probability simplex into cells of TV diameter h; a coordinate mesh $2h/D_0$ suffices.

In each cohort choose a sufficiently large ordered family of disjoint intervals $J_{a,t} \subset (0,1)$ with smooth probability densities $\rho_{a,t}$ supported inside. For a tuple $T \in \prod_a \binom{H_a}{b}$, draw one score per selected interval, attach each possible ordered word tuple w, and let $Q_{T,w}$ be the rounded rule's output law. Color T by the simplex cells of *all* 2^{bg} laws. Lemma A.3 gives n intervals per cohort and a reference tuple T_0. The word-only rule $\mathcal{L}(\psi(w)) = Q_{T_0,w}$ satisfies

$$\mathrm{TV}(Q_{T,w}, \mathcal{L}(\psi(w))) \le h \quad \text{for every selected } T \text{ and every } w. \tag{39}$$

Set $f_a = n^{-1} \sum_{t \in V_a} \rho_{a,t}$ and use response η_n. Its equal-mass intervals are exactly the rank cells, so labels are independent of numerical locations within a cell. Except with probability $q_n = g\binom{b}{2}/n$, no cohort repeats an interval. On that event the score law is exactly the law defining $Q_{T,w}$. Thus the estimator is within $h + q_n$ of the word-only rule on actual words, and those words are within another q_n of $p_b(\eta)^{\otimes g}$. Bounded loss, output rounding, and target Lipschitzness give

$$\mathbb{E}_{\eta_n,F}\ell(A, T(\eta_n)) \ge \mathbb{E}_\eta \ell(\psi, T(\eta)) - 2h - \frac{2g\binom{b}{2} + \max_{k \in \mathbf{K}} \binom{k}{2}}{n}. \tag{40}$$

The CDFs and ψ were fixed from (A, n, h) *before* choosing η. Hence this controls the response supremum, not a different calibration tailored to each response. Take $\sup_\eta$, then $n \to \infty$, $h \downarrow 0$, and finally $\inf_A$.

For smooth positive densities replace f_a by $(1-\zeta)f_a + \zeta q$, with q smooth and positive on $(0,1)$. Couple scores and then labels. A changed score costs probability ζ; on coincident scores the step response changes only within ζ of a rank-cell boundary. Consequently

$$\mathrm{TV}(\mathcal{L}_{\eta_n,F}(\mathrm{data}), \mathcal{L}_{\eta_n,F_\zeta}(\mathrm{data})) \le bg[1 + 2(n-1)]\zeta.$$

The target is unchanged. Send $\zeta \downarrow 0$ first at fixed (n, h). A Gaussian background similarly supplies positivity on all of $\mathbb{R}$. No complexity, modality, derivative, or finite-precision bound is asserted.

Restriction to uniformly bounded-variation response classes. Let $\mathcal{H}$ have common variation bound $V < \infty$. Uniform-cell averages obey $\sup_{\eta \in \mathcal{H}} \|\eta - \eta_n\|_1 \le V/n$, even when the averages do not belong to $\mathcal{H}$. Keep the same CDFs and word rule constructed from (A, n, h) before choosing η. Coupling labels at identical uniform ranks changes either numerical or word data by at most bgV/n in TV; a target at K changes by at most KV/n. Applying the intermediate step-response comparison and returning to the original response gives

$$\mathbb{E}_{\eta,F}\ell(A, T\eta) \ge \mathbb{E}_\eta \ell(\psi, T\eta) - 2h - q_n - 2(bg + K_{\max})V/n.$$

Take the supremum on $\mathcal{H}$ before $n \to \infty$, $h \downarrow 0$ and the infimum over A. Sorting gives the reverse bound. This contains the uniformly Lipschitz extension, for which $V \le L$, and all bounded nondecreasing responses, for which $V \le 1$.

For a positive-density background, the retained score's percentile shifts by at most ζ. Integrating the variation measure over the interval of radius ζ bounds label mismatch by $2V\zeta$, even at jumps. Changed scores cost another ζ. Thus data TV is at most $bg(1 + 2V)\zeta$, uniformly on the class. Send $\zeta \downarrow 0$ at fixed (n, h) first. This proves only the stated fixed-finite-list minimax equality; no infinite-target experiment equivalence is assumed.

Observed iid identities. For M iid draws from J observed categories, uniform sampling, common η, and one arbitrary F_x per category, let $p_{M,J}$ be the repeat probability. Then

$$R^{\mathrm{word}}_{B,M}(K) - p_{M,J} \le R^{\mathrm{iid,J}}_{B,M}(K) \le R^{\mathrm{word}}_{B,M}(K), \quad p_{M,J} \le \min\{1, \binom{M}{2}/J\}. \tag{41}$$

The upper bound holds conditional on repeats. For $J \ge M$, color products over all J categories by the finite table indexed by every ordered distinct sampled-category tuple and every word tuple,

ignoring inactive coordinates. Product Ramsey fixes all CDFs simultaneously. Average the resulting word rules over category tuples conditional on no repetition, independently of words. Equation (40) and nonnegative loss on repeats give $(1 - p_{M,J})\mathbb{E}_\eta \ell(\psi, T) - o(1)$; bounded loss proves the claim. For $J < M$ the lower bound is trivial. This retains identities, rather than assuming they are hidden.

Infinite repetition at fixed cohort depth. For fixed b, K, write a best approximant as $v = \sum_{j=1}^{b} a_j b_{j,b}$ and $r = f_K - v$. The affine estimator $g^{-1} \sum_{a,j} a_j W_{aj} + \frac{1}{2} \int r$ has bias at most $\Delta_{b,K}/2$ and expected absolute fluctuation at most $\sum_j |a_j|/(2\sqrt{g})$. Clip and combine with the exact collision lower bound to obtain

$$\lim_{g \to \infty} R^{\mathrm{num}}_{b,g}(K) = \Delta_{b,K}/2.$$

The nuisance class gains an unrelated CDF with each cohort. This is not pointwise nonidentification in an infinite sample from one fixed CDF; the large-domain limitation of Ramsey reductions is discussed by Diakonikolas et al. (2024, Section 3.4).

A.5 Shape-stable identification and the boundary of saturation

A.5.1 Proof of Theorem 4.1: fixed shape constraints

Proof strategy. The obstacle is exact collision under fixed shape constraints: smoothing a collision need not preserve its word law, and a derivative kernel alone gives only first-order equality. We first locate a maximal-rank smooth response with strict shape slack, then correct an invisible perturbation to keep every word coordinate fixed. Finally, two exponential families with incommensurate rates restrict identifiable kernels to polynomials. This is a class-wide statement; individual fibres can still be singletons (Hàn et al., 2021, Theorem 5 and Remark 26).

Fix κ, L and extend $F = p_B$ polynomially to $C^1[0, 1]$. Its maximal derivative rank $r \geq 1$ is attained, since the possible ranks form a finite set.

A maximal-rank response in the shape class. For every $c > 0$, the spaces $E_{c,m} = \mathrm{span}\{1, e^{cu}, \ldots, e^{mcu}\}$ have dense union in C^1: substitute $x = e^{cu}$, approximate the derivative by polynomials, then integrate and match one value. Approximate a maximal-rank response and r witnessing directions in one $E_{c,m}$. The resulting nonzero derivative minor is a polynomial in the response coefficients. It cannot vanish on the nonempty open set where

$$\eta_c(u) = a_0 + \sum_{j=1}^{m} a_j e^{jcu} \in \mathcal{H}_{\kappa,L}, \qquad a_j > 0 \ (j \geq 1).$$

Such a set exists by taking a_0 near $1/2$ and all positive coefficients small. Thus η_c may be chosen of rank r, for every fixed $L > 0$ and, likewise, in any open admissible strip of response levels.

Derivative kernels and exact fibres. Differentiating one factor of a word integral splits its remaining variables into an ordered prefix and suffix. Multiplication and integration preserve nonnegative exponential rates; evaluation at one adds only constants. There are at most $B - 1$ integrations. Hence, including $B = 1$,

$$DF_w(\eta_c)[h] = \int k_{w,c} h, \qquad k_{w,c} \in V_{c,m,B} := \mathrm{span}\{u^q e^{jcu} : 0 \leq q < B,\ 0 \leq j \leq m(B-1)\}.$$

If T_g is outside the derivative row span, annihilator duality supplies $h_0 \in \ker DF(\eta_c)$ with $\int g h_0 \neq 0$. Choose r smooth compensating directions h_i with an invertible minor on r output coordinates. Approximate h_0 smoothly and correct those coordinates using the h_i; the correction tends to zero, preserving its nonzero target pairing and making every derivative coordinate zero.

The implicit function theorem applied to $\eta_c + t h_0 + \sum_i z_i(t) h_i$ fixes the selected outputs, with $z(0) = z'(0) = 0$. Strict C^1 slack preserves the shape constraints. The selected minor remains invertible, while global maximality bounds the full rank by r. Thus every other derivative row lies in their span along the curve: *all* word coordinates stay fixed, although the target derivative is $\int g h_0 \neq 0$. An identifiable kernel must therefore belong to $V_{c,m,B}$ almost everywhere.

Intersection and converse. Apply this conclusion for $c = 1$ and $c = \sqrt{2}$, with possibly different m. Their positive rates are disjoint. Independence of distinct-rate exponential polynomials follows by applying $\prod_{\lambda \neq \lambda_0}(D - \lambda)^B$: on the remaining polynomial, every factor becomes $(D + \alpha)^B$, $\alpha \neq 0$, and is invertible. Consequently the two spaces intersect in exactly $\mathcal{P}_{B-1}$; almost-everywhere equality of their analytic representatives is equality everywhere. Rank marginals recover this space, including under mixing; Dirac populations give the reverse inclusion. Class inclusion then gives the nondecreasing and unrestricted classifications.

The result concerns identifiable targets over a class, not a quantitative gap or nonidentification of each response. It is distinct from particular-response finite forcing (Hàn et al., 2021, Theorem 5), growing-level inversion of coordinatewise monotone paths (Chang et al., 2017), and prior-based reconstruction (Kang et al., 2026). Coordinatewise monotonicity of the path holds for every bounded response and does not require monotonicity of η.

A.5.2 Attainment and strict nonsaturation

Attainment and an exact saturation criterion. For $w \in L^1$, let $\mathcal{C}_B$ be the set of differences of responses with equal p_B. Define $\Omega_B^{\mathrm{sig}}(w) = \max_{d \in \mathcal{C}_B} |\langle w, d\rangle|$ and $D_B(w) = \min_{v \in \mathcal{P}_{B-1}} \|w - v\|_1$. Both extrema are attained. For the first, approximate the simplex indicator in $L^1([0,1]^B)$ by finite sums of product kernels. Their integrals are products of weak-star continuous affine functionals, and the approximation is uniform over bounded responses. Thus p_B is weak-star continuous and its collision set compact. Finite-dimensional norm compactness gives the second attainment. Since $d \perp \mathcal{P}_{B-1}$ and $|d| \leq 1$,

$$\Omega_B^{\mathrm{sig}}(w) \leq D_B(w), \qquad \Omega_B^{\mathrm{sig}}(w) = D_B(w) \iff \sigma d(w - v_*) = |w - v_*| \text{ a.e.} \tag{42}$$

for a best approximant v_*, a collision d, and $\sigma \in \{-1, 1\}$. This is exactly the equality condition in $|\langle w - v_*, d\rangle| \leq \|w - v_*\|_1$. A residual nonzero almost everywhere forces deterministic complementary responses.

A positive kernel for which the marginal bound is strict. Split $[0, 1]$ into 32 cells indexed from zero, and set $h = 1$ on

$$\{1, 3, 4, 6, 9, 10, 12, 13, 16, 21, 22, 23, 24, 26, 27, 31\},$$

with $h = -1$ elsewhere. The sums of i^r on this set are half the full sums, namely $(16, 248, 5208, 123008)$ for $r = 0, 1, 2, 3$. Cell integrals are polynomials of the same degrees, so $h \perp \mathcal{P}_3$. The density $w = 1 + h/2$ has $D_4(w) = 1/2$, attained by approximant 1 and dual sign h.

The deterministic cell word and its complement have respectively 2590 and 2594 scattered occurrences of 0101. These counts follow by initializing $N_0 = 1$, $N_{1:4} = 0$ and, for each bit b_i, updating $N_j \leftarrow N_j + \mathbf{1}\{b_i = (0101)_j\} N_{j-1}$ in descending j. Alternation forces distinct cells, and therefore

$$p_4((1 + h)/2)[0101] - p_4((1 - h)/2)[0101] = 4!(2590 - 2594)/32^4 = -3/2^{15} \neq 0. \tag{43}$$

Equality in (42) would force precisely this pair, which fails to collide. Thus $\Omega_4^{\mathrm{sig}}(w) < D_4(w)$. Both widths are 1-Lipschitz in L^1. Positive continuous approximations followed by normalized nonnegative Bernstein approximants preserve the strict gap once the error is below half that gap. Thus strictness also occurs for positive mixtures of fixed-rank beta densities. This limits saturation without a separate structured-response theory.

A.6 Quantitative regularity and the entire deep curve

Here $\mathcal{M}_L$ and the exact-law widths are as in Section 4.1. The class is compact in the uniform norm; the word map and every target are continuous, and $\sup_K |\theta_K(\eta) - \theta_K(\widetilde{\eta})| \leq \|\eta - \widetilde{\eta}\|_\infty$. A fibre's coordinatewise midpoints therefore give half its sup-norm diameter as exact-law minimax radius. The forecast need not itself be generated by a single response. Coordinate extrema are measurable functions of the finite observation, and the budget set is countable.

A.6.1 One canonical sign determines every marginal target

Put $n = B - 1$ and $w(u) = u(1 - u)$. Among monic degree-n polynomials let q_n minimize $\int w|q_n|$, with $q_0 = 1$, and set $\chi_B = \operatorname{sgn} q_n$. Existence follows from finite-dimensional norm coercivity. First

variation gives

$$\int_0^1 w(u)p(u)\chi_B(u)du = 0 \quad (p \in \mathcal{P}_{B-2}). \tag{44}$$

Here and below $\mathcal{P}_{-1} = \{0\}$. If q_n had fewer than n interior sign changes, a polynomial in $\mathcal{P}_{n-1}$ could have its sign, contradicting (44). Thus its roots are simple and lie in $(0, 1)$. Equality in the triangle inequality shows any two minimizers have the same sign, hence the same roots; monicity makes q_n unique. This is the classical canonical-L^1-interpolation mechanism (Pinkus, 1989; Dryanov & Petrov, 2012); the statistical consequences below use the derivative constraint.

For $1 \le K \le B$, the feasible polynomial u^K gives zero width. For every $K \ge B$, let $P_{B,K} \in \mathcal{P}_B$ interpolate u^K at 0, 1 and these $B-1$ roots. For $K = B$ it is exact. For $K > B$, the interpolation remainder has sign $\operatorname{sgn}[u(u-1)q_n] = -\chi_B$ because the $(B+1)$st derivative of u^K is positive on $(0, 1)$. Admissible polynomial differences are in $w\mathcal{P}_{B-2}$, so (44) proves optimality and the exact formula

$$e_{B,K} := \inf_{\substack{P\in\mathcal{P}_B \\ P(0)=0,\ P(1)=1}} \|u^K - P\|_1 = \|u^K - P_{B,K}\|_1 = \int_0^1 (u - u^K)\chi_B(u)du. \tag{45}$$

The same sign works for every target, rather than choosing a dual extremizer separately. Indeed $e_{B,K+1} - e_{B,K} = \int w u^{K-1}\chi_B > 0$ for $K \ge B$: interpolate u^{K-1} at the n roots by degree below n and use its positive nth derivative; for $n = 0$ positivity is immediate. Dominated convergence gives

$$e_{B,K} \uparrow \lambda_B := \int_0^1 u\chi_B(u)du = \min_{\substack{P\in\mathcal{P}_B \\ P(0)=0,\ P(1)=1}} \|P\|_1, \qquad \frac{1}{(B+1)^2} \le \lambda_B \le \frac{4}{B^2}. \tag{46}$$

For the equality, $P_\infty = uq_n/q_n(1)$ has sign χ_B, and every admissible P obeys $\int P\chi_B = \int u\chi_B$. For the lower bound, the shifted Legendre reproducing identity gives $1 = |P(1)| \le \sum_{j=0}^{B}(2j+1)\|P\|_1 = (B+1)^2\|P\|_1$. For the upper bound, with $m = \lfloor (B-1)/2 \rfloor$, the polynomial

$$E_1(u) = u\left[\frac{\sum_{j=0}^{m}(2j+1)L_j(u)}{(m+1)^2}\right]^2$$

has degree at most B, endpoint values $(0, 1)$, and $\int E_1 \le (m+1)^{-2} \le 4/B^2$ by Legendre orthogonality.

Exact width and a single worst pair for the whole curve. For any admissible P, integration by parts yields

$$\left|\theta_K - \int P'\eta - \frac{L}{2}\int (P - u^K)\right| \le \frac{L}{2}\|P - u^K\|_1. \tag{47}$$

The quantity $\int P'\eta$ is determined by the rank marginals. Define, once for all K,

$$d(u) = L\left[\int_0^u \chi_B(t)dt - \int_0^1 (1-t)\chi_B(t)dt\right], \qquad \eta_\pm(u) = \frac{1}{2} + \frac{L}{2}(u - \tfrac{1}{2}) \pm \frac{d(u)}{2}.$$

Their derivatives are $(L \pm L\chi_B)/2$. For $L \le 1$, left endpoints lie in $[1/2 - L/2, 1/2]$ and right endpoints in $[1/2, 1/2 + L/2]$, so both belong to $\mathcal{M}_L$. Their target gap at q is $L\int (u - u^q)\chi_B$, zero for $q \le B$ and equal to $Le_{B,K}$ at every $K \ge B$. Consequently

$$W_{B,K}^{\mathrm{mar}}(L) = Le_{B,K}, \qquad W_B^{\mathrm{mar,curve}}(L) = L\lambda_B. \tag{48}$$

The pair is independent of K, but is only asserted to match marginals, not full words. Figure 3 compares the ambiguity bounds. For any finite $L > 0$, (47) still holds. Taking $P = u^B$ recovers the block-winner rule with drift $L[(B+1)^{-1} - (K+1)^{-1}]/2$ and curve risk $1/(2\sqrt{M}) + L/[2(B+1)]$; the canonical choice improves the uniform bias order.

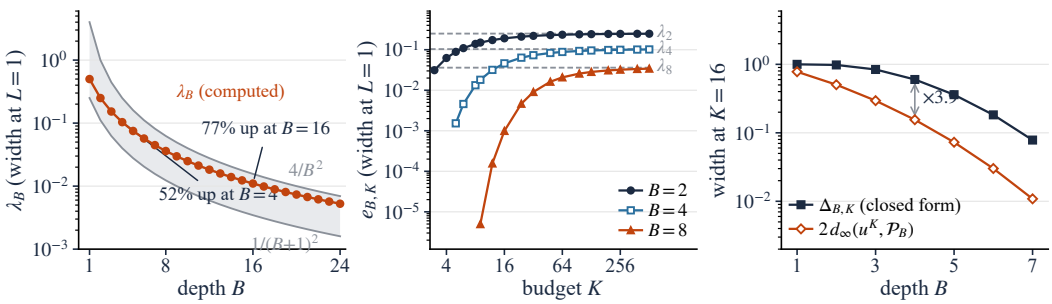


Figure 3: **Shape constraints and ambiguity bounds.** Left: λ_B within $[(B+1)^{-2}, 4B^{-2}]$. Centre: marginal width $e_{B,K}$ tending to λ_B. Right: unrestricted word width at $K = 16$ versus the monotone marginal benchmark $2\,\mathrm{dist}_\infty(u^{16}, \mathcal{P}_B)$, about 3.9 times smaller at $B = 4$. The monotone full-word width is not plotted; it is strictly below this benchmark at $(4, 16)$ (Corollary A.4). Only $\Delta_{B,K}$ is closed form; the other curves use fixed-grid linear programmes with reported accuracy about 4×10^{-3}.

A.6.2 A FULL NONLINEAR LOWER BOUND UNDER THE SAME SLOPE BUDGET

Lemma A.2 with $s = 1$ gives

$$\eta_\pm(u) = \frac{1}{4} + \frac{L}{2}u \pm \frac{L}{4} J h_{2B}(u), \quad p_B(\eta_+) = p_B(\eta_-), \quad |\theta_K(\eta_+) - \theta_K(\eta_-)| = \frac{L\Delta_{2B,K+1}}{2(K+1)}.$$

The slopes lie in $[L/4, 3L/4]$ and responses in $[1/4, 3/4]$. At $K_B = 2(2B+1)^2 - 1$, the alternating residual bound gives $\Delta_{2B,K_B+1} \geq 1 - 2e^{-2}$. Therefore

$$W_B^{\mathrm{word,curve}}(L) \geq \frac{1 - 2e^{-2}}{36} \frac{L}{B^2}.$$

The marginal upper bound follows from (48) and (46); words determine marginals. This proves (8) with width constants $c = (1 - 2e^{-2})/36$, $C = 4$. For mixtures, the mean response remains in $\mathcal{M}_L$, and a degenerate population supplies the lower bound. For general $L > 1$, the upper bound $4L/B^2$ still holds, while the class $\mathcal{M}_1$ supplies the lower bound c/B^2. Thus every fixed finite positive L has the same B^{-2} order, without claiming uniform linear dependence on large L.

Exact smooth witnesses. Let $R = n + 1$ and

$$f_{n,\varepsilon}(\vartheta) = \mathrm{sgn}(\sin R\vartheta) \exp[-\varepsilon^2 / \sin^2(R\vartheta)], \qquad h_{n,\varepsilon}(u) = f_{n,\varepsilon}(\arccos(2u - 1)),$$

with value zero at zeros. Flatness there makes $h_{n,\varepsilon} \in C^\infty[0, 1]$, including the endpoints. It is bounded by one and converges almost everywhere to h_n. The smooth odd function $f_{n,\varepsilon}$ is antiperiodic with shift π/R, so its absolutely convergent Fourier series contains only $\sin(kR\vartheta)$ for odd k. The coefficient-independent harmonic cancellation in Yang & Liu (2008, arXiv v1, Eq. (20) and its harmonic proof) therefore gives exactly (24) with $B = n$ for each $\varepsilon > 0$. The change $u = (1 - \cos\vartheta)/2$ contributes only an overall sign; the allowed total degree is unchanged.

Lemma A.2 consequently applies to $h_{n,\varepsilon}$. For $n = 2B$, the pair $\eta_{\pm,\varepsilon} = 1/4 + Lu/2 \pm (L/4) J h_{2B,\varepsilon}$ has equal complete depth-B laws, range $[1/4, 3/4]$ and slopes in $[L/4, 3L/4]$. The one-point cancellation gives $J h_{2B,\varepsilon}(1) = 0$, and

$$|\theta_K(\eta_{+,\varepsilon}) - \theta_K(\eta_{-,\varepsilon})| = \frac{L}{2} \left| \int_0^1 u^K h_{2B,\varepsilon}(u)\, du \right| \longrightarrow \frac{L\Delta_{2B,K+1}}{2(K+1)}.$$

Thus the same lower bound holds in the supremum over smooth strictly increasing responses. The displayed Δ value is a limit, not an identity at finite ε; no common bound on higher derivatives is imposed.

A.6.3 WHY QUALITATIVE SMOOTH MONOTONICITY IS INSUFFICIENT

Fix $B, \kappa < 1/2$ and $\varepsilon > 0$ with $t_0 = 1 - 2\kappa - 3\varepsilon > 0$. By the admissible-strip construction in Appendix A.5.1, choose an analytic maximal-rank response $\kappa + \varepsilon < \eta_0 < \kappa + 2\varepsilon$ with $\eta_0' > 0$. The r independent derivative kernels remain independent on any $I \Subset (0, 1/2)$ by analyticity. Thus some $h_i \in C_c^\infty(I)$ make their derivative matrix A invertible.

Choose a smooth nondecreasing q_a equal to zero below $1-a$ and one above $1-a/2$. For the selected word coordinates F_*, let

$$G_a(t, z) = F_*(\eta_0 + tq_a + \sum_i z_i h_i) - F_*(\eta_0), \qquad 0 \le t \le t_0.$$

Multilinearity of the word integrals and $\|q_a\|_1 \le a$ give, uniformly over this fixed t-interval,

$$\|G_a(t, 0)\| \le Ca, \qquad \|D_z G_a(t, z) - A\| \le C(a + \|z\|).$$

Consequently $z \mapsto z - A^{-1}G_a(t, z)$ contracts a radius-$C'a$ ball into itself for small a, uniformly in t. Its fixed point $z_a(t)$ is C^1 and satisfies $z_a(0) = 0$. No small C^1 norm is required of the ramp.

The selected minor remains invertible throughout this tube and the full rank cannot exceed r. The exact-fibre argument of Appendix A.5.1 therefore fixes every word coordinate, not just the selected ones. The interior correction has C^1 size $O(a)$ and is supported away from the ramp, so strict monotonicity and range persist. At t_0 this yields

$$\eta_a \in \mathcal{I}_\kappa, \qquad p_B(\eta_a) = p_B(\eta_0), \qquad \eta_a(1) - \eta_0(1) = t_0.$$

For this fixed pair, continuity at one gives $\theta_K(\eta_a) - \theta_K(\eta_0) \to t_0$. Sending $\varepsilon \downarrow 0$ proves (9), whose reverse bound is $1 - 2\kappa$. Slopes need not have a common bound, and the construction gives no threshold uniform in B; neither particular finite-forced fibres nor the fixed-L rate is contradicted.

The thin-tail illustration in Figure 1. For the separate depth-three example, set $a = 10^{-5}$, $t = .96$ and $\eta_0(u) = .01 + .01e^{2u-2}$. Define $v(s) = e^{-1/s}/(e^{-1/s} + e^{-1/(1-s)})$ for $0 < s < 1$, extended by zero and one outside this interval, and let $q(u) = v(2(u - 1 + a)/a)$. Write L_j for the shifted Legendre polynomials, $J_j(u) = \int_0^u L_j$, $Q(u) = \int_0^u q$, and $F_0(u) = \int_0^u \eta_0$. Set $c_j = -(2j+1)t\int_0^1 qL_j$ for $j = 0, 1, 2$ and $D_0 = tQ + \sum_{j=0}^2 c_j J_j$. Choose the small root c_3 of

$$\int_0^1 \{2F_0(D_0 + c_3J_3) + (D_0 + c_3J_3)^2\} = 0, \qquad \eta_1 = \eta_0 + tq + \sum_{j=0}^3 c_j L_j.$$

This preserves $m_j = \int u^j \eta$ for $j = 0, 1, 2$ and $\int F^2$. These determine all depth-three words: with $A_{ij} = \mathbb{E}W_iW_j$, $A_{12} = 3\int F^2$, $A_{23} = -3m_0^2 + 6m_0m_1 + A_{12}$, $A_{13} = 3m_0^2 - A_{12} - A_{23}$, and $\mathbb{E}W_1W_2W_3 = m_0^3$; expanding the binary word indicators completes the argument. Numerical quadrature gives $c_3 \approx -5.81821 \times 10^{-5}$, range bounds $.01123 < \eta_1 < .98013$, and a derivative lower bound .00174, using $|L_j| \le 1$ and $|L_j'| \le j(j+1)$. The plotted endpoint difference is $\eta_1(1) - \eta_0(1) \approx .959877$, rounded to 96.0 percentage points. It is a numerical illustration of the exact integral construction above, not a certified finite-K bound; the gap is its $K \to \infty$ limit.

A.6.4 A FINITE-BUDGET INSTANCE AT DEPTH TWO

Guarantees. For $B = 2$, the canonical node is $1/2$ and $\lambda_2 = 1/4$. Interpolating at $0, 1/2, 1$ gives

$$P_K(u) = a_K u^2 + (1 - a_K)u, \qquad a_K = 2 - 2^{2-K},$$
$$P_K' = \beta_{1,K}b_{1,2} + \beta_{2,K}b_{2,2}, \qquad (\beta_{1,K}, \beta_{2,K}) = \tfrac{1}{2}(1 - a_K, 1 + a_K).$$

Here $b_{1,2} = 2(1-u)$ and $b_{2,2} = 2u$. Use the same empirical rank means $\widehat{b}_1, \widehat{b}_2$ from all M prompts for every $K \ge 2$, and set $\widehat{\theta}_K = \sum_i \beta_{i,K}\widehat{b}_i + (L/2)\int(P_K - u^K)$. Equation (47), $|\beta_{1,K}| \le 1/2$ and $|\beta_{2,K}| \le 3/2$ imply

$$\mathbb{E}\sup_{K \ge 2} |\widehat{\theta}_K - \theta_K| \le \frac{L}{8} + \frac{1}{2}\mathbb{E}|\widehat{b}_1 - b_1| + \frac{3}{2}\mathbb{E}|\widehat{b}_2 - b_2| \le \frac{L}{8} + \frac{1}{\sqrt{M}}.$$

Clipping the final estimate cannot increase loss. The bound also holds for heterogeneous responses with the same slope budget and arbitrary conditional score CDFs; independence across ranks is unnecessary. The choice $P = u^2$ gives the block-winner midpoint guarantee $L/6 + 1/(2\sqrt{M})$. The first guarantee is smaller for $M > 144/L^2$; at $L = 1, M = 1024$ the values are .15625 and .18229. Both use the same $2M$ candidates and labels without score sharing. This compares guarantees, not pointwise risks or finite-sample minimax rates.

A.7 A STRICT FINITE-TARGET BENEFIT FROM JOINT LABELS

The slope-constrained curve law compares orders of ambiguity, not equality at a particular target. Here there is no slope restriction: let $W(B, K)$ be the largest θ_K gap between homogeneous nondecreasing $[0, 1]$-valued responses with the same depth-B word law, and put $E_B(g) = \inf_{P \in \mathcal{P}_B} \|g - P\|_\infty$.

Corollary A.4 (Strict improvement within the monotone class)**.** *For $K > 2$, $W(2, K) = 2E_2(u^K)$. At depth four,*

$$0.02921638 < W(4, 16) < 2E_4(u^{16}), \qquad 2E_4(u^{16}) \approx 0.15526963. \tag{49}$$

The last decimal approximates the marginal width, not $W(4, 16)$.

Proof. The marginal benchmark. Represent a nondecreasing response almost everywhere by a probability CDF $\eta(u) = \mu([0, u])$, putting missing mass at 1. Fubini gives $\theta_q = \int (1 - t^q)\, d\mu(t)$, so equal marginals through B are equivalent to degree-B moment matching. Their largest target gap is $2E_B(u^K)$: subtracting a polynomial gives the upper bound; a norming signed measure of total variation one annihilating $\mathcal{P}_B$ has positive and negative masses $1/2$, and doubling its Jordan parts gives attaining probability measures. This is the dual characterization of uniform approximation on $C[0, 1]$. For $m_j = \int u^j \eta(u) du$, the depth-two probabilities in order $00, 01, 10, 11$ are

$$((1 - m_0)^2,\ 2m_1 - m_0^2,\ 2(m_0 - m_1) - m_0^2,\ m_0^2).$$

Thus words and marginals impose exactly the same constraint at depth two. At depth four we give a nonzero exact collision, then exclude the unique marginal extremizer using one joint coordinate. Compactness converts that exclusion into the strict upper bound.

An algebraic depth-four collision. The following rational recipe specifies the witness without large expanded polynomial coefficients. For a response with heights $v \in \mathbb{R}^{20}$ on the intervals $((j - 1)/20, j/20)$, define the word polynomials by

$$p_w(v) = 4![x_{w_1} \cdots x_{w_4}] \prod_{j=1}^{20} \exp\left(\frac{(1 - v_j)x_0 + v_j x_1}{20} \right), \qquad w \in \{0, 1\}^4,$$

where the letters are noncommuting, the product is ordered with j increasing, w_1 indexes the lowest order statistic, and only degree four is needed. Write $F_S(v) = \sum_{w:\, w_i = 1\ (i \in S)} p_w(v)$. Let c have the following entries, read row by row and divided by 10^6:

$$\begin{pmatrix} 21346 & 22346 & 135902 & 136902 & 329161 \\ 330161 & 331161 & 465090 & 529333 & 530333 \\ 531333 & 688310 & 689310 & 692853 & 708091 \\ 709091 & 807679 & 810757 & 811757 & 934673 \end{pmatrix}.$$

Form the rational 10×20 matrix A with rows $\nabla F_S(c)$ for the four singleton and six two-element subsets of $\{1, 2, 3, 4\}$. Then $\operatorname{rank} A = 6$, and the reduced row echelon form has pivots in columns $1, \ldots, 6$ with its remaining four rows zero. For $1 \le j \le 14$, let z_j be the unique vector satisfying $Az_j = 0$ and $(z_j)_{6+k} = \mathbf{1}\{j = k\}$ for $1 \le k \le 14$; set $\xi_j = z_j / \|z_j\|_\infty$, so that the pinned coordinate of ξ_j is $1/\|z_j\|_\infty$ rather than one. Define $e = \sum_j \alpha_j \xi_j$ and $d = \sum_j \beta_j \xi_j$ by

j	α_j	β_j	j	α_j	β_j
1	9435958/8527369	−1	8	811517222/8314525	−1
2	−880622/573815	−1	9	255742452/2443685	−1
3	−78079434/6849349	1	10	1313247541/9091941	−1
4	−213201345/9103301	2	11	−217479927/886163	2
5	−402865509/9481436	2	12	−1892551861/5659549	1
6	186117398/4230513	1	13	−2773020059/6442736	−2
7	382812441/5577158	1	14	639700036/1017455	−2

.

For an indeterminate t, put $v_\pm(t) = c\pm(e+td)$ and $Q(t) = p_{0001}(v_+(t)) - p_{0001}(v_-(t))$. Expansion over the rationals gives a cubic Q and, in lexicographic order,

$$\big(p_w(v_+(t)) - p_w(v_-(t))\big)_w = Q(t)(0,1,-3,2,3,-4,0,1,-1,0,4,-3,-2,3,-1,0).$$

This is a polynomial identity for the displayed rational construction, not a small-residual comparison. Consequently the depth-four total variation is $14|Q(t)|$, and vanishes exactly where Q does. Set $a = 0.0004793916368534$ and $b = 0.0004793916368535$, interpreted as exact rationals. Rational evaluation gives $Q(a) > 0 > Q(b)$ and $-1.264 \cdot 10^{-6} < Q'(t) < -1.263 \cdot 10^{-6}$ on $[a,b]$. Thus there is a unique root $t_* \in (a,b)$ and all sixteen word probabilities agree at t_*. Each height $v_{\pm,j}$ is affine in t, so its range over $[a,b]$ is spanned by the endpoint values, and evaluating those gives, for both signs and every applicable j,

$$6 \cdot 10^{-5} < v_{\pm,j}(t) < 1 - 10^{-7}, \qquad v_{\pm,j+1}(t) - v_{\pm,j}(t) > 2 \cdot 10^{-4} \quad (t \in [a,b]).$$

Hence both responses are admissible and nondecreasing at t_*. Their deployment gap is

$$G(t_*) = 2\sum_{j=1}^{20}(e_j + t_* d_j)\frac{j^{16} - (j-1)^{16}}{20^{16}}.$$

The same rational interval evaluation yields $0.02921638735166 < G(t_*) < 0.02921638735168$, proving the lower bound in (49).

Certifying the unique marginal maximizer. The following rational enclosure supplies the strict upper bound; it does not optimize the word-constrained width. Write $P(t) = \sum_{j=0}^{4} a_j t^j$, set $t_0 = 0, t_5 = 1$, and let $z = (a_0, \dots, a_4, E, t_1, \dots, t_4)$. Define ten polynomial equations by

$$G_i(z) = P(t_i) - t_i^{16} - (-1)^i E \quad (0 \le i \le 5), \qquad G_{5+i}(z) = P'(t_i) - 16t_i^{15} \quad (1 \le i \le 4).$$

Use the following exact decimal rationals as the centre z_0:

a_0	0.077634815763693	E	0.077634815763693
a_1	−2.882600644661631	t_1	0.126240468812194
a_2	16.727938476312622	t_2	0.431572750685938
a_3	−31.091149033544870	t_3	0.744925588531164
a_4	18.090541570366492	t_4	0.938556436730422

Let $\mathcal{B} = z_0 + [-h,h]^{10}$, $h = 10^{-9}$, and $C = DG(z_0)^{-1}$, computed over the rationals. Ordinary rational interval evaluation of the displayed polynomials and their derivatives gives

$$\|CG(z_0)\|_\infty < 10^{-12}, \qquad \sup_{z \in \mathcal{B}} \|I - CDG(z)\|_\infty < 10^{-2}. \tag{50}$$

Here the matrix norm is the maximum absolute row sum. These bounds use interval addition, multiplication by endpoint products, and reciprocal intervals not containing zero; no floating-point residual is assumed zero. The map $z \mapsto z - CG(z)$ is a contraction of $\mathcal{B}$ into itself, since $10^{-12} + 10^{-2}h < h$. Hence G has a unique zero in $\mathcal{B}$, with $E > 0$ and $0 < t_1 < \cdots < t_4 < 1$.

The residual $r = P - t^{16}$ has exactly these four interior critical points: the fourth derivative of r' is strictly negative on $(0,1)$, so Rolle's theorem excludes five distinct zeros of r'. Consequently $|r| \le E$, with equality only at the six alternating nodes $t_0, \dots, t_5$. Define barycentric coefficients and positive masses by

$$b_i = \prod_{j \ne i}(t_i - t_j)^{-1}, \qquad \lambda_i = \frac{|b_i|}{\sum_{j \text{ even}} |b_j|}.$$

The vector (b_i) annihilates degrees zero through four. Its alternating signs imply that the even and odd masses each sum to one. Let $\mu_+ = \sum_{i \text{ even}} \lambda_i \delta_{t_i}$ and $\mu_- = \sum_{i \text{ odd}} \lambda_i \delta_{t_i}$. Then $\int r\, d(\mu_+ - \mu_-) = 2E$, proving that P is a best approximant. Moreover, any pair attaining the marginal width must place its two probability measures on these opposite contact sets: otherwise its residual pairing is strictly below $2E$. In measure coordinates the matched depth-four marginals pin $m_0, \dots, m_3$, which is the four-dimensional space $V = \{p \in \mathcal{P}_4 : p(1) = 0\}$ rather than all of $\mathcal{P}_4$. Every element of V vanishes at $s = 1$, and so does the objective $1 - s^{16}$, so mass placed at $t_5 = 1$ is unconstrained and alters no integral. The four V-constraints together with $\mu_+([0,1]) = 1$ give five conditions on

the five masses at $t_0, \dots, t_4$, which forces them uniquely. Thus the marginal-maximizing response pair is unique almost everywhere, the mass at t_5 being immaterial.

A joint coordinate excludes that pair. For $\eta_\pm(u) = \mu_\pm([0,u])$, let $Y(\eta) = \int_{s<t} \eta(s)\eta(t)(t-s)\,ds\,dt$. If $d_i = t_{i+1} - t_i$, $c_i = (t_i + t_{i+1})/2$ and q_i is the response height on that interval, its exact value is

$$Y = \sum_{i=0}^{4} q_i^2 d_i^3/6 + \sum_{0 \le i < j \le 4} q_i q_j d_i d_j (c_j - c_i).$$

Using the same node intervals above, barycentric masses, and interval arithmetic gives the strictly positive enclosure

$$69 \cdot 10^{-6} < Y(\eta_+) - Y(\eta_-) < 71 \cdot 10^{-6}. \tag{51}$$

Indeed $p_3(101) = 6Y - m_0^3$, and the two m_0 agree. Deleting a uniformly chosen position from a depth-four word gives the depth-three law, so this pair cannot have equal depth-four laws. Finally, bounded nondecreasing responses are sequentially compact in L^1 by Helly selection, and word probabilities and θ_{16} are continuous by dominated convergence. The word-matched subset is closed, hence its maximum is attained. Since it excludes the unique marginal maximizer, its width is strictly smaller. □

The certificate proves neither the exact value of $W(4,16)$ nor strict improvement at every depth $B \ge 3$. It is an exact-law, homogeneous finite-target result, not a finite-sample minimax bound or a monotone complete-word horizon law.

B Prediction under exact and approximate score sharing

Appendix B.1 proves recovery under exact sharing; Appendix B.2 characterizes the pooling bias when sharing fails. Appendices B.3–B.6 connect full-data moment concealment, approximation, and observed-data correction to the joint risk bound. Appendix B.7 derives range and precision; Appendix B.8 retains the distinct fixed-depth integrated-pilot proof.

B.1 Proof of Theorem 5.1: shared unknown CDF

The subset weights are from Nakano et al. (2021, Appendix I). We prove their cross-prompt validity and the two sampling floors.

U-statistic variance. For an iid sample of size N and a symmetric $[0,1]$-valued kernel h of order $K \le N$, Hoeffding's canonical decomposition (Hoeffding, 1948) gives

$$\mathrm{Var}(h) = \sum_{r=1}^{K} \binom{K}{r} \sigma_r^2, \qquad \mathrm{Var}(U_N) = \sum_{r=1}^{K} \frac{\binom{K}{r}^2}{\binom{N}{r}} \sigma_r^2.$$

Canonical kernels integrate to zero in each argument, making distinct index sets orthogonal. Since $\binom{K}{r}/\binom{N}{r} \le K/N$ for every $r \ge 1$, including growing K,

$$\mathrm{Var}(U_N) \le (K/N)\,\mathrm{Var}(h) \le K/(4N). \tag{52}$$

Unbiasedness and dependence. A shared atomless CDF makes $U_{ib} = F(S_{ib})$ iid uniform and independent of all prompt states. Score and percentile order agree almost surely, including with flat CDF intervals. Rank j wins $\binom{j-1}{K-1}$ subsets. Conditional on a subset's ranks, averaging prompt states gives winner mean $\bar\eta(\max U)$, even when prompt identities repeat within the subset; integration proves unbiasedness for the original target.

An exact covariance comparison. Set $w_j = 0$ for $j < K$, $a_{N,K} = \sum_j w_j^2$, and $c_{M,B} = (B-1)/(N-1)$ for $N > 1$. Conditional on ordered ranks Q, the assignment to the M named prompt blocks is uniform, balanced and independent of Q and the prompt states. Distinct ranks share a block with probability $c_{M,B}$. Writing $C_\eta(u,v) = \mathrm{Cov}_X(\eta_X(u), \eta_X(v))$, conditional independence gives

$$\mathbb{E}[G_{(r)} G_{(s)} \mid Q] = \bar\eta(Q_r)\bar\eta(Q_s) + c_{M,B} C_\eta(Q_r, Q_s), \qquad r \ne s.$$

Compare with T_{iid}, the same weighted statistic of iid rank–label pairs with Bernoulli response $\bar{\eta}$. The means and diagonal second moments coincide, giving (14). As a complete subset-winner U-statistic, T_{iid} has variance at most $K/(4N)$ by (52). Since $C_\eta \le 1/4$ and $2\sum_{r<s} w_r w_s = 1 - a_{N,K}$, this proves the first inequality in (12); $c_{M,B} \le 1/M$ proves the second. The covariance need not have a constant value or positive sign. For $N = 1$ use the single-label bound; for $K > N$ use $1/2$.

Matching floors. For known uniform scores and one prompt type, take $\eta_\pm = 1/2 \pm t\mathbf{1}_{[1-1/K,1]}$ with $t = \min\{1, \sqrt{K/N}\}/16$. The target gap is at least $2t(1 - e^{-1})$. Bernoulli KL is bounded by $(64/3)t^2$ on a tail of mass $1/K$, giving full-data KL at most $1/12$ and TV below $1/4$. The absolute-loss common-mass bound yields $c_0 \min\{1, \sqrt{K/N}\}$, where $c_0 = (1 - e^{-1})/64$.

For the prompt floor, let the observed state be $X \sim \mathrm{Bernoulli}(1/2 \pm 1/(16\sqrt{M}))$, all labels equal X and scores be independent uniforms. Labels are conditionally independent despite being degenerate. The data augment only M Bernoulli bits: KL is at most $1/12$, giving risk at least $3/(64\sqrt{M})$. Combining the floors with the upper bound proves

$$\frac{c_0}{\sqrt{2}} \min\{1, \sqrt{M^{-1} + K/N}\} \le R^{\mathrm{shared}}_{B,M}(K) \le \min\{1, \sqrt{M^{-1} + K/N}\}. \tag{53}$$

With a common response, the rank–label pairs are iid: (52) gives $K/(4N)$, matched by the singleton floor. All guarantees are pointwise in K.

For the resource example $B = 4, M = 1000, K = 16$, the exact-law floor is $\Delta_{4,16}/2 = .3004086018\ldots$. Standard-deviation bounds are $.0316227766\ldots$ for a common response and $.0353553391\ldots$ for heterogeneous responses; retaining $a_{N,K}$ tightens the latter to $.0344551733\ldots$. Direct evaluation uses 250 independent winner labels with bound $1/(2\sqrt{250})$. These are bounds, not measured errors or a model-independent risk ordering. Direct acquisition's 250 depth-16 winners match the first bound with sixteen times fewer labels, not risk dominance. Indeed, common-response pooling Rao–Blackwellizes random-partition winners: at $N = 6, K = 2, \eta \equiv 1/2$, its MAE is $49/240 < 1/4$. Conversely, $B = 4, M = 3, K = 6$, independent uniform scores and each prompt's fair bit as every label give pooled MAE $7/22 > 1/4$ for direct acquisition. Neither method uniformly dominates.

B.2 When cross-prompt pooling preserves the deployment target

Fix the iid prompt law and atomless conditional score CDFs F_x. Write $H = \mathbb{E}F_X$, $p = F_X(S)$ and $h = H(S)$. For a measurable response $g_x(s) = \mathbb{E}[G \mid X = x, S = s] \in [0, 1]$, let θ^{mix}_K be winning correctness when each of the K candidates draws a fresh independent prompt. This differs from the deployment target θ_K, whose candidates share one prompt, and from the finite pooled statistic considered below. Define

$$\mathcal{D}_{\mathrm{cal}} = \mathbb{E}_X \int (F_X(s) - H(s))^2 \, dH(s). \tag{54}$$

This integrated conditional-CDF variance is classical: with a real-valued encoding preserving the prompt's sigma-field and continuous scores, $6\mathcal{D}_{\mathrm{cal}}$ is the population functional of Chatterjee (2021, Theorem 1.1, Eq. (1.2)), who also notes its earlier antecedents. Our result concerns its budget-dependent consequence for selection, not a new dependence measure.

Proposition B.1 (Response-uniform pooling bias)**.** *For every fixed score family and integer $K \ge 2$,*

$$\Xi_K := \sup_{0 \le g_x \le 1} |\theta^{\mathrm{mix}}_K - \theta_K| = \frac{K}{2} \mathbb{E}|h^{K-1} - p^{K-1}|, \tag{55}$$

$$\Xi_K \le \min\left\{1, \frac{K(K-1)}{2\sqrt{2K-3}} \sqrt{\mathcal{D}_{\mathrm{cal}}}\right\}. \tag{56}$$

Moreover, $\Xi_K = 0$ if and only if $F_X = H$ almost surely. Necessity is uniform over responses, not a claim about each particular response. For every $K \ge 4$ and $0 < d \le 1/(6K^3)$, two atomless prompt types attain $\mathcal{D}_{\mathrm{cal}} = d$ and

$$\Xi_K \ge c_0 K^{3/2} \sqrt{d}, \qquad c_0 = \frac{3e^{-2}}{8} \sqrt{\frac{3}{2}}. \tag{57}$$

Thus the budget dependence in (56) *is sharp up to universal factors in this range. At* $K = 1$ *both targets always agree.*

Proof. Both p and h are marginally uniform, but need not be independent. Exchangeability gives $\theta_K = K\mathbb{E}[Gp^{K-1}]$ and $\theta_K^{\mathrm{mix}} = K\mathbb{E}[Gh^{K-1}]$. Their difference has a centered integrand. Maximizing over $g \in [0,1]$ takes its positive part, attained by $g_x(s) = \mathbf{1}\{H(s) > F_x(s)\}$, proving (55). It is the total variation between the two winning-candidate laws on (X, S) and hence at most one.

For each x, $q_x = F_x - H$ is continuous of bounded variation and zero at both infinite endpoints. The Stieltjes chain rule gives $\int q_x^2\, d(F_x - H) = [q_x^3]/3 = 0$, so

$$\mathbb{E}(h-p)^2 = \mathbb{E}_X \int (F_x - H)^2\, dF_x = \mathcal{D}_{\mathrm{cal}}. \tag{58}$$

Put $k = K - 1$. Since $h^k - p^k = (h-p)k\int_0^1((1-t)p + th)^{k-1}dt$, Jensen and convexity give

$$\mathbb{E}\left[\left(k\int_0^1((1-t)p+th)^{k-1}dt\right)^2\right] \le k^2\int_0^1 \mathbb{E}[((1-t)p+th)^{2k-2}]dt \le \frac{k^2}{2k-1}.$$

The integrands are one for $k = 1$. Cauchy–Schwarz and (58) prove (56). If $\Xi_K = 0$, strict increase of u^{K-1} implies $p = h$ almost surely, hence $\mathcal{D}_{\mathrm{cal}} = 0$. For almost every x, continuity then gives $F_x = H$ on the support of H: a strict difference at a support point would persist on a neighborhood of positive H-mass. Across each gap, monotonicity and equality at its endpoints force the same constant value; the tails follow likewise. Thus $F_x = H$ everywhere for almost every x. The converse follows directly.

For sharpness let $w = 1/K$ and let b be the triangular hat of height one on $[1-2w, 1]$, peaking at $1 - w$. With $0 < a \le w/2$, take equiprobable CDFs $F_\pm(s) = s \pm ab(s)$ on $[0,1]$, extended by zero and one. Their densities are at least $1 - a/w \ge 1/2$, $H(s) = s$, and $\mathcal{D}_{\mathrm{cal}} = a^2\int b^2 = 2a^2/(3K)$. Set $G = 1$ on the minus type and $G = 0$ on the plus type. This attains (55) and has $\theta_K = 1/2$. Integration by parts gives

$$\Xi_K = \theta_K^{\mathrm{mix}} - \frac{1}{2} = \frac{K(K-1)a}{2}\int_0^1 s^{K-2}b(s)\, ds \ge \frac{3e^{-2}}{8}Ka.$$

Indeed, $s^{K-2} \ge (1-2/K)^{K-2} \ge e^{-2}$ on the hat's support, by $\log(1-z) \ge -z/(1-z)$; also $\int b = 1/K$ and $K - 1 \ge 3K/4$. Taking $a = \sqrt{3Kd/2}$ proves (57); its allowed range is exactly $a \le 1/(2K)$. □

Fixing $a = \alpha/K$, $0 < \alpha \le 1/2$, gives $\mathcal{D}_{\mathrm{cal}} = 2\alpha^2/(3K^3) \to 0$ but nonvanishing pooling bias. This is not a minimax lower bound: the prompt-label mean is unbiased in this example. Nor does (56) uniformly improve the $O(K\delta)$ bound under a declared uniform radius.

Observing average discrepancy. Draw S_1, S_2 conditionally iid at one random prompt and an independent S_0 with marginal CDF H. Then

$$\mathcal{D}_{\mathrm{cal}} = \Pr\{S_0 > \max(S_1, S_2)\} - \frac{1}{3}, \qquad 0 \le \mathcal{D}_{\mathrm{cal}} \le \frac{1}{6}. \tag{59}$$

The probability equals $\mathbb{E}_X\int F_x^2\, dH$, while $\int H^2 dH = 1/3$. The upper bound uses $\mathrm{Var}_X(F_X(s)) \le H(s)(1-H(s))$ and $\int H(1-H)dH = 1/6$. For n disjoint triples from $2n$ independent prompt draws, set $Y_i = \mathbf{1}\{S_{i,0} > \max(S_{i,1}, S_{i,2})\}$ and $\widehat{\mathcal{D}}_{\mathrm{cal}} = n^{-1}\sum_i Y_i - 1/3$. For $0 < \alpha < 1$,

$$\Pr\left\{|\widehat{\mathcal{D}}_{\mathrm{cal}} - \mathcal{D}_{\mathrm{cal}}| > \sqrt{\frac{\log(2/\alpha)}{2n}}\right\} \le \alpha. \tag{60}$$

For completeness, the second derivative of $\log \mathbb{E}e^{tY}$ for $Y \in [0,1]$ is a tilted variance at most $1/4$. Integrating twice bounds the centered moment-generating function by $e^{t^2/8}$. Independence and exponential Markov with $t = 4\epsilon$ give the two-sided bound $2e^{-2n\epsilon^2}$ and hence (60). This statistic reads $3n$ scores and no labels; uniform depth-two acquisition charges $4n$ candidates and, under our full-label accounting, $4n$ labels.

Average discrepancy does not certify a uniform radius. Under the declared-radius model,

$$\mathcal{D}_{\mathrm{cal}} = \int \mathrm{Var}_X(F_X(s))\, dH(s) \le \mathbb{E}_X \int (F_X - F_0)^2\, dH \le \delta^2. \tag{61}$$

A positive lower confidence bound on $\mathcal{D}_{\mathrm{cal}}$ therefore lower-bounds δ, but a small average does not upper-bound it. For example, $F_A(s) = s$ and $F_B(s) = s + s(1-s)$ on $[0,1]$, with probabilities $1-\varepsilon, \varepsilon$, give

$$\mathcal{D}_{\mathrm{cal}} = \frac{\varepsilon(1-\varepsilon)}{30} \to 0, \qquad \inf_{F_0} \operatorname*{ess\,sup}_x \|F_x - F_0\|_\infty = \frac{1}{8} \quad (0 < \varepsilon < 1).$$

For the first identity integrate $(s(1-s))^2$ against $dH = d[s + \varepsilon s(1-s)]$; the derivative term is a zero endpoint term. For the second, $\|F_A - F_B\|_\infty = 1/4$, the triangle inequality gives the lower bound, and $(F_A + F_B)/2$ attains it. At $n = 48$, the 95% radius in (60) is about .196, larger than the full possible range of $\mathcal{D}_{\mathrm{cal}}$. The identity therefore gives no claim of useful precision at the current natural-bank size. Reusing prompt draws across triples is not covered by this independence proof; neither are transformations fitted on the same evaluation observations. Ties are outside the atomless assumptions of the proposition.

A uniform radius can nevertheless be bounded below directly. If each of M prompts has n iid responses, the Dvoretzky–Kiefer–Wolfowitz inequality with Massart's constant and a Bonferroni correction places every empirical CDF within $\epsilon = \{\log(2M/\alpha)/(2n)\}^{1/2}$ of its CDF with probability at least $1-\alpha$ (Massart, 1990). Two prompts whose empirical CDFs differ by D then have CDFs at least $D - 2\epsilon$ apart, and the triangle inequality gives $\delta \ge (D - 2\epsilon)/2$ for every reference F_0. This bound needs no labels and holds for ties; Appendices D.3.1 and D.3.7 apply it.

Relating the mixed target to the finite pooled record. For $K \le MB$, a uniformly sampled K-subset repeats a prompt block with probability

$$q_{M,B,K} = \begin{cases} 1 - \binom{M}{K} B^K / \binom{MB}{K}, & K \le M, \\ 1, & K > M. \end{cases} \tag{62}$$

Conditional on distinct blocks, its score–label pairs are iid from the marginal law. On the complement, the winning label is in $[0,1]$. Consequently

$$|\mathbb{E}\widehat{\theta}_K - \theta_K| \le \Xi_K + q_{M,B,K}, \qquad q_{M,B,K} \le \min\left\{1, \binom{K}{2} \frac{B-1}{MB-1}\right\} \tag{63}$$

for $MB > 1$; the latter is the pairwise union bound. Thus the limiting mean for fixed B, K and $M \to \infty$ is θ_K^{mix}, not in general the same-prompt target. Exact sharing is sharper: Theorem 5.1 gives zero bias even when $q_{M,B,K} > 0$.

B.3 Approximate sharing: common reductions

Proof strategy for Theorem 5.2. There are two observation-model obstacles. For recovery, the estimator must use an estimated marginal CDF, not unobserved conditional ranks, and account for dependent anchor comparisons. Only the analysis recenters at the true conditional CDF: the resulting order-$(r+1)$ kernels have projections only at orders $r, r+1$, retaining the depth factor $(B)_r$. For the converse, concealment must survive numerical scores and observed prompt identities. In the two-band submodel, positions conditional on band counts are parameter-free; moments through degree B determine the count law. Functional-estimation duality applies to this submodel (Polyanskiy & Wu, 2021). Freezing a large finite prompt population then transfers its lower bound to the stated observation model. The following reductions establish this transfer before the estimator and risk assembly.

For $\mathfrak{P}_\delta$, the unknown F_0 specifies comparability but is not supplied to the estimator. The observable marginal CDF $H = \mathbb{E}F_X$ instead satisfies $|F_x - H| \le 2\delta$. Conditional ranks $p = F_X(S)$ are uniform, and $\theta_K = K\mathbb{E}[Gp^{K-1}]$.

Moment truncation, high-order correction, and monomial approximation are established tools (Klaassen & Mnatsakanov, 2020; Orlitsky et al., 2016; Robins et al., 2008; Newman & Rivlin, 1976; Saibaba, 2021). The proof here must implement the correction without conditional CDFs and retain pool depth in its variance. We give the common reductions before the approximation, construction, and risk assembly.

Pooling bias is not an information lower bound. The true ranks are iid uniform independently of prompts. The pooled statistic ordered by those unobserved ranks, $\widetilde{\theta}_K$, obeys the shared-CDF variance bound. Under $\|F_x - F_0\|_\infty \le \delta$, numerical and true-rank winners agree whenever the top two true ranks differ by more than 2δ. For $0 \le e \le 1$, that event has probability $K\int_e^1 (u-e)^{K-1}du = (1-e)^K$. Averaging the bounded winner discrepancies gives

$$\mathbb{E}|\widehat{\theta}_K - \theta_K| \le \tfrac{1}{2}\sqrt{M^{-1} + K/N} + \beta_K(\delta), \quad \beta_1 = 0, \quad \beta_K = 1-(1-2\delta)_+^K \le \min\{1, 2K\delta\}. \tag{64}$$

The last expression applies to $2 \le K \le N$; continuity of F_0, not strict increase, suffices.

For a procedure-specific sharp example, take $B = 1$, equiprobable $X \in \{0,1\}$, $G = X$, and scores uniform on $[0, 1-\delta]$ and $[0,1]$, respectively. The true same-prompt target is $1/2$. Below $1-\delta$ the mixture CDF is $s(2-\delta)/(2(1-\delta))$. Integrating its winning type gives

$$\mathbb{E}\widehat{\theta}_K - \theta_K = \tfrac{1}{2}[1 - (1-\delta/2)^{K-1}] = \Theta(\min\{1, (K-1)\delta\}). \tag{65}$$

This is pooling bias, not minimax risk: averaging prompt labels is unbiased in this example. Implementation of pooling uses one sort and the recurrence $w_N = K/N$, $w_{j-1} = w_j(j-K)/(j-1)$; it uses neither calibration nor the CDF.

Observed identities in lower bounds. Suppose priors $\nu_\pm$ on valid conditional models, all obeying one reference-CDF bound, induce one-prompt laws $Q_\pm$ and target means with gap Γ. Assign J named categories iid parameters from either prior once, freeze the assignment, then sample categories uniformly. Conditional on no repeated identity, the predictive data have law $Q_\pm^{\otimes M}$. The repeat probability is at most $\binom{M}{2}/J$, and the frozen population target differs from its prior mean in expected absolute value by at most $1/(2\sqrt{J})$. The common-mass bound therefore gives

$$R_\delta \ge \frac{\Gamma}{2}\left[1 - \mathrm{TV}(Q_+^{\otimes M}, Q_-^{\otimes M}) - \frac{2\binom{M}{2}}{J}\right] - \frac{1}{2\sqrt{J}}. \tag{66}$$

Taking $J \to \infty$ at each fixed finite experiment yields

$$R_\delta \ge \frac{\Gamma}{2}[1 - \mathrm{TV}(Q_+^{\otimes M}, Q_-^{\otimes M})]. \tag{67}$$

This is a Bayes lower bound over deterministic frozen models with identities observed, not a hidden parameter resampled when a known prompt reappears.

Moment concealment with full numerical scores. Scores uniform on $(0, 1/2)$ and $(1/2, 1)$, with masses $1-p, p$ and label $G = \mathbf{1}\{S > 1/2\}$, have deployment target $g_K(p) = 1-(1-p)^K$. For $p \in [0, A]$, their CDFs are within $A/2$ of $F_{A/2}$ and their densities are bounded by two. Given the upper-band count, all positions, labels, and sample indices have a parameter-free law. Thus matching moments below degree h on $[0, A]$, $A \le 2\delta$, makes the full numerical laws identical when $h = B + 1$. Otherwise,

$$\mathrm{TV}(Q_+, Q_-) \le 2\sum_{j=h}^{B}(2eBA/j)^j \le \frac{2q^h}{1-q}, \qquad q = 2eBA/h < 1. \tag{68}$$

For $C \sim \mathrm{Bin}(B, p)$ and $|\varphi| \le 1$, the Newton expansion is $\mathbb{E}_p\varphi(C) = \sum_{j=0}^B \binom{B}{j}\Delta^j\varphi(0)p^j$, with $|\Delta^j\varphi(0)| \le 2^j$. Matched terms vanish, and $\binom{B}{j} \le (eB/j)^j$ gives the bound.

Moment duality turns a best uniform approximation error e_h into a target gap $2e_h$: a norm-one Hahn–Banach annihilating measure has positive and negative masses $1/2$ because constants are annihilated. For the complete fixed-error range we also retain the explicit implication

$$KA \ge 64h^2 \quad \Longrightarrow \quad \Gamma \ge 1/2. \tag{69}$$

Otherwise a degree-$(h-1)$ approximant to $(1-Ax)^K$ with error below $1/4$ has norm at most $5/4$ and changes by at least $3/8$ between 0 and $\log 8/(KA)$. Its derivative is at least $3KA/(8\log 8)$, contradicting Markov's bound $\frac{5}{2}(h-1)^2$. This includes $h = 1$.

B.4 THE RESIDUAL SHARED BY RECOVERY AND THE CONVERSE

Write $L = \log(eM)$, $\tau = B\delta$, $a = K\delta$, $\Phi(s, \lambda) = s\log(1 + s/\lambda)$, and $E_* = \Phi(D, a)$. Constants in the following proof are universal. The first lemma controls an integrated residual for the estimator; the second controls a uniform residual for moment duality.

Lemma B.2 (Interval residual and exponential envelope)**.** *Let $d, K \geq 1$, $0 < A \leq 1$, $N_A \sim \mathrm{Bin}(K, A)$, and let S_K be a walk with independent increments taking $(-1, 0, 1)$ with probabilities $(A/4, 1 - A/2, A/4)$. Then*

$$\mathcal{D}_{d,K}(A) := \inf_{\deg P < d} \int_{1-A}^{1} |f_K - P| = \mathbb{E}\Delta_{d,N_A}, \tag{70}$$

$$= 2(d+1) \sum_{m \in \mathbb{Z}} \Pr\{S_K = (2m+1)(d+1)\}, \tag{71}$$

$$\mathcal{D}_{d,K}(A) \leq \min\{1, 2\pi\sqrt{KA}/d, 8e^{-\Phi(d,KA)/8}\}. \tag{72}$$

Proof. Expand $Af_K(1 - A + Ax) = \sum_{j=1}^{K} \Pr(N_A = j) f_j(x)$. Interpolation at the d Uhrig nodes (Uhrig, 2007) reproduces $j \leq d$, and every other residual has the same sign h_d. Their positive mixture attains the triangle upper bound and its dual certificate, proving the exact identity. The bound $\Delta_{d,j} \leq 2\pi\sqrt{j}/d$ and Jensen give the square-root envelope. The odd-residue formula (34), also valid for $j \leq d$ with both sides zero, expresses $\Delta_{d,j}$ through a centered $\mathrm{Bin}(2j, 1/2)$ walk. Independently thinning its increments gives (71), including $N_A = 0$.

The increment mass and its convolution are symmetric and unimodal. To verify preservation, express a symmetric unimodal sequence as a positive mixture of uniforms on symmetric integer intervals; counting overlaps proves the convolution claim. Put $s = d + 1$. For each positive odd multiple k of s, the interval $[k - \lfloor s/2 \rfloor, k]$ contains at least $s/2$ masses no smaller than the mass at k. These intervals are disjoint, so $\mathcal{D}_{d,K}(A) \leq 4\Pr(|S_K| \geq \lceil s/2 \rceil)$.

For $\lambda = KA$ and $t = \log(1 + u/\lambda)$,

$$\log \mathbb{E}e^{tS_K} \leq \frac{\lambda}{2}(\cosh t - 1) = \frac{u(u/\lambda)}{4(1 + u/\lambda)} \leq \frac{u}{4}\log(1 + u/\lambda).$$

Chernoff and symmetry give $\Pr(|S_K| \geq u) \leq 2e^{-3\Phi(u,\lambda)/4}$. Since $\Phi(s/2, \lambda) \geq \Phi(d, \lambda)/4$, the exponential bound follows. The zero approximant supplies the bound one. □

Lemma B.3 (Uniform endpoint approximation)**.** *For $0 < A \leq 1/2$ and integers $1 \leq h \leq K$,*

$$\inf_{\deg P < h} \|1 - (1 - p)^K - P(p)\|_{\infty,[0,A]} \geq ce^{-C\Phi(h,KA)}. \tag{73}$$

Proof. Set $p = A(1 - \cos t)/2$. The Fourier coefficients of $(1 - A/2 + (A/2)\cos t)^K$ are $p_z = \Pr(S_K = z) \geq 0$. For the integral-one Fejér kernel F_n, $V = 2F_{2h-1} - F_{h-1}$ has coefficient one below frequency h, coefficients in $[0, 1]$ thereafter, and zero at $|z| \geq 2h$. Its L^1 norm is at most three. Thus $\delta_0 - V(t)dt/(2\pi)$ annihilates degrees below h and has total variation norm at most four. Nonnegative Fourier coefficients give

$$\inf_{\deg P < h} \|(1 - p)^K - P\|_{\infty,[0,A]} \geq \tfrac{1}{4}\Pr(|S_K| \geq 2h). \tag{74}$$

We need the tail estimate

$$\Pr(|S_K| \geq u) \geq ce^{-C\Phi(u,KA)}, \qquad 1 \leq u \leq K. \tag{75}$$

Put $\lambda = KA$. If $u \geq \lambda/8$, exactly u positive increments and no negative increment have probability at least $(\lambda/(4u))^u e^{-2\lambda/3}$, using $\binom{K}{u} \geq (K/u)^u$ and $\log(1 - A/2) \geq -2A/3$. For $x = u/\lambda \geq 1/8$, the ratio $[\log(4x) + 2/(3x)]/\log(1 + x)$ is bounded above, proving the claim.

If $u < \lambda/8$, the number N of nonzero increments has mean $\lambda/2 > 4$. Paley–Zygmund and Markov give $\Pr(\lambda/8 \leq N \leq 4\lambda) \geq 13/40$. Conditional on $N = n > u$, the symmetric walk tail is at least $ce^{-Cu^2/n}$. For $n \leq 4u$ the all-positive event suffices; otherwise sum admissible parity masses

in $[u, u+\sqrt{n}]$, each at least $cn^{-1/2}e^{-C(u^2/n+1)}$ by Stirling. On the stated event $n \geq \lambda/8$, giving $ce^{-Cu^2/\lambda}$, which has the required form because $\Phi(u,\lambda) \asymp u^2/\lambda$ in this branch.

When $2h \leq K$, combine (74)–(75) and $\Phi(2h,\lambda) \leq 4\Phi(h,\lambda)$. Otherwise the approximant has zero Fourier coefficient at K, while $p_K = (A/4)^K$. For $h > K/2$, $K\log(4/A) \leq 6h\log(1+h/(KA))$, using $8z \leq (1+z)^3$ at $z = 1/(2A) \geq 1$. This includes $h = K = 1$. Subtracting one does not change the approximation distance. □

B.5 Constructing correction from observed scores and labels

We use a coarse gate when $K\delta$ is comparable to or larger than D, and centered comparisons when it is smaller. Neither estimator observes true conditional ranks. Split the prompts into $m = \lfloor M/2 \rfloor$ pilot and $n = M - m$ analysis prompts. Both constructions charge all MB candidates and labels. Independence of all pilot scores is used only conditional on prompt states.

B.5.1 Coarse upper-tail localization

For $\ell > 0$ put $t = 4\delta + 2\ell/K$. If $t > 1/4$, set $\widehat{c} = -\infty$ and $A = 1$. Otherwise require $mBt \geq 10$, take $\widehat{c}$ to be pilot order statistic $mB - \lfloor mBt \rfloor$, and put $A = 9t/4$. Reference quantiles $F_0^{-1}(1-2t)$ and $F_0^{-1}(1-t/2)$ have conditional tail probabilities at least $7t/4$ and at most $3t/4$. Conditional Chernoff bounds, using $\lfloor mBt \rfloor \geq .9mBt$, give

$$\Pr(\mathcal{E}^c) \leq 2e^{-mBt/100}, \quad \mathcal{E} = \{t/4 \leq 1 - F_x(\widehat{c}) \leq A \text{ a.s.}\}, \quad A \geq \delta, \quad KA \leq 16(a+\ell). \tag{76}$$

An omitted anchor contributes at most $e^{-Kt/4}$. Flat CDF intervals carry no score mass. If the count requirement fails the midpoint is available; later choices verify it wherever this gate is used.

Turning an interval approximant into an observed statistic. For a best interval approximant write $P(1-Ax) = \sum_{r<d} a_r x^r$, $1 \leq d \leq B$. Its $L^1[0,1]$ norm is at most $2/A$. The shifted Legendre coefficient of degree j is at most $(2j+1)2/A$, and its monomial coefficient norm is at most 7^j: initial norms are 1 and 3, and the recurrence bounds the next by six times its predecessor plus the previous one. Therefore

$$\sum_{r<d} |a_r| \leq 2d^2 7^d / A. \tag{77}$$

Interpolation nodes and the binomial expansion determine these coefficients by a finite linear system, independently of all unknown CDFs. For $C_b = \#\{q \neq b : S_q > S_b\}$, use

$$Z_P = \frac{1}{B}\sum_b G_b \mathbf{1}\{S_b > \widehat{c}\} \sum_{r<d} \frac{a_r}{A^r} \frac{\binom{C_b}{r}}{\binom{B-1}{r}}, \qquad \widehat{\theta} = \Pi_{[0,1]}\left(\frac{1}{n}\sum_i Z_{P,i}\right). \tag{78}$$

Conditioning on the labeled anchor makes C_b binomial with parameter $1 - F_x(S_b)$, so its mean evaluates P at the true rank.

For $k = r+1$, the comparison part is exactly

$$T_k = \frac{1}{B}\sum_b G_b \mathbf{1}\{S_b > \widehat{c}\} \frac{\binom{C_b}{k-1}}{\binom{B-1}{k-1}} = \frac{1}{k\binom{B}{k}} \sum_{|J|=k} G_{\arg\min_{j\in J} S_j} \mathbf{1}\{\min_{j\in J} S_j > \widehat{c}\}. \tag{79}$$

Each subset has one minimum, and $B\binom{B-1}{k-1} = k\binom{B}{k}$.

Count overlaps instead of assuming independent anchors. On $\mathcal{E}$, a pair of subset kernels is bounded by the event that their union exceeds the gate. If Z is their overlap, then conditional on the prompt

$$\mathbb{E}T_k^2 \leq A^{2k}k^{-2}\mathbb{E}A^{-Z}, \qquad \mathbb{E}A^{-Z} = \sum_{l=0}^{k} (A^{-1}-1)^l \binom{k}{l} \frac{(k)_l}{(B)_l} \leq (1 + k/(BA))^k.$$

Here $(k)_l/(B)_l \leq (k/B)^l$ for falling factorials. Minkowski and (77) give

$$\|Z_P\|_2 \leq Q_{d,B,A} := 2d^2 7^d (1 + d/(BA))^{d/2}. \tag{80}$$

The rarity of the anchor and companions cancels the coefficient factor A^{-1}. No independence among anchor statistics is used. Clipping only the final average and charging loss at most one on pilot failure yields

$$R_\delta \le e^{-Kt/4} + \mathcal{D}_{d,K}(A) + Q_{d,B,A}/\sqrt{n} + 2e^{-mBt/100}. \tag{81}$$

In the ungated branch omit the first and last terms and use $A = 1$.

B.5.2 Tail-local estimation and centered comparisons

The pilot needs accuracy only in the winning tail, but its bound must retain both between-prompt and within-prompt sampling.

Lemma B.4 (Tail-local empirical CDF). *Let $\widehat{H}$ use all mB pilot scores and let $s_0 = H^{-1}(1-v)$, $0 < v \le 1$, taking $s_0 = -\infty$ at $v = 1$. Then*

$$\Pr\left\{\sup_{s \ge s_0} |\widehat{H}(s) - H(s)| > \delta\right\} \le C(1 + v/\delta)\left[e^{-cm} + e^{-cmB\delta^2/(v+\delta)}\right]. \tag{82}$$

Proof. At fixed $s \ge s_0$, let $z = 1 - H(s) \le v$. The independent prompt tail probabilities $q_i = 1 - F_{X_i}(s)$ lie in an interval of length 4δ. Their average differs from z by more than $\delta/8$ with probability at most $2e^{-m/512}$. Conditional on the states, the mB indicators are independent with total variance at most $mB(v + 2\delta)$. Bernstein bounds a further deviation $\delta/8$ by $2e^{-cmB\delta^2/(v+\delta)}$. Use an H-quantile grid of mesh $\delta/4$ from $1 - v$ to one, with at most $2 + 4v/\delta$ points including s_0 and $+\infty$. The union bound and monotonicity interpolate error to at most $\delta/2$. Continuity suffices because flat intervals have zero score mass. □

Choose $T > 3a$, let $v = \min(1, T/K + 2\delta)$, and retain anchors with $q(s) = \mathbf{1}\{\widehat{H}(s) > 1 - T/K\}$, or all anchors when $T \ge K$. On the good event of Lemma B.4, retained anchors lie in the controlled tail: if $v < 1$, $\widehat{H}(s_0) \le 1 - T/K - \delta$. Thus

$$|p - h| \le 3\delta \text{ on retained anchors}, \qquad p = F_x(s), \quad h = \widehat{H}(s). \tag{83}$$

Omitted anchors have $p \le 1 - T/K + 3\delta$ and contribute at most e^{-T+3a}, or zero if $T \ge K$.

For any center h_b, define

$$V_{b,\ell}^{h_b} = \binom{B-1}{\ell}^{-1} \sum_{\substack{A \subseteq [B]\setminus\{b\} \\ |A| = \ell}} \prod_{q \in A} (\mathbf{1}\{S_q \le S_b\} - h_b), \qquad V_{b,0}^{h_b} = 1. \tag{84}$$

Given the anchor, conditional independence gives $\mathbb{E}V_{b,\ell}^{h_b} = (p - h_b)^\ell$ and Jensen gives $\mathbb{E}(V_{b,\ell}^{h_b})^2 \le [p(1-p) + (p-h_b)^2]^\ell$. For $1 \le j \le B$, $j \le (K-1)/2$, set

$$Z_j = \frac{1}{B}\sum_b q(S_b) G_b K \sum_{\ell < j} \binom{K-1}{\ell} h_b^{K-1-\ell} V_{b,\ell}^{h_b}, \quad h_b = \widehat{H}(S_b), \quad \widehat{\theta}_j = \Pi_{[0,1]}\left(\frac{1}{n}\sum_i Z_{j,i}\right). \tag{85}$$

Lemma B.5 (Only two conditional Hoeffding orders remain). *Let $K > B$ and $P_j = (2 + \sqrt{j})/(1 - e^{-1/2})$. On a good pilot, uniformly in the prompt state,*

$$|\mathbb{E}[Z_j \mid X, \widehat{H}] - \theta_K(X)| \le b_j := e^{-T+3a} + e^{3a}(3ea/j)^j, \tag{86}$$

$$\mathrm{Var}(Z_j \mid X, \widehat{H}) \le e^{12a} P_j^2 (eK/B)^j. \tag{87}$$

Omit the first bias term when $T \ge K$. If $b_j \le 1$, the risk is at most

$$b_j + \frac{2 + e^{6a} P_j (eK/B)^{j/2}}{\sqrt{n}} + C(1 + v/\delta)\left[e^{-cm} + e^{-cmB\delta^2/(v+\delta)}\right]. \tag{88}$$

The relaxation $e^{12a}(64eK/B)^j$ also follows since $P_j \le 8^j$, but the explicit factor is retained for finite-budget comparisons.

Proof. **Bias.** Comparison means form the Taylor polynomial of KGp^{K-1} around h. Integrating its remainder in the conditionally uniform rank bounds it by $\binom{K}{j}(3\delta)^j(1+3\delta)^{K-j} \le e^{3a}(3ea/j)^j$. Add the omitted contribution.

Recenter the proof, not the estimator. The implemented center remains $\widehat{H}$. In the analysis, let $p(s) = F_x(s)$. The identity

$$V_\ell^h = \sum_{r\le\ell} \binom{\ell}{r}(p-h)^{\ell-r}V_r^p$$

and anchor/subset counting express Z_j as a sum of complete order-$(r+1)$ U-statistics, $0 \le r < j$, with kernels

$$k_r(z_0,\dots,z_r) = \frac{1}{r+1}\sum_{i=0}^{r} G_i w_r(S_i)\prod_{q\ne i}(\mathbf{1}\{S_q \le S_i\} - p(S_i)),$$
$$w_r(s) = q(s)K\binom{K-1}{r}\sum_{t<j-r}\binom{K-1-r}{t}h^{K-1-r-t}(p-h)^t.$$

Since $h + |p-h| \le p + 6\delta$, $K-1-r \ge K/2$, and $u = 1-p$,

$$|w_r(s)| \le e^{6a}K^{r+1}e^{-Ku/2}/r!. \tag{89}$$

The surviving projections. For $r \ge 1$, conditioning on fewer than r observations leaves an unobserved centered companion in every term, even if its anchor is also unobserved. Hence projections below order r vanish. At order r only the unobserved-anchor term remains, with canonical projection

$$g_r(u_1,\dots,u_r) = \frac{1}{r+1}\int_0^1 \eta_x(p)w_r(p)\prod_i(\mathbf{1}\{u_i \le p\} - p)\,dp.$$

Using $\min(p,q) - pq \le \min(1-p, 1-q)$ and (89),

$$\|g_r\|_2^2 \le \frac{e^{12a}K^{2r+2}}{(r+1)^2(r!)^2}\int_0^\infty\int_0^\infty e^{-K(u+v)/2}\min(u,v)^r\,du\,dv = \frac{4e^{12a}K^r}{(r+1)^2r!},$$
$$\mathbb{E}k_r^2 \le e^{12a}K^{r+1}/r!.$$

The second bound uses Jensen across anchors and integrates $w_r(p)^2[p(1-p)]^r$. Only orders $r, r+1$ remain, so the variance identity underlying (52) gives

$$\mathrm{Var}(U_{B,r+1}(k_r)) \le \frac{(r+1)^2\|g_r\|_2^2}{\binom{B}{r}} + \frac{\mathbb{E}k_r^2}{\binom{B}{r+1}} \le e^{12a}\left[\frac{4K^r}{(B)_r} + \frac{(r+1)K^{r+1}}{(B)_{r+1}}\right].$$

For $r = 0$ the bound is $e^{12a}K/B$. Since $(B)_r \ge (B/e)^r$, Minkowski and a geometric sum, with $x = eK/B > e$, bound the total standard deviation by $e^{6a}x^{j/2}(2+\sqrt{j})/(1-e^{-1/2}) \le e^{6a}(64eK/B)^{j/2}$. No orthogonality between distinct r-terms is assumed.

Prompt-conditional means lie in $[-b_j, 1+b_j]$ and hence have variance at most $9/4$ when $b_j \le 1$. Total variance, conditional independence across analysis prompts, and loss bounded by one on pilot failure yield (88). □

B.6 Assembling the error-uniform risk bound

Recall $L = \log(eM)$, $\tau = B\delta$, $a = K\delta$, and $E_* = \Phi(D,a)$. All choices below use only (B, M, K, δ) and pay the pilot cost. We prove the upper and lower bounds of Theorem 5.2.

Upper bound: bounded and dense regimes. For $M < M_0$, enlarge a universal prefactor using $\rho \ge M_0^{-1/2}$; for $K > MB$ use the midpoint. Otherwise split with $m, n \asymp M$.

If $L \le \tau$, elementary bounds on the root equation give $\tau L \le r_*^2 \le 3\tau L$ and $L \le r_* < 2\tau \le B/2$. Thus $D = r_*$, $E_* \asymp BL/K$, and the rank-additive estimator from the sampling bound (28) has risk at most Ce^{-cE_*}. Its upper construction applies to heterogeneous responses because its mean is linear in η_x and its range uniform over words; no homogeneous converse is transferred.

If $D < D_0$, choose $c \le 1/(2D_0)$. For $a \le 1$, $e^{-cD\log(1+D/a)}$ is bounded below by a constant times $\sqrt{a}$; for $a \ge 1$ it is bounded below by a positive constant. The pooling bound $C\rho + 2\min(1, a)$ suffices. For all remaining cases $L > \tau$, $D \ge D_0$, and

$$D > \tau, \quad D\log(1 + D/\tau) \le L, \quad E_* \le L, \quad \tau \ge De^{-L/D}, \quad D \le L/\log 2. \tag{90}$$

Upper bound: centered correction, $a \le \eta D$. Choose a small universal $\gamma > 0$, then $\eta > 0$ such that

$$\frac{6\eta}{\log 2} + \frac{\gamma}{2}\left(1 + \frac{1}{\log 2}\right) < \frac{1}{16},$$
$$\log\frac{\gamma x}{6e} \ge \tfrac{1}{2}\log(1 + x) \quad (x \ge 1/\eta), \qquad 3\eta \le \tfrac{\gamma}{8}\log(1 + 1/\eta).$$

Take $j = \lfloor \gamma D \rfloor$ and $T = 12a + 2E_*$. Increase D_0 so $j \ge \gamma D/2 \ge 1$ and $2e^{-\gamma D_0 \log(1+1/\eta)/8} \le 1$. Since $K > B \ge D$, $j \le B$ and $j \le (K - 1)/2$. The choices give

$$b_j \le e^{-2E_*} + e^{3a - j\log(j/(3ea))} \le 2e^{-\gamma E_*/8} \le 1, \qquad 6a + \log P_j + \frac{j}{2}\log(eK/B) \le L/16.$$

For the latter, use $K/B = a/\tau \le D/\tau$ and (90). Here $\log P_j = O(\log L)$; increasing D_0 absorbs it into the strict slack of the choice of γ, η. The analysis-sample term is $CM^{-7/16} \le Ce^{-cE_*}$.

For the pilot, $v + \delta \le (T + 3a)/K$ and $T + 3a \le CL$, so

$$\frac{mB\delta^2}{v + \delta} \ge \frac{cm(K/B)\tau^2}{L} \ge \frac{cD^2M^{1-2/D}}{L}, \qquad \log(1 + v/\delta) \le C + \log L + L/D \le CL.$$

For $D_0 \ge 8$, both concentration exponents dominate the grid cost and $2\log M$ after a universal increase of M_0. Pilot failure is then at most M^{-2}. Equation (88) completes this regime.

Upper bound: coarse correction, $a > \eta D$. Use $d = \lfloor \gamma_0 D \rfloor$ and $\ell = \alpha E_*$, where $0 < \alpha \le \eta/\log(1 + 1/\eta)$, so $\ell \le a$. For $E_* < 1/\alpha$ the midpoint suffices. Otherwise increase D_0 so $d \ge \gamma_0 D/2$. The gate satisfies $A \ge \delta$ and $KA \le 32a$. Using $\log(1 + cx) \ge c\log(1 + x)$ for $0 \le c \le 1$ gives

$$\Phi(d, KA) \ge \frac{\gamma_0^2}{128}\Phi(D, a).$$

For sufficiently small universal γ_0,

$$\log Q_{d,B,A} \le 4d + \frac{d}{2}\log(1 + d/\tau) \le L/16.$$

The residual is Ce^{-cE_*} and omission at most $e^{-\ell/2}$. In the gated branch $mBt \ge 4m\tau \ge cDM^{1-1/D}$, which verifies the count requirement and makes failure at most M^{-2}. Neither is charged in the ungated branch. Equation (81) proves the claimed upper bound. Comparing every branch with the midpoint gives the cap $1/2$.

One concealment choice for accuracy and horizon. For $K > B$, choose moment order h and band width A. If $L \le \tau$, take $h = \lceil 4L \rceil \le B$ and $A = h/(16eB) \le 2\delta$. Then $q = 1/8$ in (68),

$$M\,\mathrm{TV}(Q_+, Q_-) \le (16/7)M8^{-h} < 1/4, \qquad D^2 \in [\tau L, 3\tau L],$$

and $\Phi(h, KA) = h\log(1 + 16eB/K) \le 80eBL/K \le CE_*$.

If $L > \tau$, put $u = \max(1, r_*)$ and $h_0 = \lceil 64u \rceil$. For $h_0 \le B$, choose $h = h_0$ and $A = 2\delta$. Then $D = u$, $q < 1/4$, and

$$h\log\frac{h}{4e\tau} \ge 64L,$$

because $h/(4e\tau) \ge 16r_*/(e\tau) \ge 1 + r_*/\tau$. Thus the product TV is below $1/4$. For $h_0 > B$, choose $h = B + 1$, $A = 2\delta$, giving identical group laws. In both sparse cases $h \le 65D$ and $h \le K$, so $\Phi(h, KA) \le \Phi(65D, 2a) \le CE_*$. All choices have $A \le 1/2$ and preserve the full numerical observations.

Lower bound and endpoints. Moment duality and Lemma B.3 give gap $\Gamma \geq ce^{-C\Phi(h,KA)}$. The frozen-identity bound (67) gives $R_\delta \geq 3\Gamma/8 \geq ce^{-CE_*}$, including arbitrarily small positive gaps. The exact-sharing submodel independently gives $R_\delta \geq c\rho$. At $\delta = 0$, Theorem 5.1 gives $R_0 \asymp \rho$. For $K \leq B$, average within-prompt all-subsets winning labels: each prompt contribution is in $[0, 1]$ and variance of the average is at most $1/(4M)$. The observed-bit prompt submodel gives the matching lower bound.

Taking logarithms in the two risk inequalities yields $\log(1/R_\delta) \asymp 1 + \min\{\log(1/\rho), E_*\}$. Multiplicative constants become additive logarithmic constants, and the midpoint supplies the $\log 2$ floor when the displayed minimum is bounded. This concludes Theorem 5.2.

B.7 Horizon and precision are distinct consequences

We prove success at every target in the prefix, without assuming monotonicity of finite-sample risk. Explicit constant gaps preserve the stated tolerance range $0 < \epsilon < 1/8$.

Fixed-error horizon. Put $Z_* = \min(MB, D^2/\delta)$. If $K \leq \alpha Z_*$, $0 < \alpha \leq 1$, then

$$\rho \leq \sqrt{M^{-1} + \alpha}, \qquad E_* \geq D \log(1 + 1/(\alpha D)) \geq \log(1 + 1/\alpha).$$

The last inequality follows because $x \log(1 + c/x)$ increases for $x > 0$. Choose α_ϵ small and M_ϵ large so the upper bound is at most ϵ. For $K \leq B$, use the within-prompt average. Every $K \leq \alpha_\epsilon Z_*$ and $K = 1$ is controlled, so $\mathcal{C}_\epsilon \geq \max(1, \lfloor \alpha_\epsilon Z_* \rfloor) \geq \alpha_\epsilon Z_*/2$.

For the whole tolerance range, an unspecified logarithmic prefactor is insufficient. With known uniform scores compare $\eta_0 = 0$ and $\eta_1 = \mathbf{1}_{(1-1/K,1]}$. Their gap is at least $1 - e^{-1}$ and data TV at most MB/K. At $K = 4MB$, risk exceeds $1/8$, so

$$\mathcal{C}_\epsilon < 4MB. \tag{91}$$

For the information restriction, use the same (h, A) as above at $K_0 = \lceil 64h^2/A \rceil$. Equation (69) and product TV below $1/4$ give risk at least $3/16$. In the dense branch $K_0 \leq 5120eBL + 1$; in the sparse branch $K_0 \leq 135200D^2/\delta + 1$. Since $D^2/\delta \geq 4$ and is at least BL in the dense branch, both imply $\mathcal{C}_\epsilon < 140000D^2/\delta$. A failing target bounds the genuine prefix without assuming monotone risk. At $\delta = 0$, the pooled bound controls all $K \leq c_\epsilon MB$ for $M \geq M_\epsilon$, the mean handles $K = 1$, and the same sampling converse applies. This proves (19).

The root satisfies $r_* \asymp \sqrt{\tau L}$ for $L \leq \tau$, and $r_* \asymp L/\log(e + L/\tau)$ otherwise. In particular,

$$M = B^2, \quad \delta = B^{-1} \quad \Longrightarrow \quad \mathcal{C}_\epsilon = \Theta_\epsilon \left[B(\log B/\log\log B)^2 \right]. \tag{92}$$

Both sample count and limiting horizon instead scale as B^3. If $\mathcal{C}_\epsilon \geq a_0 B^2/\delta$ for fixed $a_0 > 0$, the explicit converse forces $D \geq \sqrt{a_0/140000}\, B$, hence for large B, $\log(eM) \gtrsim_{a_0} B \log(1 + 1/\delta)$. Conversely a sufficiently large ϵ-dependent multiple of this logarithmic budget makes $D = B$ and $MB \geq B^2/\delta$, attaining a fixed fraction of the limit.

Infinite repetition. Discarding extra prompts makes $r_{B,\delta}(k) = \lim_M R_\delta(B, M, k)$ well defined. For its genuine horizon $V_\epsilon^\infty = \sup\{T : \max_{1 \leq k \leq T} r_{B,\delta}(k) \leq \epsilon\}$,

$$V_\epsilon^\infty(B, \delta) = \Theta_\epsilon(B^2/\delta) \quad (B \geq 1,\ 0 < \delta \leq 1/4,\ 0 < \epsilon < 1/8). \tag{93}$$

The lower bound follows from the finite horizon once $D = B$ and $MB \geq B^2/\delta$. For the reverse, the preceding two-band construction with $h = B + 1$, $A = 2\delta$ and $K = \lceil 32(B + 1)^2/\delta \rceil$ has identical group laws and gap at least $1/2$. Equation (67) gives risk at least $1/4$ for every finite M and its limit; no fixed-population supremum is interchanged with this limit. At $\delta = 0$, pooling gives an infinite limit horizon.

Consistency under polynomial resources. Fix $\alpha \in (0, 1]$, $\gamma > 0$ and sequences $B \leq K^{1-\alpha}$, $M \leq K^\gamma$, $0 \leq \delta \leq 1/4$ as $K \to \infty$. Then $R_\delta \to 0$ if and only if $M \to \infty$, $K/(BM) \to 0$ and $K\delta \to 0$. The exact-sharing submodel makes the first two necessary; all three make (15) vanish, with $K \leq BM$ eventually. If $K\delta \geq a > 0$ on a subsequence, nested classes give $R_\delta \geq R_{a/K}$. At radius a/K, $\tau = Ba/K \leq aK^{-\alpha}$ and $L \leq 1 + \gamma \log K$. For fixed $r_0 > \max\{1, \gamma/\alpha\}$, eventually

$r_0 \log(1 + r_0/\tau) > L$, so $D \le r_0$ and $E_* \le r_0 \log(1 + r_0/a)$ is bounded. Theorem 5.2 prevents vanishing risk, proving necessity.

The power-law gap cannot be replaced by $B = o(K)$. For $B = \lceil K/\sqrt{\log K} \rceil$, $M = \lceil K^3 \rceil$ and fixed $\delta \in (0, 1/4]$, the dense root gives $D \sim \sqrt{B\delta \log(eM)}$ and $E_* \sim 3\sqrt{\log K}$, so risk tends to zero. This statement concerns consistency, not a calibrated finite-budget method choice.

Why horizon does not encode precision. Take $B = n^2$, $\delta = n^{-2}$, $M = (n+1)^n$, $K = \lfloor n^3/\log n \rfloor$, and $Z = \min(MB, D^2/\delta)$. Then $K > B$ eventually and $Z/K \asymp n \log n$, whereas

$$R_\delta(B, M, K) = \exp\{-\Theta(n \log\log n)\}. \tag{94}$$

Indeed $\tau = 1$, $L = 1 + n\log(n+1)$, $D \sim n$, $Z \asymp n^4$, $a \sim n/\log n$, $E_* \sim n \log\log n$, and $\log(1/\rho) \sim n \log n/2$. Apply Theorem 5.2. A bound $R_\delta \le C[\rho + \exp(-cZ/K)]$ with universal $c, C > 0$ would instead be $\exp[-\Omega(n \log n)]$, contradicting the lower bound.

For small ϵ, the risk theorem requires $\rho \le \epsilon/c$ and $E_* \ge C^{-1} \log(c/\epsilon)$; it is sufficient that $\rho \le \epsilon/(2C)$ and $E_* \ge c^{-1} \log(2C/\epsilon)$. For either positive threshold s, inversion gives the main-text equivalence in Section 5.4. Its right side is asymptotic to $D^2/(\delta s)$ when $s/D \to 0$ and to $(D/\delta)e^{-s/D}$ when $s/D \to \infty$. Sampling still requires $M \gtrsim \epsilon^{-2}$ and $K \lesssim \epsilon^2 MB$ with their respective constants. These are extrapolation statements only above B; exponent constants are not identified.

B.8 Fixed-depth correction and equal-cost accuracy

Appendix B.7 characterizes consistency under polynomial resources. We now refine the small-mismatch risk at fixed B, with constants allowed to depend on B. At equal candidate and label budgets $N = 2K^3$ and $\delta = K^{-5/4}$, let $R_b = R_\delta(b, N/b, K)$ for $b = 1, 2$. Both designs are consistent, but only depth two can eventually meet the shrinking worst-case expected-error requirement $\epsilon_K = K^{-3/8}$. The final paragraph derives both rates and the applicability threshold; neither is a finite-budget crossing or an advantage over unrestricted acquisition.

Theorem B.6 (Optimal correction order at fixed pool depth)**.** *For every fixed integer $B \ge 1$, all $M \ge 1$, $K \ge 2B + 2$ and $0 \le K\delta \le 1/16$,*

$$R_\delta(B, M, K) \asymp_B \min\left\{1, \min_{1 \le j \le B} \left[(K\delta)^j + \sqrt{K^j/M}\right]\right\}. \tag{95}$$

The upper rule uses empirical marginal ranks and within-prompt comparisons; the converse retains all numerical data, randomization and observed identities.

Proof. **1. Integrated-pilot upper bound.** Put $a = K\delta$, $d = 2\delta$, and first suppose $M \ge 8K$. Form $\widehat{H}$ from one score from each of $m = \lfloor M/2 \rfloor$ pilot prompts; these scores are iid from $H = \mathbb{E}F_X$. With $n = M - m$ and the comparison kernel (84), use the ungated rule

$$Z_{i,j} = \frac{1}{B} \sum_b K G_{ib} \sum_{\ell < j} \binom{K-1}{\ell} h_b^{K-1-\ell} V_{b,\ell}^{h_b}, \qquad \widehat{\theta}_{K,j} = \Pi_{[0,1]}\left(\frac{1}{n}\sum_i Z_{i,j}\right). \tag{96}$$

This is the same observed comparison kernel as before, but a different, one-score pilot. All MB candidates and labels remain charged.

For $Y \sim \mathrm{Bin}(m, z)$, the Bernoulli mgf bound $\mathbb{E}e^{t(Y-mz)} \le e^{mzt^2}$, $|t| \le 1$, with $|t| = \min\{u/(2mz), 1\}$ yields a two-sided tail bound $2e^{-\min\{u^2/(4mz), u/2\}}$. Tail integration gives

$$\mathbb{E}|Y/m - z|^r \le C_r[(z/m)^{r/2} + m^{-r}]. \tag{97}$$

At an anchor put $v = 1 - p$, $z = 1 - H(S)$, $Q = 1 - h$, so $|z - v| \le d$. For $m \ge 4K$ and $K/2 \le \lambda \le 2K$, exponential tilting gives $\mathbb{E}[e^{-\lambda Q} f(Q)] = D_0^m \mathbb{E} f(Y'/m)$, where $D_0 = 1 - z + ze^{-\lambda/m}$, $Y' \sim \mathrm{Bin}(m, z')$ and $z' = ze^{-\lambda/m}/D_0$. Since $d \le 1/(8K)$, $D_0^m \le e^{Kd/4} e^{-Kv/4}$, $z' \le v + d$, and $|z - z'| \le 4K(v+d)/m$. Thus

$$\mathbb{E}[e^{-\lambda Q}|Q - v|^r] \le C_r e^{-Kv/4}\left[d^r + \left(\frac{v+d}{m}\right)^{r/2} + m^{-r} + \left(\frac{K(v+d)}{m}\right)^r\right]. \tag{98}$$

The zeroth moment is at most $Ce^{-Kv/4}$; without tilting omit the exponential and final term. Also $\int_0^1 e^{-Kv/4}(v+d)^t dv \le C_t K^{-t-1}$.

The conditional Taylor remainder is bounded by $K\binom{K-1}{j}|Q-v|^j[e^{-(K-1-j)v}+e^{-(K-1-j)Q}]$. Because $K-1-j \ge K/2$, integration in the uniform anchor rank and (98) give

$$\mathbb{E}_{\widehat{H}}|\mathbb{E}[Z_{i,j} \mid \widehat{H}] - \theta_K| \le C_B[(K\delta)^j + (K/m)^{j/2}]. \tag{99}$$

With a true marginal CDF the bias is at most $\binom{K}{j}(2\delta)^j[1+2(K-j)\delta]$, but the empirical bound is required here. Jensen across anchors and $\mathbb{E}(V_\ell^h)^2 \le [p(1-p)+(p-h)^2]^\ell$ give

$$\mathbb{E}Z_{i,j}^2 \le C_B K^2 \sum_{\ell<j} K^{2\ell} \int_0^1 \mathbb{E}\Big[e^{-2(K-1-\ell)Q}(v^\ell + |Q-v|^{2\ell})\Big]\, dv \le C_B K^j.$$

The two integrals are bounded by $C_\ell K^{-\ell-1}$ and $C_\ell K^{-2\ell-1}$ using (98), with its zeroth version at $\ell = 0$. Conditional independence across analysis prompts supplies fluctuation $C_B\sqrt{K^j/n}$. For $M < 8K$, every branch is at least $1/\sqrt{8}$, so the midpoint suffices. Minimizing over the declared finite envelope proves the upper bound.

2. Marked three-band lower bound. For $a > 0$, use three fixed score intervals with uniform conditional positions and masses $(1-w-p, w, p)$, $w = p_0 = 1/K$, $p \in [p_0, p_0+\delta]$. A prompt bit z labels only the middle interval, so

$$t_K(p,z) = z\psi_K(p), \qquad \psi_K(p) = (1-p)^K - (1-p-1/K)^K. \tag{100}$$

Densities are at most three and the midpoint-CDF radius is at most $\delta/2$. For $1 \le r \le B$, set $p_l = p_0 + \delta l/r$ and $\mu_\pm = 2^{1-r}\sum_{l \text{ even/odd}} \binom{r}{l}\delta_{p_l}$. They match degrees below r. The priors $\nu_A = \frac{1}{2}(\mu_+,1)+\frac{1}{2}(\mu_-,0)$ and $\nu_B = \frac{1}{2}(\mu_-,1)+\frac{1}{2}(\mu_+,0)$ have identical score-parameter marginals. Their signed integral of f is $(-1)^r 2^{1-r}\int_{[0,\delta/r]^r} f^{(r)}(p_0+\sum t_i)dt$. The derivative of ψ_K has constant sign, with bracket $(K-r)\int_p^{p+w}(1-v)^{K-r-1}dv \ge e^{-5}/2$, using $K-r \ge K/2$ and $(1-33/(16K))^K \ge e^{-5}$. Therefore the deployment gap obeys

$$\alpha_r a^r \le d_r \le a^r, \qquad \alpha_r = \frac{e^{-5}}{2(4r)^r}. \tag{101}$$

Retain each candidate's band and label as category y; positions carry no remaining parameter information. Write $Q_0 = (Q_A+Q_B)/2$, $D_y = (Q_A - Q_B)/2$. If no middle candidate occurs, $D_y = 0$. Otherwise let middle/upper/lower counts be $m_0 \ge 1, h_0, l_0$, and put $g(p) = p^{h_0}(1-w-p)^{l_0}$. Then $Q_0(y) = w^{m_0}\int g\, d(\mu_+ + \mu_-)/4$ and $D_y = \pm w^{m_0}\int g\, d(\mu_+ - \mu_-)/4$. For $r = B$, $\deg g < B$ gives identical full-data laws. For $r < B$, $g \ge 4^{-B}w^{h_0}$ and Leibniz gives $|g^{(r)}| \le (4B)^B w^{(h_0-r)_+}$. Each information summand is at most $4^B(4B)^{2B}\delta^{2r}w^{m_0+2(h_0-r)_+-h_0}/8$; the exponent is at least $1-r$ and there are at most $2\cdot 3^B$ categories. Hence

$$I_r := \sum_y D_y^2/Q_0(y) \le A_B a^{2r}/K^{r+1}, \qquad A_B = 12^B(4B)^{2B}. \tag{102}$$

Contract about the midpoint by $\zeta = (8\sqrt{A_B})^{-1}\min\{1, K^{(r+1)/2}/(a^r\sqrt{M})\} \le 1/2$. Since $|D_y| \le Q_0(y)$, group KL is at most $8\zeta^2 I_r$, product KL at most $1/8$, and TV at most $1/4$. The observed-identity construction (67) then yields $R_\delta \ge c_B\min\{a^r, \sqrt{K^{r+1}/M}\}$ for $r < B$; the exact $r = B$ collision gives $R_\delta \ge \alpha_B a^B/2$. The known-uniform shared tail pair adds $cB^{-1/2}\min\{1, \sqrt{K/M}\}$, including $a = 0$.

Let $L_* = \max\{a^B, \max_{0\le r<B}\min(a^r, \sqrt{K^{r+1}/M})\}$, with $a^0 = 1$. For any j, terms with $r < j$ are at most $\sqrt{K^j/M}$ and those with $r \ge j$ at most a^j. Conversely, if $L_* < 1$, choose the first j with $a^j \le L_*$; the $r = j-1$ term forces $\sqrt{K^j/M} \le L_*$. Thus the risk envelope is between L_* and $2L_*$. This includes $B = 1$ and $a = 0$. All constants may depend on B; they are not numerical crossing guarantees. For example $c_B = 3e^{-5}/[256\, 12^{B/2}(4B)^{2B}]$ is permitted in the lower comparison; the upper constants remain uninstantiated. □

Equal-cost accuracy and finite-order interpretation. For $N = 2K^3$, $\delta = K^{-5/4}$ and $M = N/B$, the depth-one envelope is $K^{-1/4} + 2^{-1/2}K^{-1}$, while the depth-two branches are $K^{-1/4}+K^{-1}$ and $2K^{-1/2}$. Hence $R_1 \asymp K^{-1/4}$, $R_2 \asymp K^{-1/2}$, and their ratios to $\epsilon_K = K^{-3/8}$ are respectively of orders $K^{1/8}$ and $K^{-1/8}$. The theorem's local condition first holds at $K \geq 65536$, hence $N \geq 2^{49}$; this is a sufficient applicability threshold, not a necessary budget for the phenomenon. Unspecified constants may delay the crossing.

The finite-order comparisons in Appendix D.3.6 use the explicit P_j factor in (87); neither those comparisons nor the uninstantiated constants in (95) establish a nonvacuous real-data risk guarantee.

C WHY JOINT LABELS MATTER FOR A STRUCTURED RESPONSE CLASS

The unrestricted width in Section 3 is not a sufficiency claim. This appendix gives a same-depth statistical separation on a monotone cubic class, using one additional joint statistic and no numerical scores. This finite-dimensional restriction is stronger than smooth monotonicity: nonidentification on the larger class in Theorem 4.1 does not preclude recovery here.

In the homogeneous hidden-rank channel, let $\mathcal{C}_3 = \{\eta \in \mathcal{P}_3 : 0 \leq \eta \leq 1, \eta' \geq 0\}$, and let $\mathcal{C}_{3,\gamma}$ additionally require $c_1 = 3\int_0^1 u(1-u)\eta'(u)\,du \geq \gamma$, $0 < \gamma \leq 1/4$. For M iid three-label words W, compare the full records with only their final empirical mean $\bar{W}$. Both cost $3M$ candidates and labels once. Denote their minimax expected $\sup_{4\leq k\leq T} |\widehat{\theta}_k - \theta_k|$ by $R^{\mathrm{word}}_{M,T}$ and $R^{\mathrm{mean}}_{M,T}$.

Theorem C.1 (Dependence retention separates curve consistency)**.** *For $M \geq 1$ and $T \in \{4, 5, \ldots\} \cup \{\infty\}$,*

$$R^{\mathrm{word}}_{M,T}(\mathcal{C}_3) \asymp M^{-1/4}, \qquad R^{\mathrm{word}}_{M,T}(\mathcal{C}_{3,\gamma}) \asymp_\gamma M^{-1/2}.$$

One additional statistic, $M^{-1}\sum_i W_1^{(i)}W_2^{(i)}$, attains both rates when retained with the three means. Yet $\liminf_{M\to\infty} R^{\mathrm{mean}}_{M,T}(\mathcal{C}_{3,\gamma}) > 0$. Constants are independent of M, T.

The proof first identifies the missing coefficient, then bounds the inverse and proves both rates. The aggregation converse concerns the actual final vector $\bar{W}$, including its fluctuations, not its expectation, running means, or block means.

Identification and branch selection. Write $\eta = \sum_{j=0}^3 c_j L_j$ in the shifted Legendre basis $(1, 2u-1, 6u^2-6u+1, 20u^3-30u^2+12u-1)$, and set $m_j = \mathbb{E}W_j$, $A = \mathbb{E}W_1W_2$. Rank integration gives

$$(c_0, c_1, c_2) = ((m_1+m_2+m_3)/3,\ m_3 - m_1,\ 5(m_1 - 2m_2 + m_3)/3), \tag{103}$$

$$A = \mathcal{B} + (c_3^2 - 3c_1c_3)/210,$$

$$\mathcal{B} = c_0^2 - c_0c_1/2 - c_0c_2/10 + c_1^2/10 + c_2^2/70, \tag{104}$$

$$c_3 = 3c_1/2 - \sqrt{D}, \qquad D = 9c_1^2/4 + 210(A - \mathcal{B}). \tag{105}$$

The joint identity follows from $A = 3\int F^2$, $F(u) = \int_0^u \eta$ (Hàn et al., 2021, Lemma 25). Monotonicity selects its root: it gives $c_1 = 3\int u(1-u)\eta' \geq 0$ and $2c_1 - 3c_3 = \eta'(1/2) \geq 0$; hence $3c_1/2 - c_3 \geq 5c_1/6$, selecting the root in (105). If $c_1 = 0$, then $\eta' \equiv 0$. This use of dependence cannot start at depth two: $\eta_0 = 1/4 + u/2$ and $\eta_1 = \eta_0 + L_2/32$ share its rank means and joint moment $(\int \eta)^2$, but their $K = 3$ targets differ by $1/320$.

One estimator for the whole curve. Insert $(\bar{W}, M^{-1}\sum_i W_1^{(i)}W_2^{(i)})$ in the inverse, use $\sqrt{\widehat{D}_+}$, and output the single curve

$$\widehat{\theta}_k = \Pi_{[0,1]} \sum_{j=0}^3 \widehat{c}_j g_j(k), \qquad g_0 = 1, \quad g_j(k) = \prod_{r=1}^j \frac{k-r}{k+r}, \quad k \geq 4. \tag{106}$$

Rodrigues' formula gives $\int f_k L_j = g_j(k)$ and $|g_j(k)| \leq 1$, so sup loss, even for $T = \infty$, is at most $\sum_j |\widehat{c}_j - c_j|$. For the maximum error e of the four moments, $\mathbb{E}e \leq 2/\sqrt{M}$. The first three

coefficients and the polynomial D are Lipschitz on the empirical cube. The only loss of regularity is therefore in the square root: $|\sqrt{x_+} - \sqrt{y}| \le \sqrt{|x-y|}$ gives $CM^{-1/4}$. On $\mathcal{C}_{3,\gamma}$, where $c_1 \ge \gamma$, the bound $\sqrt{D} \ge 5\gamma/6$ instead gives $C_\gamma M^{-1/2}$. The estimator does not use γ, and the bounded coefficients $g_j(k)$ make both guarantees uniform in the curve endpoint T.

Matching full-word lower bounds. For $0 < r \le 1/6$, the monotone responses $\eta_{0,r} = 1/2 + rL_1$, $\eta_{1,r} = \eta_{0,r} + (r/2)L_3$ lie in $[1/4, 3/4]$ and have equal means. In lexicographic word order,

$$p_3(\eta_{1,r}) - p_3(\eta_{0,r}) = -r^2(0, 1, -2, 1, 1, -2, 1, 0)/168. \tag{107}$$

Indeed $\delta A_{12} = -r^2/168$, reflection gives $\delta A_{23} = \delta A_{12}$, and $\sum_{i<j} A_{ij} = 3(\int \eta)^2$ gives $\delta A_{13} = -2\delta A_{12}$; the triple moment is unchanged. Each word mass is at least $1/64$, so one-word KL is at most $4r^4/147$. Taking $r = M^{-1/4}/6$, product KL, Pinsker, and the $K = 4$ gap $r/70$ give risk at least $M^{-1/4}/1680$. For the regular rate, on every $\mathcal{C}_{3,\gamma}$, the pair $1/2 \pm t + L_1/4$, $t = (32\sqrt{M})^{-1}$, has target gap $2t$ and KL at most $3/28$ even with all ranks revealed, giving $M^{-1/2}/64$.

Why empirical means still lose information. Equal expectations are insufficient: fluctuations of $\bar{W}$ may distinguish the responses. Compare the full laws of $S_M = M\bar{W}$ for $\eta_0 = 1/2 + L_1/4$, $\eta_1 = \eta_0 + L_3/8$. Their common mean is $(3/8, 1/2, 5/8)$; the first word law is $p_0 = (20, 31, 18, 31, 11, 18, 11, 20)/160$ and the difference in (107) with $r = 1/4$. Their covariances are

$$\Sigma_0 = \frac{1}{13440}\begin{pmatrix} 3150 & 84 & 42 \\ 84 & 3360 & 84 \\ 42 & 84 & 3150 \end{pmatrix}, \quad \Sigma_1 = \frac{1}{13440}\begin{pmatrix} 3150 & 79 & 52 \\ 79 & 3360 & 79 \\ 52 & 79 & 3150 \end{pmatrix}. \tag{108}$$

Both matrices are positive definite. Fourier inversion gives the uniform lattice local limit $\sup_s |M^{3/2} \Pr_i(S_M = s) - \varphi_{\Sigma_i}((s - M\mu)/\sqrt{M})| \to 0$: full cube support makes the characteristic function strictly smaller than one off the origin; near zero its centered expansion is $1 - t^\top \Sigma_i t/2 + O(\|t\|^3)$, dominated by $e^{-a\|t\|^2}$. After $v = \sqrt{M}t$, dominated convergence handles the near region and $O(M^{3/2}\rho^M)$ the complement. To pass from this local approximation to total variation, sum on $\|s - M\mu\| \le R\sqrt{M}$, bound the tails by $\mathrm{tr}(\Sigma_i)/R^2$, and then let $R \to \infty$. Consequently

$$\begin{gathered} \mathrm{TV}(\mathcal{L}_0(S_M), \mathcal{L}_1(S_M)) \to v = \mathrm{TV}(N(0, \Sigma_0), N(0, \Sigma_1)) < 1, \\ \liminf_M R^{\mathrm{mean}}_{M,T} \ge g_3(k)(1 - v)/16 > 0 \end{gathered} \tag{109}$$

for every $4 \le k \le T$. Gaussian affinity $(\det \Sigma_0 \det \Sigma_1)^{1/4}/\det((\Sigma_0 + \Sigma_1)/2)^{1/2}$ gives $v < 1/500$ by the displayed rational matrices: the floor exceeds .029 at $k = 16$ and $499/8000$ for $T = \infty$. Weak convergence alone would not prove this TV statement.

Scope. Four summaries attain rates, not sufficiency; depth-four marginals already identify cubics. Polynomial forcibility and signature recovery are classical (Hàn et al., 2021; Chang et al., 2017; Pfeffer et al., 2019; Améndola et al., 2019); the contribution here is the shape-selected inverse and same-depth curve separation.

D Experimental evidence

The three questions of Section 6 organize the evidence: shallow-summary inference (Appendix D.1), recovery mechanisms under enforced assumptions (Appendix D.2), and applicability to natural scores (Appendix D.3). Empirical screens and simulated errors are distinct from minimax guarantees.

D.1 What do shallow summaries support?

We distinguish agreement with a deeper reference from what shallow summaries support, accounting separately for plug-in compatibility and sampling error. These analyses use rank means or pooled pair records, not optimal full-word estimators.

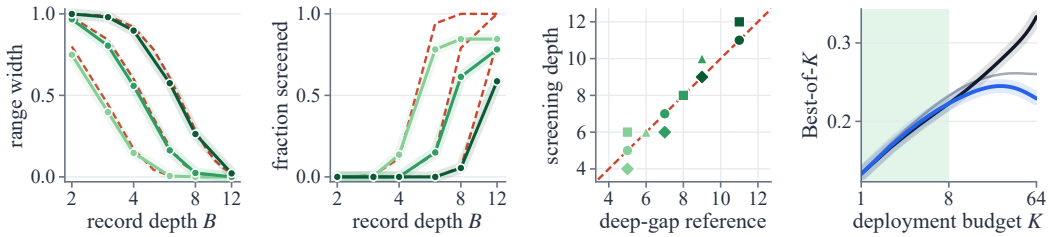


Figure 4: **Rank-mean ranges and retrospective comparison depths.** Light to dark: $K = 8, 16, 32$. Left to right: median plug-in width (red: $\Delta_{B,K}$); screened fraction (red: gaps above $\Delta_{B,K}$); screening versus deep-gap reference depth (red: equality; circle: MATH, square: MMLU-Pro, triangle: GPQA, diamond: MBPP-Plus); independent Mistral references for GRM (blue), OffsetBias (dark) and Skywork v0.2 (grey), with approximately agreeing Best-of-2, 4 and 8 summaries (shaded: $K \leq 8$). Screening has no coverage guarantee.

Data. The public Best-of-K sets of Frick et al. (2025) contain 512 rows per domain (507 for MBPP-Plus), each with 32 labeled candidates from one prompt and one of four generators. All rows contain both labels; MMLU-Pro (Wang et al., 2024), GPQA (Rein et al., 2024) and MBPP-Plus (Austin et al., 2021; EvalPlus Team, 2023) additionally restrict correctness to 10–90%, whereas 98 MATH (Hendrycks et al., 2021) rows fall outside that range. Estimates describe the filtered banks. Label-based row selection can violate conditional iid sampling, so these ranges are descriptive plug-in diagnostics rather than population identification guarantees. Since release order correlates with length, candidate splits follow a fixed random permutation within each row. We evaluate 17 scorers (Table 2; Liu et al., 2025), using inputs as in Appendix D.3.5. Separate unfiltered MATH-500 checks use the historical and prospective banks of Appendices D.3.1 and D.3.7.

Table 2: Reward models scored on the public Best-of-K sets.

model (Hugging Face)	size	domains
RLHFlow/ArmoRM-Llama3-8B-v0.1	8B	MATH
sfairXC/FsfairX-LLaMA3-RM-v0.1	8B	MATH
Ray2333/GRM-Llama3-8B-rewardmodel-ft	8B	MATH
Ray2333/GRM-Llama3.2-3B-rewardmodel-ft	3B	MATH, MMLU-Pro, GPQA, MBPP-Plus
Ray2333/GRM-gemma2-2B-rewardmodel-ft	2B	MATH, MMLU-Pro, GPQA, MBPP-Plus
NCSOFT/Llama-3-OffsetBias-RM-8B	8B	MATH
weqweasdas/RM-Gemma-2B	2B	MATH, MMLU-Pro, GPQA, MBPP-Plus
Skywork/Skywork-Reward-Llama-3.1-8B	8B	MATH
Skywork/Skywork-Reward-V2-Llama-3.2-1B	1B	MATH, MMLU-Pro, GPQA, MBPP-Plus
Skywork/Skywork-Reward-V2-Llama-3.1-8B-40M	8B	MATH
Skywork/Skywork-Reward-V2-Qwen3-0.6B	0.6B	MATH, MMLU-Pro, GPQA, MBPP-Plus
Skywork/Skywork-Reward-V2-Qwen3-1.7B	1.7B	MATH, MMLU-Pro, GPQA, MBPP-Plus
Skywork/Skywork-Reward-V2-Qwen3-4B	4B	MATH, MMLU-Pro, GPQA, MBPP-Plus
Skywork/Skywork-Reward-Llama-3.1-8B-v0.2	8B	MATH
Skywork/Skywork-Reward-V2-Llama-3.2-3B	3B	MATH, MMLU-Pro, GPQA, MBPP-Plus
Skywork/Skywork-Reward-V2-Llama-3.1-8B	8B	MATH
LxzGordon/URM-LLaMa-3.1-8B	8B	MATH

D.1.1 RANK SUMMARIES AND COMPARISON RULES

Pairwise accuracy largely agrees with deeper model rankings: Kendall $\tau = .71$–1.00 at $K = 16$ and .36–1.00 at $K = 32$, lowest on GPQA; no point-estimate reversal at $K = 32$ is resolved.

Under the iid model, for population rank means $m_r = \int \beta_{r,B}\eta$, define $\mathcal{I}_{B,K}(m) = \{\theta_K(\eta) : 0 \leq \eta \leq 1,\ \int \beta_{r,B}\eta = m_r\}$. This contains the target and encloses the full-word range; its monotone version restricts η to be nondecreasing. The empirical calculation substitutes rank means using *all* candidates in each row, averaged over all B-subsets and pooled over rows. It is neither a B-sample experiment nor a confidence interval. Infeasible point estimates and bootstrap draws are reported without projection.

Screening rule. A pair passes if one plug-in range lies above the other in at least 97.5% of paired prompt-cluster bootstrap draws. This per-pair, unadjusted screen has no coverage guarantee. A reference difference is resolved when its 95% bootstrap interval excludes zero; an opposite sign contradicts the screen. References use the same candidates, disjoint 16-candidate halves, or independent 64-candidate banks. Screen failure does not establish impossibility for other estimators.

No unconstrained Best-of-16 ordering passes through depth four (Table 3). Adding the joint success probability of the pooled depth-two record still selects none at $K = 8, 16, 32$, with widths .74–1.00 of the marginal widths. Prompt-specific pair laws, complete depth-three/four words and the RewardBench 2 statistic are not assessed. Median widths are .82–1.00 of $\Delta_{B,K}$ at depths 2–8, versus .89–1.00 for a base-rate-matched uninformative scorer. This is not universal: Skywork-V2 Qwen3-4B on MBPP-Plus reaches .39–.53 at depths 4–8, $K \geq 16$, versus .98–.99 for its uninformative counterpart (Figure 4, far left).

Table 3: Box-class screens: reference-consistent screened / reference-resolved pairs by B, K. Columns use the same candidates, disjoint halves or independent MATH-500 references. Unresolved reference differences are excluded. Dashes mark rank-mean infeasibility at the point estimate or in over 20% of bootstrap draws. No screen contradicts a resolved reference.

	same candidates			disjoint halves		banks	
B	K=8	16	32	8	16	8	16
2	0/188	0/176	0/154	0/173	0/150	0/23	0/22
3	1/188	0/176	0/154	1/173	0/150	0/23	0/22
4	30/188	0/176	0/154	25/173	0/150	0/23	0/22
6	172/188	33/176	0/154	151/173	24/150	22/23	0/22
8	186/188	135/176	12/154	158/173	113/150	22/23	12/22
12	186/188	172/176	129/154	–	–	–	–

D.1.2 ACTUALLY OBSERVING B CANDIDATES PER PROMPT

At $K = 16$, random nested prefixes of each historical protocol record supply $B = 4, 8$ candidates for 96 prompts and eight scorers. Twenty partitions measure sensitivity on the same banks, not independent replication. Costs are 384 or 768 scored, labeled candidates rather than 3072 for all-subset estimates; separate references do not form ranges or select depth.

For each replay, we enlarge rank moments by $\rho = \sqrt{\log(2C/.05)/(2M)}$, $M = 96$, $C = 2 \cdot 8 \cdot (4 + 8) = 192$. Independent prompt blocks and bounded labels yield simultaneous 95% Hoeffding coverage over banks, scorers and depths under iid candidate sampling. With fixed prompts, the target is their average response; population generalization additionally requires representative sampling. We optimize over $0 \leq \eta \leq 1$ subject to $|\int \beta_{r,B}\eta - \hat{m}_r| \leq \rho$. A 4096-cell relaxation uses exact density integrals, enlarges each constraint by cell width h times the total variation of $\beta_{r,B}$, and expands target endpoints by Kh. Summing within-cell oscillation bounds justifies these enlargements without assuming piecewise-constant responses. Doubling the grid on one partition checks sensitivity, not interval-arithmetic certification.

Table 4: Best-of-16 range widths: all-subset plug-in versus shallow replay. Medians use eight scorers (all subsets) or scorers and 20 partitions (replay). Shallow plug-in widths condition on feasibility; the last column includes conservative simultaneous sampling uncertainty.

		all-subset	shallow plug-in		uncertainty
bank	B	width	feasible	width	width
Qwen	4	.356	160/160	.334	.859
Qwen	8	.016	48/160	.007	.625
Mistral	4	.334	142/160	.289	.887
Mistral	8	.012	2/160	.007	.686

All 32 all-subset estimates are feasible, but many genuinely shallow depth-eight estimates are not. No model pair has disjoint uncertainty-enlarged ranges in any replay. Table 4 thus separates plug-in precision from finite-data uncertainty; its conservative concentration bound does not measure irreducible information loss or the optimal sampling cost.

D.1.3 Sensitivity to depth, shape and candidate splits

Depth. The screening depth is the smallest $B \in \{2, \dots, 12\}$ selecting at least half the reference-resolved pairs with the correct sign. The retrospective reference depth is the first B with $\Delta_{B,K}$ below the domain's median estimated deep gap; it is not a shallow-only selector. They agree in 6 of 11 evaluable cells and differ by at most one in all 11 (Table 5). GPQA at $K = 32$ is censored: the reference depth is 13 and no tested depth screens half its three resolved pairs.

The unfiltered check merges protocol and reference banks (Qwen/Mistral: 96×96; prospective: 128×80), deliberately giving up reference independence, and tests depths 2–14. Across ten bootstrap seeds, all 89 evaluable combinations differ by at most one, with 5–7 of nine cells agreeing per seed and one censored combination (Table 6). These are sensitivity checks, not new replications. No screen contradicts a resolved sign, but only 5–20 pairs per cell are resolved, mostly involving the process verifier on Qwen and the prospective bank.

Table 5: Screening / deep-gap reference depth (median estimated Best-of-K gap).

domain	$K = 8$	$K = 16$	$K = 32$
MATH	5/5 (.072)	7/7 (.088)	11/11 (.098)
MMLU-Pro	6/5 (.048)	8/8 (.055)	12/11 (.061)
GPQA	6/6 (.021)	10/9 (.026)	–/13 (.021)
MBPP-Plus	4/5 (.165)	6/7 (.181)	9/9 (.193)

Table 6: Unfiltered MATH-500 banks, first bootstrap seed: screening / deep-gap reference depth (resolved pairs of the bank's verifier pairs).

bank	$K = 8$	$K = 16$	$K = 32$
Mistral (96×96)	5/6 (17)	7/8 (17)	11/11 (17)
Qwen (96×96)	6/6 (7)	9/9 (6)	14/14 (7)
prospective (128×80)	6/6 (12)	8/9 (12)	12/13 (9)

At population moments, each range has width at most $\Delta_{B,K}$, so a true gap exceeding $2\Delta_{B,K}$ separates two ranges. Table 7 instead records the distinct benchmark $\Delta_{B,K} < g$: for $g = .05$–$.1$, depths are within two of $2\sqrt{K}$. Neither benchmark estimates an unknown gap or accounts for shallow sampling error.

Table 7: Smallest record depth B with $\Delta_{B,K} < g$ (exact L^1 distances).

$g \setminus K$	8	16	32	64	128	256
.20	4	6	9	13	19	27
.10	5	7	10	15	21	31
.05	5	8	12	17	24	34
$2\sqrt{K}$	5.7	8.0	11.3	16.0	22.6	32.0

Shape. Assuming monotonicity screens 34 of 136 MATH pairs at depth four, without contradiction. To test this assumption, we use the $\hat{\Sigma}^{-1}$-weighted squared distance of rank means to the monotone moment set, with covariance from 20,000 prompt-cluster bootstrap draws. A χ^2_B reference and 1000 draws recentred at the fitted monotone point give p-values Holm-adjusted within each domain over scorers and $B = 3, 4, 6, 8$. Only the two MBPP-Plus scorers with pairwise accuracy at most .50 (GRM-3B and RM-Gemma-2B) reject: recentred adjusted $p = .032$ throughout; the χ^2 test agrees except GRM-3B at depth eight (.054). The other 39 model–domain pairs do not reject.

Non-rejection gives little support against small violations. In PPE-like simulations (512 rows, 430 prompt clusters, 32 candidates, logistic difficulty, with/without filtering; 200 datasets per setting), the depth-eight recentred test detects top-fifth response drops of .10 in 90–95% of datasets, but .05 in only 24.5–26.5% (χ^2: 35–35.5% and 0.5–1%). It rejects three monotone profiles in at most 7%, and a score-independent profile in 5–10%. These finite checks do not provide uniform calibration or license a monotonicity assumption.

Shallow agreement and deeper differences. Four Mistral pairs have Best-of-2, 4 and 8 difference intervals within $\pm.03$. The largest independent Best-of-64 gap is .104 (OffsetBias over GRM; $[.021, .188]$), with unadjusted $p = .021$ but Holm $p = .082$ over four pairs (Figure 4, far right). None of seven Qwen pairs has a resolved deep gap. On disjoint public-data halves, none of 17 selected pairs differs after within-domain Holm correction; one pair is Holm-significant in three of ten alternative halvings. These split-sensitive contrasts are exploratory: approximate shallow-summary agreement does not establish equal full word laws.

Fixed label cost. With 1024 MATH labels and $B = 2, 4, 8, 16$, pairwise-accuracy selection gives mean Best-of-16 regret $.0053, .0018, .0009, .0008$ over ten random halvings of the same data; selecting by Best-of-B gives $.0053, .0055, .0098, .0163$. The former's $B = 8$ and 16 results are indistinguishable within noise. Depth benefits depend on the selection rule, not a universally optimal depth.

D.2 Which additional structure changes recovery?

These studies test retained joint information, sharing-based recovery and the statistical costs of dependence and correction. Each isolates a response or score condition, not a stage of one estimator.

D.2.1 What is lost when joint records are compressed

At depth three, joint labels enable cubic recovery but incur a finite-sample cost (Theorem C.1). We sample exact word histograms at $M \in \{2^{12}, 2^{16}, 2^{20}, 2^{24}\}$ for constant $1/2$, affine $1/2 + L_1/4$, cubic $1/2 + L_1/4 + L_3/8$, and out-of-class quartic $.3 + .4u^4$ responses. Each rule uses $3M$ candidates and labels; loss is maximum error over $K \in \{4, 8, 16, 32, 64, 128, \infty\}$. The affine and cubic responses share population rank means $(3/8, 1/2, 5/8)$, not word or empirical-mean laws.

We compare the four-counter inverse (106) with the three-mean quadratic fit ($c_3 = 0$). Separate finite-family decoders use equal priors on the first three responses, Gaussian summary likelihoods or exact multinomial word likelihoods, and return posterior-mean curves. These are restricted estimators, not optimal full-class comparisons.

On the cubic, the quadratic fit retains error .125 while the inverse overtakes it only at larger M; the fit wins on the constant and affine controls (Table 8; Figure 5, far left). The final sixteenfold

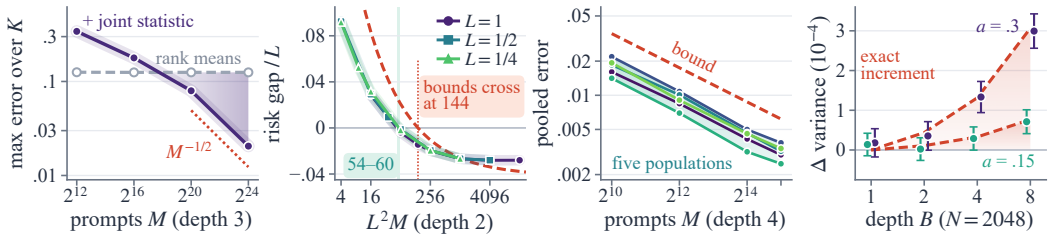


Figure 5: **Controlled effects of information and structure.** Red: theory. Left to right: monotone-cubic recovery from three means (error near .125) versus one additional joint statistic (small-sample cost; last-step decrease $M^{-1/2}$); canonical minus block-winner mean risk over eight slope profiles (guarantee crossover $L^2M = 144$, measured 54–60); exact-sharing pooled error at $B = 4, K = 16$ ($M^{-1/2}$ across five heterogeneous populations); variance under increasing prompt dependence at a fixed deployment curve, with the covariance prediction and simultaneous Monte Carlo intervals. The first two panels use grid-maximum loss.

Table 8: Record retention: mean grid-maximum error times 10^3, 8,192 repetitions per cell; Monte Carlo standard errors are below 2.83 on this scale, and zeros are rounded. Finite-family decoders receive extra model knowledge unavailable to the general inverse and quadratic fit; compare record types within that group.

Record and rule	2^{12}	2^{16}	2^{20}	2^{24}
Constant				
Three means: quadratic fit	27.76	6.85	1.73	0.43
Four counters: cubic inverse	242.64	181.15	93.26	46.52
Three means: finite-family	0.00	0.00	0.00	0.00
Four counters: finite-family	0.00	0.00	0.00	0.00
Full word: finite-family	0.00	0.00	0.00	0.00
Affine				
Three means: quadratic fit	26.76	6.74	1.65	0.42
Four counters: cubic inverse	343.68	176.62	56.81	13.51
Three means: finite-family	62.50	62.50	62.50	62.50
Four counters: finite-family	62.33	60.06	37.40	0.13
Full word: finite-family	61.61	49.86	4.12	0.00
Cubic				
Three means: quadratic fit	125.33	124.93	125.03	125.01
Four counters: cubic inverse	341.17	178.12	79.62	20.53
Three means: finite-family	62.50	62.50	62.50	62.50
Four counters: finite-family	62.36	59.96	36.97	0.21
Full word: finite-family	61.47	49.84	4.06	0.00
Quartic (outside class)				
Three means: quadratic fit	48.85	45.65	45.72	45.72
Four counters: cubic inverse	335.04	194.62	92.38	24.11
Three means: finite-family	151.55	150.06	150.00	150.00
Four counters: finite-family	151.46	150.01	150.00	150.00
Full word: finite-family	151.56	150.09	150.00	150.00

sample increase reduces inverse errors by factors 2.0, 4.2 and 3.9, respectively, consistent with the theorem's pointwise bounds, not evidence of minimax rates. Finite-family decoding retains error .15 under quartic misspecification.

D.2.2 A DEPTH-TWO PRECISION TASK: GUARANTEES AND MEASURED ERROR

The guarantees in Appendix A.6.4 need not order measured accuracy. We observe two rank means from $2M$ labeled candidates, without sharing, at $L = 1$. Responses are $\eta(u) = c_0 + \int_0^u \psi(t)dt$, $c_0 = (1 - \int_0^1 \psi)/2$, for eight predeclared slopes: $0, 1/4, 1/2, 3/4, 1, (1-\cos(2\pi u))/2, \mathbf{1}\{P_{16}(u) > u^{16}\}$ and $\mathbf{1}\{u > 1/2\}$; the last two coincide almost everywhere and both enter averages. Loss is maximum error on $\{2, 3, 4, 6, 8, 12, 16, 24, 32, 64, 128\}$.

All rules use the same unclipped estimator, replacing a_K by a: canonical (a_K), block-winner (1), read-all (0, not score pooling), shrinkage ($(1 + a_K)/2$), or the label-free minimizer on $[0, 2]$ of $\|au^2 + (1 - a)u - u^K\|_1/2 + \max\{1, |a|\}/(2\sqrt{M})$. At $M = 16, 1024$, we use 4,000 paired repetitions. The $M = 1024$ comparison was prespecified; tolerance $\epsilon = .16$ is post hoc.

Table 9: Certification does not order measured accuracy. Finite-grid errors at $L = 1$. Parentheses give one Monte Carlo standard error of the eight-profile mean. "Max" is the largest estimated profile risk, not a class-wide risk bound.

	$M = 16$		$M = 1024$	
Rule	Mean (MCSE)	Max	Mean (MCSE)	Max
Canonical	.1840 (.0013)	.2083	.07018 (.00011)	.1213
Block-winner	.1555 (.0009)	.2036	.09690 (.00008)	.1631
Read-all	.1536 (.0005)	.2486	.12914 (.00007)	.2464
Shrinkage	.1635 (.0011)	.1878	.08284 (.00009)	.1214
Bound-selected	.1555 (.0009)	.2036	.07553 (.00010)	.1212

Monotonicity of the canonical bias bound and absolute rank weights in K extends the argument of Appendix A.6.4 to grid-maximum expected loss. For maximum $T \geq 2$, canonical and block-winner have bounds

$$C_T(M) = \tfrac{1}{8} - \tfrac{1-2^{-T}}{2(T+1)} + \tfrac{2-2^{2-T}}{2\sqrt{M}}, \qquad D_T(M) = \tfrac{1}{6} - \tfrac{1}{2(T+1)} + \tfrac{1}{2\sqrt{M}}.$$

At $T = 128$, $C_T(662) = .15999$ while D_T has floor .1628; only canonical certifies .16 (requiring $M = 817$ for all $K \geq 2$). Yet at $M = 16$, block-winner's profile-mean error is .156 versus canonical's .184; its largest profile error .204 still exceeds the tolerance. Canonical wins on average at $M = 1024$, but not on every profile, and neither it nor bound selection dominates the other across budgets (Table 9).

Rescaling to $L \in \{1, 1/2, 1/4\}$ over $M = 16$–$16{,}384$ gives interpolated measured crossings at $L^2M = 54, 60, 59$, versus 144 for the guarantees (Figure 5, centre left). All measured errors respect their bounds. Thus the scale of the crossover is informative, but certification neither ranks finite-sample methods nor establishes a minimax rate.

D.2.3 RECOVERY WHEN EXACT SHARING IS ENFORCED

To test Theorem 5.1, we use $B = 4$ and common scores $S = 7 + 0.3\,\mathrm{logit}(U)$ with conditionally iid uniform U. Five Bernoulli-response populations comprise random power curves $a + (b - a)u^c$, a score-independent Bernoulli(.6) prompt bit, thresholds $\mathbf{1}\{u \geq X\}$ for X uniform on $\{.8, .9, .95, .99\}$, and Qwen/Mistral shapes from Appendix D.3.1. The last two are simulations enforcing sharing, not real-score tests.

At $M \in \{1024, 4096, 16384, 32768\}$, each rule pays $4M$ candidates and labels; we use 128 repetitions (256 at the largest M). Coverage uses the predeclared 95% radius $\sqrt{(1 + 3K/B)/(4M(.05))}$, looser than the refined theorem. Baselines are pooled depth-four plateau, block winners and clipped $1.75\widehat{\theta}_4 - .75\widehat{\theta}_2$ extrapolation.

Pooled error decreases as $M^{-1/2}$ in all five populations (Table 10; Figure 5, centre right), with 100% coverage in 20 cells (exact-binomial 95% interval $[.985, 1]$ at the largest M). Unbiasedness does not imply pointwise superiority: at $M = 32768$, block winners have the lowest reported MAE in the bit control. Extrapolation wins at the two smallest budgets on Qwen shapes.

Table 10: Enforced sharing: mean absolute error times 10^3. The final three columns use $M = 32768$. Block winners have the lowest reported MAE in the bit control; these are conditional examples, not uniform superiority.

	Pooled $K = 16$, prompt draws M				Same-record baselines		
Population	1024	4096	16384	32768	Plateau	Block	Extrap.
Smooth	16.071	8.442	4.122	2.985	96.207	96.149	42.793
Bit	21.874	10.786	4.989	3.810	2.538	2.216	2.827
Threshold	18.533	10.084	4.647	3.215	333.963	333.982	241.755
Q shape	14.121	6.965	3.176	2.491	18.373	18.378	4.017
M shape	19.313	9.063	4.591	3.398	94.803	94.755	55.299

Matched deployment curves isolate prompt dependence. To isolate the variance term in Theorem 5.1, take fair $X \in \{-1, 1\}$, common scores $S = U \sim \mathrm{Unif}[0, 1]$, and conditionally independent labels with $\eta_X(u) = .3 + .4u + aX$, $a \in \{0, .15, .3\}$. The marginal law of (S, G) and all targets $\theta_K = .3 + .4K/(K + 1)$ are unchanged. Here $C_\eta = a^2$, so Eq. (14) gives

$$\mathrm{Var}_a(\widehat{\theta}_K) - \mathrm{Var}_0(\widehat{\theta}_K) = a^2 \frac{B-1}{N-1} \left(1 - \sum_{r=1}^{N} w_{r,K}^2\right). \tag{110}$$

At $N = 2048$, $K = 16$, $B \in \{1, 2, 4, 8\}$ (16,384 records per depth), all eight predictions lie within Bonferroni 95% Monte Carlo intervals; only three of six positive effects exclude zero, with two $B = 1$ zero controls. At $B = 8, a = .3$, the increment is 2.99×10^{-4} ($[2.56, 3.43] \times 10^{-4}$), versus 3.07×10^{-4} predicted, or 35% of baseline variance (Figure 5, far right).

D.2.4 Paired comparability with unchanged deployment

A hidden prompt-specific score warp varies comparability without changing any deployment winner. For fair iid $a \in \{-1, 1\}$, let $w = 1/16$, $b_w(u) = \max\{0, 1 - |u - (1 - w)|/w\}$ and

$$S = T_{a,\delta}(U) = U + a\delta b_w(U), \qquad 0 \le \delta < w.$$

Given a, percentile–label pairs are iid. The increasing bijection preserves ranks while $\|F_{a,\delta} - F_0\|_\infty = \delta$ for uniform F_0, not necessarily the closest reference. Types, transforms and percentiles are unobserved; a known invertible warp would not lose information.

Design. At $B = 4$, $K = 16$, $M \in \{512, 4096\}$ and $\delta \in \{0, 1/256, 1/64, 1/32\}$, six prespecified responses give 48 cells:

	Original	Parameter-shifted
Quadratic	$.2 + .4u^2 + .15a$	$.15 + .5u^2 + .10a$
Threshold	$\mathbf{1}\{u \ge t_a\}$, $(t_+, t_-) = (.85, .97)$	$\mathbf{1}\{u \ge t_a\}$, $(t_+, t_-) = (.87, .98)$
Bit	$(a + 1)/2$	$.5 + .4a$.

Each cell uses 128 evaluation records, paired across δ and methods, independent of the development pilot. All rules pay $C = L = MB$, including pilot labels: pooling; physical-block plateau; centered-gated orders 1–4 (85) (equal prompt split, all pilot scores, $T = 8$); half-shrinkage (average unclipped orders 1 and 2, then clip); and ungated orders 2–4 and half-shrinkage with a one-score pilot.

Bias reduction need not improve accuracy. Across nine correction/shrinkage variants, paired Student-t intervals at $.05/432$ compare simulated MAE with pooling (maximum Monte Carlo standard error .007). Order two improves in 7 cells and worsens in 37; orders 1, 3, 4 and shrinkage improve in 0, 3, 0 and 9, respectively, with the same counts for ungated variants (Tables 11–12). At $M = 4096, \delta = 1/32$, it improves both thresholds but worsens both quadratics: for the original threshold, bias/standard deviation changes from $.057/.011$ to $-.018/.040$. Plateau wins both bit controls. All order-two gains occur at reference radii $\delta \ge 1/64$; even the smallest possible radii $1/80, 1/48$ give $K\delta \ge 1/5$, outside the local regime of Theorem B.6. These are fixed-method comparisons, not deployment guarantees.

Table 11: Original response parameters: MAE $\times 10^3$ on independent evaluation data. P: pooling; C: centered-gated order two; S: fixed half-shrinkage; L: physical-block plateau. These are fixed methods, not test-selected winners.

	$M = 512$				$M = 4096$			
δ	P	C	S	L	P	C	S	L
Quadratic response								
0	26.1	84.8	58.9	89.6	9.5	33.9	22.5	89.6
$1/256$	26.5	85.4	58.8	89.6	10.0	34.4	22.6	89.6
$1/64$	29.4	88.0	58.6	89.6	15.0	35.8	23.2	89.6
$1/32$	35.3	92.9	58.8	89.6	25.1	38.8	23.8	89.6
Threshold response								
0	27.3	95.9	70.3	360.8	9.1	28.3	21.8	360.2
$1/256$	27.6	97.8	70.6	360.8	11.8	28.8	21.7	360.2
$1/64$	39.9	103.9	72.2	360.8	35.4	30.6	24.3	360.2
$1/32$	57.5	112.0	73.9	360.8	57.0	34.2	26.6	360.2
Bit control								
0	30.6	71.1	51.7	17.8	9.9	23.6	16.7	5.6
$1/256$	31.4	72.9	51.3	17.8	14.7	24.3	18.1	5.6
$1/64$	52.1	78.4	52.4	17.8	48.8	26.4	27.3	5.6
$1/32$	88.5	85.3	57.5	17.8	90.8	30.5	42.5	5.6

Table 12: Prespecified alternative parameters: MAE $\times 10^3$ on independent evaluation data. Method definitions match Table 11.

	$M = 512$				$M = 4096$			
δ	P	C	S	L	P	C	S	L
Changed quadratic response								
0	24.6	90.1	66.0	109.0	9.0	27.4	20.4	110.9
$1/256$	25.2	90.1	65.8	109.0	9.2	27.4	20.3	110.9
$1/64$	27.1	92.1	65.9	109.0	11.5	27.7	20.5	110.9
$1/32$	30.2	96.0	65.9	109.0	16.6	30.1	20.7	110.9
Changed threshold response								
0	27.5	80.8	61.1	334.0	9.7	27.6	20.6	331.8
$1/256$	29.0	82.0	62.1	334.0	13.9	28.1	20.7	331.8
$1/64$	44.4	85.6	65.0	334.0	42.4	30.2	24.4	331.8
$1/32$	68.2	91.4	68.8	334.0	72.7	33.3	30.8	331.8
Noisy bit control								
0	27.0	79.1	53.1	16.8	10.7	25.1	17.8	6.3
$1/256$	28.3	80.4	53.3	16.8	13.0	25.5	19.4	6.3
$1/64$	43.9	84.8	54.8	16.8	38.8	26.8	26.5	6.3
$1/32$	74.0	92.2	57.9	16.8	72.3	29.4	37.6	6.3

D.2.5 Correction order inside the approximate-sharing model

We separate exact bias reduction from the cost of estimating correction (Figure 6). Family W uses the warp in Appendix D.2.4, with $w = 1/K$ and bit, scaled-threshold or quadratic responses. T1 is the three-band construction of Appendix D.2.7, indistinguishable at depth one; T2 is a variant indistinguishable through depth two.

The prespecified finite-budget grid uses $B \in \{2, 4, 8, 16\}$, $M \in \{2^{10}, 2^{12}, \ldots, 2^{18}\}$, $K \in \{32, 64\}$ and $K\delta \in \{0, 1/64, 1/16, 1/4, 1/2\}$, with bit and threshold responses and 200 records per (B, M, K). The primary rule uses a one-score pilot, orders $j \leq \min\{B, 5\}$ and clipped RMSE. Equal-cost comparisons use $N = 2K^3$, $\delta = K^{-5/4}$, $K = 8$–256, at depths one and two.

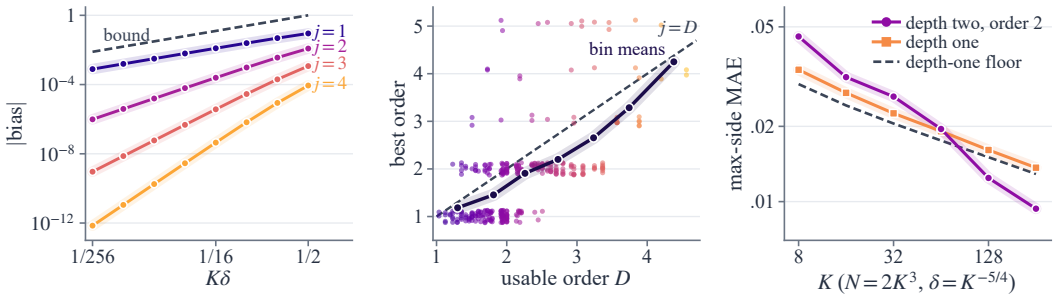


Figure 6: **Bias reduction and finite-data correction.** Left: exact order-j bias versus $K\delta$ for a bit response, known marginal CDF and $K = 128$ (dashed: pooling bound $\min\{1, 2K\delta\}$). Centre: RMSE-optimal order with a one-score pilot versus D in Eq. (16): 320 jittered cells, bin means and $j = D$. Right: max-side MAE of depth-one pooling and depth-two order-two correction at equal cost $N = 2K^3$, $\delta = K^{-5/4}$, with the exact lower bound for every depth-one rule on the depth-one-indistinguishable pair.

Bias and estimable order. Exact calculations at $K = 16$–128, $K\delta = 1/256$–$1/2$ give pooling bias 0.06–$0.20\,K\delta$, respecting (64). With the marginal CDF known, orders 1–4 have log-slopes $1.00, 2.00, 2.99, 3.99$: each companion adds a power of $K\delta$. Finite data change the comparison. Across 320 positive-δ cells, the RMSE-optimal order correlates with D at .63 (depth-preserving null 97.5% quantile .42), but with depth alone at .77, so this grid does not isolate D from depth. The best correction beats pooling in 109 cells; pooling wins 129 of 140 cells in the local regime ($K\delta \le 1/16$, $K \ge 2B + 2$; Theorem B.6) and all 80 exact-sharing cells.

Equal-cost depth. On T1, depth-one pooling is within 6–14% of its exact lower bound and has slope $-.25$ over $K = 16$–256. Depth-two correction slopes are $-.44$ to $-.51$ across families (95% intervals within $[-.60, -.38]$); the T1 slope gap .21 ($[.14, .28]$) is consistent with $1/4$. Post hoc, depth two loses for $K \le 32$ but beats the depth-one bound at $N = 2^{25}$ (.0094 versus .0129, simultaneous intervals). On T2 the same rule is 4–12 times worse than depth-one pooling. These cells have $K\delta \in [.25, .59]$, outside the local theorem, and all designs meet tolerance $K^{-3/8}$; they illustrate a precision mechanism, not that theorem's guarantee.

D.2.6 Fixed total cost: depth versus independent prompts

At fixed $N = MB$, greater depth permits higher correction order but reduces prompt coverage (Figure 7). We measure this tradeoff for fixed rules, not minimax risk.

Using the three original populations of Appendix D.2.4, we hold the deployment target fixed across B and δ: $K = 16$, $N \in \{4096, 16384\}$, $B \in \{2, 4, 8, 16\}$ and $\delta \in \{0, 1/256, 1/64, 1/32\}$, with 256 repetitions per $(N, B, \text{population})$ paired across δ and methods. All rules pay $C = L = N$, including pilot labels: pooling, centered-gated orders one/two, half-shrinkage, block winners, inverse-budget extrapolation and rank-polynomial fitting. The latter fits rank means to $\prod_{i<q}(r + i)/(B + 1 + i)$, $q \le \min\{2, B - 1\}$, and returns $\Pi_{[0,1]}[\sum_q c_q K/(K + q)]$. No rule is selected using test errors.

Table 13 reports prespecified contrasts with Bonferroni $.05/8$ Monte Carlo intervals. At $\delta = 1/32$, increasing B from 2 to 8 helps order two on thresholds but hurts block winners on bits. At $N = 16384$, threshold standard deviation falls from .068 to .030 while bias stays near $-.016$; constant bit labels gain no information from deeper pools. The other contrasts favor pooling over unnecessary correction at $B = 2$ under exact sharing, and rank-polynomial fitting over order two on the quadratic response at $B = 8, \delta = 1/32$.

Direct depth-16 evaluation attains threshold MAE .012, versus .026 for depth-eight correction, even when charged $N = 16384$ labels instead of the 1024 winner labels it needs. There is no demonstrated saving over unrestricted acquisition. The largest radius and $B = 8$ fall outside the local

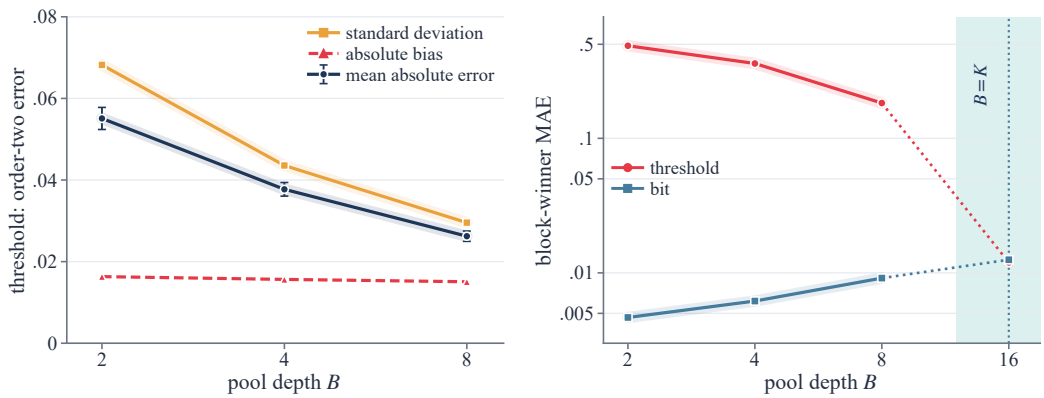


Figure 7: **Depth and prompt coverage at fixed cost.** $N = 16384$ candidates and labels, $K = 16$, $\delta = 1/32$; 256 repetitions. Left: order-two bias and standard deviation on the threshold response. Right: block-winner MAE on threshold and bit responses as $M = N/B$ decreases (log scale). Shading and dotted connections distinguish direct evaluation at $B = K$ from extrapolation. Bars show one Monte Carlo standard error of MAE only, not uncertainty in bias or standard deviation.

Table 13: Fixed-cost comparisons: MAE differences and simultaneous Monte Carlo intervals, all multiplied by 10^3. A negative difference favors the first named rule. Both improvements and reversals persist at both costs.

Contrast	N	Difference	Lower	Upper
Threshold depth	4096	−76.27	−93.54	−59.00
	16384	−28.86	−37.07	−20.65
Bit depth	4096	8.95	6.36	11.54
	16384	4.48	3.15	5.81
Exact sharing	4096	92.45	78.95	105.95
	16384	38.02	31.72	44.32
Quadratic baseline	4096	−27.65	−34.05	−21.25
	16384	−16.17	−19.65	−12.68

theorem ($K\delta > 1/16$, $K < 2B + 2$); these comparisons describe measured error, not its local exponent.

D.2.7 Finite-cost estimation and decisions from paired records

A controlled family separates the information in equal-cost depths one and two, while exposing the bias–noise tradeoff of fixed correction rules. These are not the tail-gated estimator of Theorem 5.2.

Model and rules. Let $K = 8$, $w = p_0 = 1/K$. Scores are uniform within the three thirds of $[0, 1]$, with masses $(1 - w - p, w, p)$ and $p = p_0 + t_x\delta$. Iid fair types t_x are frozen on $J = 2^{62}$ observed identities and hidden from estimators. Only the middle band can be correct: for type zero on side A and type one on side B. CDFs lie within $\delta/2$ of their midpoint. Writing $\psi_K(p) = (1-p)^K - (1-p-w)^K$, the prior target centers are $\vartheta_A = \psi_K(p_0)/2$, $\vartheta_B = \psi_K(p_0 + \delta)/2$; expected frozen-population deviation is at most $e_J = 1/(2\sqrt{J})$.

All rules pay $C = L = N$, with $M = N/B$. Pooling uses the full record. Corrections use first scores from half the prompts to estimate $\widehat{H}$; an anchor (s, G) and companion s' give $h = \widehat{H}(s)$, $I = \mathbf{1}\{s' \le s\}$ and scores

$$KGh^{K-1}, \qquad KG\big[h^{K-1} + (K-1)h^{K-2}(I-h)\big], \tag{111}$$

averaged over analysis prompts and anchors, then clipped.

Table 14: Same-record $B = 2, K = 8$ comparison: maximum sidewise MAE times 10^3. Bold entries resolve a winner against both other rules under the simultaneous Monte Carlo comparisons. First order is never resolved best.

	$N = 2^{18}$		$N = 2^{20}$		$N = 2^{22}$	
$K\delta$	Pooling	Order 2	Pooling	Order 2	Pooling	Order 2
0	.9059	3.3578	**.4797**	1.8280	**.2395**	.8923
$1/64$	1.2269	3.6469	.9719	1.6954	1.0141	.8036
$1/32$	2.0163	3.8454	2.0344	1.8215	1.9618	**.8983**
$1/16$	4.1676	3.4226	3.9635	**1.6856**	3.9852	**.8208**

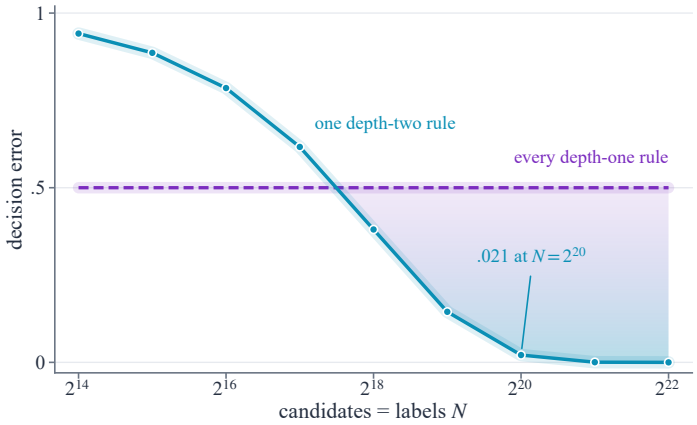


Figure 8: **Same cost and model knowledge, different information.** Analytic bounds for a declared threshold decision at $K = 8$, $\delta = 1/128$, where both designs pay N candidates and labels: the balanced-prior error of every depth-one rule stays near $1/2$ (dashed), while one depth-two counting rule, which reads within-prompt label–band pairs, has prior-averaged error below .021 on either side at $N = 2^{20}$. These are bounds, not measured errors.

Estimation. Across the depth-two grid (100 repetitions per cell), simultaneous Bonferroni Monte Carlo comparisons favor pooling in two cells, order two in three and neither in seven (Table 14). Even with known marginal H, at $K\delta = 1/16, N = 2^{22}$ the standard deviation .0010 exceeds the residual bias bound .00006. At $\delta = 1/128$, identical single-candidate mixture laws imply every depth-one rule has prior-averaged risk at least $L_N \simeq .00394$ (66). Depth-two Monte Carlo upper bounds fall below it at $N = 2^{20}, 2^{22}$ (.00198, .00107), but not 2^{18} (.00414). Under exact sharing at $N = 2^{18}$, unnecessary correction worsens MAE from .0010 to .0017/.0033 for orders one/two.

An equal-cost threshold decision. Set $\tau = (\vartheta_A + \vartheta_B)/2 \simeq .1178$ and $\Gamma = \vartheta_A - \vartheta_B > 0$ at $\delta = 1/128$. Both depths know the bands, w, p_0, δ and the two families, not the types. Equal priors, identical single-candidate laws and Markov's inequality for target crossing yield depth-one balanced error at least $\ell_N = [1 - N(N-1)/J]/2 - 2e_J/\Gamma > .4999$ at all three budgets.

A direct observable pair suffices. For $m = N/2$ depth-two prompts, $Y_i = G_{i1}\mathbf{1}\{S_{i2} \geq 2/3\} + G_{i2}\mathbf{1}\{S_{i1} \geq 2/3\} \in \{0, 1\}$ has means $q_A = 1/64$, $q_B = 17/1024$ under fresh fair types. Declare the target above τ when $\sum_i Y_i < \lceil mq_c \rceil$, $q_c = 33/2048$. Chernoff's bound, coupling until a frozen identity repeats, and the target-crossing bound give

$$R^{\text{dec}}_{\text{pair}} \leq e^{-md_*} + \frac{\binom{m}{2}}{J} + \frac{2e_J}{\Gamma}, \qquad d_* = \min_{s \in \{A,B\}} D(q_c \| q_s), \tag{112}$$

Table 15: Retrospective decisions at the analytic midpoint: errors on A/B, each out of 100 paired repetitions. The last row is a probability upper bound, not a count. True-marginal-H rules are privileged diagnostics; zero errors do not establish zero risk.

Rule	$N = 2^{18}$	$N = 2^{20}$	$N = 2^{22}$
Depth 1, pooling	47/56	59/46	55/40
Depth 1, empirical order 1	51/51	62/47	50/45
Depth 2, empirical order 1	45/44	56/46	43/42
Depth 2, empirical order 2	21/14	3/2	0/0
Depth 1, true-H order 1	53/59	58/47	55/34
Depth 2, true-H order 1	46/47	53/41	56/37
Depth 2, true-H order 2	21/14	3/2	0/0
Empirical order-2 upper bound	.365406	.130018	.071577

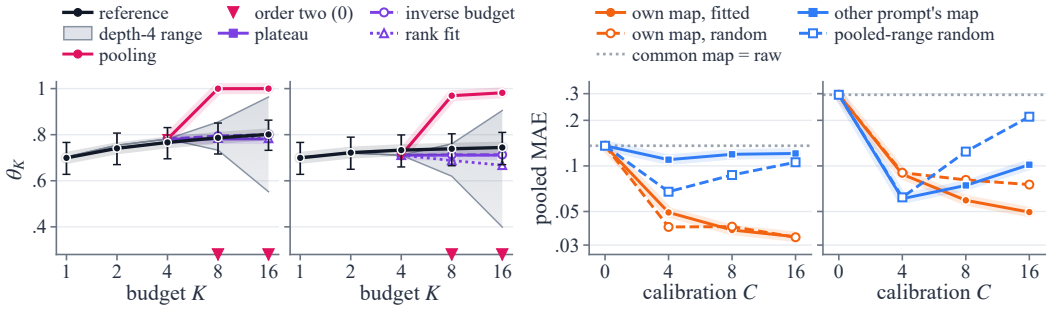


Figure 9: **Natural-score prediction and calibration.** Left pair: prospective predictions ($B = 4$, 128 prompts), process scorer and second verifier. Grey: ranges fixing four empirical rank means without shape constraints, not confidence bands or full-word ranges. Black: 64-candidate references with approximate 95% bootstrap intervals. Triangles: order-two predictions of zero at $K = 8, 16$; sharing fails ($\delta > 1/4$). Right pair: calibration MAE versus scores C per prompt, Qwen and Mistral. All maps preserve within-prompt order. Warm: own-prompt fitted or random-range maps; cool: another prompt's fitted map or a random pooled-range map; dotted: one common map, equivalent to raw pooling.

where D is binary KL. At equal cost $C = L = N = 2^{20}$, this is below .021. Figure 8 compares these analytic, side-prior-averaged bounds for the declared family. They are not per-population guarantees.

Thresholding the same correction estimates gives Table 15. Only depth-two order two has balanced error well below $1/2$; its simultaneous 95% exact-binomial upper bounds are below ℓ_N at all three budgets. True-H diagnostics show the same pattern without a certified bound. These probabilities concern the controlled prior, not full-model minimax risk.

D.3 Do natural scores satisfy the recovery conditions?

Natural-score studies test applicability and compare fixed prediction rules. Calibration and direct acquisition pay for their additional information; prospective predictions precede reference acquisition.

D.3.1 Historical banks: sharing validity and prediction error

Data and applicability. Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3 answer the same 96 MATH-500 questions, disjoint from 48 pilots (Hendrycks et al., 2021; Lightman et al., 2024; Yang et al., 2024). Independent 32-response protocol and 64-response reference banks use temperature

Table 16: Real-score aggregate-target trial at $M = 96, B = 2$. All primary reference gates fail. MAEs compare the same paid record; the reference uses a separate bank.

System, K	Reference	Order-2 MAE	Plateau MAE	Split gap	95% half-width
Q, 6	.85647	.8275	.0289	.00518	.00692
Q, 8	.85982	.8500	.0319	.00656	.00774
Q, 12	.86360	.8538	.0353	.00732	.00887
M, 6	.26878	.1902	.0976	.01499	.01466
M, 8	.28962	.2367	.1184	.01310	.01713
M, 12	.31678	.2835	.1456	.01006	.01969

.7, top-$p = .8$ and a 1,024-token cap. Math-Verify labels are paired with target-blind sums of Qwen2.5-Math-PRM-7B log rewards over blank-line-delimited steps (Zhang et al., 2025), not its default aggregation. Protocol correctness is $.80/.12$ and proxy–label AUC $.97/.89$ on Qwen/Mistral.

Simultaneous 95% CDF radii of .208 for 96 responses per prompt, together with disjoint prompt score ranges, imply $\delta \geq .292$ in each bank (Appendix B.2). Thus these tests lie outside both Theorem 5.2 and the local regime ($K\delta \geq 1.7$ for $K \geq 6$, versus $K\delta \leq 1/16$).

Prediction error. Each of 1,000 fixed-bank replays uses $M \in \{48, 64, 96\}$ distinct prompts with two protocol candidates ($C = L = 2M$), predicting $K \in \{6, 8, 12\}$. Rules are pooling, ungated corrections (111) and depth-two plateau. Corrections fall back to $1/2$ when $M < 8K$; six of 18 cells are excluded from comparisons. The hidden reference averages Best-of-K U-statistics over all 96 prompts. All six references fail the predeclared precision gate for a .01 gain: split-half gap at most .01 and bootstrap half-width at most .005 (Table 16). Order two beats plateau in none of 12 non-fallback cells, and pooling only once (Mistral, $M = 96, K = 12$: .2835 versus .2925).

Local correction fails off-regime. At $M = 96, K = 8$, 93%/62% of Qwen/Mistral estimates clip to zero (raw means $-.89/-.18$). The linearization requires $|F_x - h| \ll 1/K$. Empirical within-prompt-rank versus pooled-percentile gaps average $.262/.181$, versus .055 under sharing, and exceed $1/K$ for 79%/68% of anchors at $K = 12$. These are diagnostics, not estimates of $\mathcal{D}_{\text{cal}}$ or bounds on δ. Using privileged full-bank ranks, unexpanded errors are $.003/.014$ but linearized errors $2.98/1.08$; gating at $h > 1 - T/K$, $T \in \{8, 4, 2, 1\}$ leaves Qwen MAE above .73. This exposes Taylor error without identifying a unique cause or a same-cost remedy.

D.3.2 Paired budget gains, score aggregation, and a simple baseline

This exploratory reanalysis uses log-product, minimum and last-step scores on the same banks. Each aggregation defines its own selector and target.

Reference gains. For $d = \theta_8 - \theta_2$, approximate pointwise percentile intervals use either candidate bootstraps within the 96 fixed prompts or prompt bootstraps, always differencing estimates on the same 64 candidates. All six intervals exceed .01, but candidate-bootstrap half-widths remain above the .005 precision gate (Table 17). Thus budget dependence is detectable without supporting the prespecified precision requirement; prompt resampling alone does not establish transfer to other tasks.

Same-record extrapolation. The parameter-free inverse-budget rule is $\Pi_{[0,1]}[2(1 - 1/K)\widehat{\theta}_2 - (1 - 2/K)\widehat{\theta}_1]$, using the mean label and mean block-winner label. It has the smallest MAE in all six comparisons (Table 18), improving over plateau by .008–.010 on Qwen and .032–.035 on Mistral (paired Monte Carlo standard errors below .001); both corrections are worse under every aggregation. For d, its reference-conditional MAEs are approximately .009 and .059–.070, not guarantees on deployment error.

Table 17: Paired reference gains from depth 2 to 8. Each score has its own target. The two interval columns answer different sampling questions; neither certifies the sharing assumption or the true deployment gain.

System	Score	θ_8	d	Candidate 95% interval	Prompt 95% interval
Q	Log product	.85982	.02818	[.02203, .03406]	[.01564, .04346]
Q	Minimum	.85910	.02831	[.02208, .03469]	[.01634, .04278]
Q	Last	.85912	.02890	[.02258, .03516]	[.01313, .04571]
M	Log product	.28962	.10508	[.09009, .11270]	[.07687, .13502]
M	Minimum	.27060	.09278	[.07938, .09995]	[.06455, .12265]
M	Last	.28844	.10502	[.09013, .11190]	[.07690, .13499]

Table 18: Same-record target MAE at $M = 96, B = 2, K = 8$, charged 192 candidates and labels. Parentheses give one Monte Carlo standard error in units of the last displayed digit. Errors condition on each aggregation's noisy reference; these uncertainties exclude reference error.

System	Score	Pooling	Order 1	Order 2	Plateau	Inverse budget
Q	Log product	.1395(1)	.1481(28)	.8500(15)	.0319(6)	.0219(5)
Q	Minimum	.1401(1)	.1441(28)	.8468(17)	.0323(6)	.0223(5)
Q	Last	.1388(2)	.1438(27)	.8477(16)	.0299(6)	.0221(5)
M	Log product	.2147(21)	.2534(58)	.2367(27)	.1184(9)	.0833(12)
M	Minimum	.2120(21)	.2496(57)	.2152(28)	.1019(9)	.0695(11)
M	Last	.2182(21)	.2543(58)	.2353(27)	.1178(9)	.0833(12)

D.3.3 Disjoint calibration and its opportunity cost

Can additional, paid calibration information improve pooling? Using the historical banks at $M = 96, B = 4, K = 12$, we run 512 paired replays per system. Four candidates supply evaluation labels; the next $C \in \{0, 4, 8, 16\}$ scores fit an order-preserving piecewise-linear map to logit midranks ($C = 0$: identity), without evaluation labels/ranks or references. At equal charge $N = 96(4 + C)$ candidate–label pairs, calibrated rules use four labeled candidates, raw rules use all $4+C$, and direct evaluation uses $N/12$ depth-12 pools. No small sharing radius is assumed.

Calibrated pooling improves over both raw depth-four and equal-cost raw pooling in all twelve positive-C contrasts. Table 19 reports point estimates; the comparisons are supported by Bonferroni-adjusted 95% Monte Carlo intervals. Nevertheless, direct evaluation and raw quadratic fitting have lower estimated MAE in all six positive-C cells; correction remains unreliable (MAE .12–.86).

Four order-preserving controls isolate the information in each map: reassign fitted maps across prompts, use one common map, or fit random knots within each prompt's visible range or the pooled range (Figure 9, right). Beyond $C = 4$, own-prompt maps hold or improve while reassigned and pooled-range maps degrade; at $C = 4$, the latter give Mistral's lowest errors among these maps. At $C = 16$, fitted maps beat reassigned, common and pooled-range maps on both systems (simultaneous 95% intervals over 36 contrasts). Own-range maps are indistinguishable on Qwen (difference .0003, $[-.0050, .0055]$) and worse by .026 ($[.018, .033]$) on Mistral. At $C = 16$, these controls support a benefit from prompt-specific information, but do not establish sharing or an advantage over deeper acquisition. Intervals condition on these banks; both references fail the .005 precision gate.

D.3.4 A second verifier of a different family

Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2025) rescores the same generations and labels. Its within-prompt rank correlations with the process model are only .19/.23 on Qwen/Mistral. Each scorer defines a separate target; this comparison does not disentangle generator–scorer interactions.

At $M = 96, B = 4, K = 12$, MAEs for pooling, order two, plateau, quadratic fitting and direct evaluation are .173/.824/.015/.029/.046 on Qwen and .173/.259/.075/.051/.060 on Mistral, with

Table 19: Natural calibration: MAE to the finite reference times 10^3. Each row charges $N = 96(4+C)$ candidate–label pairs. Raw methods use all $4+C$ candidates; direct evaluation has depth 12. Calibration improves pooling but does not beat every use of the same resources.

System	C	Calibrated pooling	Raw pooling	Raw quadratic	Direct
Qwen	0	135.90	135.90	26.60	40.87
	4	49.30	135.88	12.31	25.01
	8	37.76	135.93	8.28	15.21
	16	34.06	135.91	4.85	14.72
Mistral	0	294.99	294.99	44.52	55.33
	4	88.10	292.94	29.42	33.99
	8	59.27	292.71	24.42	20.29
	16	49.55	291.17	22.42	16.32

Table 20: Sharing certificates and Best-of-16 pooling under nine scorers. Disjoint: nonoverlapping prompt score ranges among 4,560 historical or 8,128 prospective pairs. AUC: pooled prospective correct-versus-incorrect separation. Last two columns: post hoc pooled signed error and |pooled error| − |plateau error| against each scorer's own 64-candidate reference; approximate 95% intervals use Bonferroni correction over nine contrasts within each scorer.

	Disjoint pairs				Prospective, $K = 16$	
Scorer	Qwen	Mistral	Prosp.	AUC	Pooled error	Pooled − plateau
Process model	799	2	1378	.962	$+.199$ $[.137, .268]$	$+.179$ $[.082, .278]$
Skywork-V2 8B	1119	63	1621	.878	$+.238$ $[.167, .305]$	$+.204$ $[.083, .310]$
ArmoRM 8B	1647	348	1976	.611	$-.029$ $[-.196, .133]$	$-.010$ $[-.078, .237]$
OffsetBias 8B	780	56	1171	.747	$+.178$ $[.072, .270]$	$+.145$ $[-.038, .278]$
GRM 8B	568	33	894	.638	$+.056$ $[-.099, .188]$	$-.002$ $[-.106, .189]$
URM 8B	646	111	879	.794	$+.203$ $[.112, .279]$	$+.182$ $[.045, .305]$
Skywork v0.2 8B	463	39	805	.764	$+.150$ $[.035, .247]$	$+.133$ $[-.029, .273]$
Skywork-V2 3B	1274	63	2029	.867	$+.241$ $[.179, .304]$	$+.238$ $[.135, .319]$
Skywork-V2 Qwen3-8B	1573	52	2413	.907	$+.234$ $[.169, .303]$	$+.229$ $[.130, .318]$

the same ordering as the process scorer. At $C = 16$, fitted calibration again beats reassigned, common and pooled-range maps in all six simultaneous comparisons. Own-range maps beat fitted maps on Qwen but lose on Mistral.

Rank–percentile gaps are $.273/.203$ (process scorer $.262/.181$; shared-score control $.055$); disjoint prompt ranges give $\delta \geq .292$ at confidence $.95$ in each bank. Appendix D.3.7 reports the separately sealed rescoring of prospective data.

D.3.5 Seven further reward models

We rescore the same 28,672 historical and prospective responses with the nine models in Table 20, keeping generations and labels fixed. Inputs use each model's chat template, one beginning-of-sequence token, no truncation, a bf16 body and fp32 score head.

Disjoint prompt score ranges occur in all 27 scorer–bank cells, giving $\delta \geq .292$ in each historical bank and $\delta \geq .268$ prospectively ($.254/.228$ simultaneously across cells). A single disjoint pair reaches this sample-size ceiling, so the certificate cannot rank scorers. All nine discriminate correctness (AUC intervals above .5).

In the post hoc primary-panel comparison at $K = 16$, pooling over-predicts by .15–.24 under seven scorers and is resolved worse than both plateau and inverse-budget extrapolation under five (four with across-scorer correction). The two weakest discriminators have errors unresolved from zero; ArmoRM's ordering is also sensitive to bf16 batching. The simple rank rules stay within .06 of each primary reference (.066 across panels), quadratic fitting within .08; order two returns zero and no budget decision is resolved. This fixed-scorer, single-benchmark study tests applicability, not generalization across reward models or new prompts.

D.3.6 Retrospective order sensitivity

An off-regime Qwen check swaps 48 pilot and 48 analysis prompts, using one or all 32 pilot scores per prompt and consecutive B-candidate blocks (first block primary). The six (B, K) cells are $(4, 16), (4, 32), (4, 64), (8, 32), (8, 64), (16, 64)$; orders $j \leq \min\{B, \lfloor (K-1)/2 \rfloor\}$ use Eqs. (96) and (84). Across all 24 split–pilot configurations, higher orders have greater block-aggregated absolute bias and standard deviation than order one. Raw estimates alternate in sign and reach -758 at $B = 4, K = 64, j = 4$; order-one raw bias is already .07–.56. Two first-block comparisons at $B = 16, K = 64$ nominally favor order three, but its raw average is $+81$ and the paired bootstrap interval crosses zero. These checks expose off-regime instability, not a validated order-selection procedure.

D.3.7 Prospective prediction with an independent deeper reference

Predictions were frozen before deeper reference acquisition, testing practical transfer without selecting methods on their observed errors. The corrections are fixed-order rules, not the full minimax construction; the sharing-radius bounds below exclude its regime.

Protocol. We sample 128 additional MATH-500 problems, absent from our banks but not necessarily pretraining data. Qwen2.5-7B-Instruct generates 16 protocol candidates per prompt: four primary candidates and three disjoint sensitivity panels on the same prompts. Predictions and budget choices at $K = 4, 8, 16$ (16 primary) precede generation of 64 independent reference candidates. Sampling, scoring and boxed-priority labeling follow Appendix D.3.1; no response is removed or regenerated based on its outcome.

Each shallow rule pays 512 candidates and labels: pooling, ungated orders one/two (111), plateau, inverse-budget extrapolation and a quadratic fit to four rank means. At equal cost, direct evaluation uses $512/K$ depth-K prompts per target, or 32 depth-16 prompts for a curve. The reference averages all K-subset winners within each prompt's 64 candidates. Reported errors are single-run discrepancies from this noisy reference, not MAEs.

Prediction and uncertainty. Only pooling and order two have discrepancy intervals excluding zero (Table 21); orders one/two clip to one/zero in all four panels. On the remaining panels, pooling misses by .188–.199, versus at most .036 for the three rank rules. The inverse-budget discrepancy .00057 is smaller than reference uncertainty: its $K = 16$ interval is $[.732, .863]$ and independent 32-candidate halves differ by .02480.

Rules maximize $\widehat{\theta}_K - \lambda(K-1)$, breaking ties toward smaller K, for the five declared sensitivity costs in Table 22, not measured prices. At $K = 4$, shallow rules use the winner mean .781 (direct curve .757, reference .766). Reference gains for $4 \to 8$, $4 \to 16$ and $8 \to 16$ are $.020, .035, .015$, smaller than the simultaneous 95% paired-gain Hoeffding radius $\sqrt{2\log(120)/128} = .274$. No primary budget choice is resolved against all alternatives; pointwise bootstrap significance does not certify optimality.

The grey ranges in Figure 9 match four empirical rank means: $[.552, .963]$ for the process scorer and $[.399, .905]$ for the second verifier at $K = 16$, collapsing to points for $K \leq 4$. They ignore sampling error and full-word constraints; reference inclusion means compatibility, not coverage.

Table 21: Prospective primary-panel predictions. The last column is an approximate paired prompt-bootstrap 95% interval for the *signed* discrepancy $e_{16} = \widehat{\theta}_{16} - \widehat{\theta}^{\mathrm{ref}}_{16}$; it is not a risk certificate. Intervals were not computed for the direct arms. All predictions are clipped; orders one and two saturate at one and zero, respectively.

Method	$\widehat{\theta}_8$	$\widehat{\theta}_{16}$	$\lvert e_{16}\rvert$	Interval for e_{16}
Pooling	.99967	1.00000	.19864	$[.13690, .26789]$
Order 1	1.00000	1.00000	.19864	$[-.42439, .25554]$
Order 2	.00000	.00000	.80136	$[-.86310, -.73198]$
Plateau	.78125	.78125	.02011	$[-.05758, .01575]$
Inverse-budget	.79427	.80078	.00057	$[-.04294, .03982]$
Quadratic rank fit	.78516	.78393	.01742	$[-.08240, .04884]$
Direct pointwise	.79688	.78125	.02011	—
Direct curve	.77354	.78125	.02011	—
Independent reference	.78646	.80136	—	—

Table 22: Budgets chosen before reference acquisition, for every declared cost. The final row optimizes the held-out reference and was unavailable to the rules. No primary choice is certified optimal by the simultaneous reference bounds.

Method $\setminus \lambda$	0	.001	.003	.01	.03
Pooling	16	8	8	8	8
Order 1	8	8	8	8	8
Order 2	4	4	4	4	4
Plateau	4	4	4	4	4
Inverse-budget	16	8	8	4	4
Quadratic rank fit	8	4	4	4	4
Direct curve	16	8	8	4	4
Reference optimizer	16	16	8	4	4

Second verifier and scope. The outcome-verifier predictions and budgets were sealed before reference scoring on these same responses, not a new prospective sample. Against its reference .74432, discrepancies are $.23750/.177/.74432$ for pooling/orders one/two, $.03339/.03143$ for plateau/inverse-budget, and $.077/.026$ for quadratic/direct evaluation. Again only pooling and order two exclude zero; no budget choice or pointwise gain is resolved.

Under either scorer, $\delta \geq .268$ at confidence .95, outside Theorem 5.2; $K\delta \geq 4.2$ at $K = 16$ also violates the local theorem, whose $K \geq 2B + 2$ condition excludes $K = 8$. Thus the study shows neither a correction advantage nor a failure of the guarantees, and does not attribute error uniquely to CDF mismatch. Conversely, rank-rule discrepancies far below the worst-case radius $\Delta_{4,16}/2 = .3004$ do not refute the information bound: worst-case guarantees do not describe every system's error.

D.3.8 Post hoc forecasting checks

Eight variants extend the rank rules: monotone fitting and range midpoints; logistic full-word/rank-mean fits; free-asymptote, unit-asymptote and exponentiated power laws; and correctness-calibrated Gaussian reward extrapolation. They were fixed after reference inspection but before computing their errors: this is post hoc, not independent prospective validation.

Across 384 paired contrasts (8 methods, 4 panels, 2 targets, 2 scorers, 3 baselines), none shows a statistically resolved improvement over the rank rules. Six variants have primary process-scored Best-of-16 discrepancies .002–.043, statistically indistinguishable from the rank rules. Gaussian reward extrapolation over-predicts by .07–.11; the common-response logistic full-word fit misses

by .10–.20. Its fitted all-equal-label fraction is approximately 25%, versus 78% observed. Historical gains for power-law, rank-mean logistic and reward forecasts occur mainly on Mistral. Its rising reference curve is a post hoc explanation, not shape established by shallow data. The reward-tail predictor of Li et al. (2026) targets reward, not correctness; a one-sample upper tail at depth four is insufficient to fit it meaningfully.